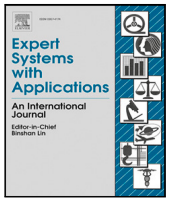
Expert Systems With Applications 231 (2023) 120679

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Towards interpretability in fingerprint based indoor positioning: May attention be with us

Andrea Brunello[1], Angelo Montanari, Nicola Saccomanno [*,1]

Department of Mathematics, Computer Science, and Physics, University of Udine, Via delle Scienze 206, 33100 Udine, Italy

ARTICLE INFO

Keywords:
Ubiquitous systems
Artificial intelligence
Neural network
Interpretability
WiFi fingerprinting
Indoor positioning

ABSTRACT

In a world increasingly pervaded by mobile and IoT devices, position-related information is gaining more and more importance. Highly accurate and standardized positioning techniques are not yet available for indoor scenarios, unlike for the outdoor case. The most commonly used method for indoor positioning is WiFi fingerprinting, which, despite its well-recognized advantages, still suffers from some notable limitations. Recently, approaches relying on deep learning showed promising results even though their lack of interpretability is still a significant drawback. In this paper, for the first time, we propose a domain-specific concept of interpretability, based on identifying the access points that are most relevant to a position estimate. The goal is to enhance the positioning process by gaining novel scientific knowledge and operational insights, without worsening the performance of the task. We show how it is possible to practically achieve both a local and a global notion of interpretability by means of a deep learning model equipped with an attention module, applied to a ranking based fingerprint representation. Since off-the-shelf application of attention does not guarantee to achieve a faithful nor plausible interpretation, we verified through a series of thoroughly designed quantitative and qualitative clustering based experiments the existence of a strong relationship between the obtained interpretations and the positioning domain. Finally, as by-product, we showed an example of how the new knowledge can be used in principle to improve positioning performance.

1. Introduction

Location-based services (LBS) have become a fundamental aspect of our everyday lives as they provide users with personalized services and real-time information based on their current location, fostering a wide range of applications, including navigation, social networking, healthcare, supply-chain management, and emergency response.

When outdoor scenarios are considered, worldwide accurate position estimations can be provided by the GNSS (Global Navigation Satellite System) service, which is supported by several mobile devices. Nevertheless, the satellite signal gets easily masked by buildings, thus it is hardly exploitable indoors, where people and objects spend most of their time. For these reasons, the research on effective alternatives applicable to the mobile and Internet of Things domains and capable of working in indoor premises is currently a highly relevant topic (Potortì, Crivello, Palumbo, Girolami, & Barsocchi, 2021).

Indoor positioning systems can rely on data collected by disparate sensors, and that refer to WiFi, Bluetooth, Ultra-wideband (UWB), RFID, cellular tower signal, magnetic field, light intensity, inertial measurements, and so on Mendoza-Silva, Torres-Sospedra, and Huerta (2019). Based on the considered information, different approaches can be followed to perform position estimation (Khalajmehrabadi, Gatsis, & Akopian, 2017; Xia, Liu, Yuan, Zhu, & Wang, 2017).

Among them, fingerprinting is by far the most studied (Mendoza-Silva et al., 2019). It is a conceptually simple technique based on two phases. The first (*offline*) one consists of the acquisition of the fingerprints, i.e., data observed by sensors at a specific time and position. The locations where the fingerprints are sampled are named Reference Points (RPs). The set of RPs is known, and it can be either defined a priori, or built in a more random fashion exploiting crowdsourcing (Wang, Guo, Wang, He, & Zhang, 2022). The second (*online*) phase refers to the actual usage of the positioning system. Here, a new fingerprint is generated at an unknown location by a user collecting the sensor data. Such a fingerprint is used to estimate the user position either by comparing it against those collected during the offline phase or by feeding it to a predictive model that had been trained on the latter data.

One of the most exploited data sources in fingerprinting is the WiFi signal emitted by the access points (APs) present in a building. In that case, each fingerprint is composed of the received signal strength (RSS)

* Corresponding author.
E-mail addresses: andrea.brunello@uniud.it (A. Brunello), angelo.montanari@uniud.it (A. Montanari), nicola.saccomanno@uniud.it (N. Saccomanno).
[1] Co-first authors.

https://doi.org/10.1016/j.eswa.2023.120679
Received 8 November 2022; Received in revised form 31 May 2023; Accepted 31 May 2023
Available online 5 June 2023

of the APs detected in a given position; the set of pairs (location, RSS vector) constructed during the offline phase is referred to as radio-map. WiFi based fingerprinting owes its popularity mainly to its low cost, as it can rely on an already existing WiFi infrastructure, and high precision (Xia et al., 2017). In particular, the large availability of WiFi APs, the possibility to directly exploit them, and the fact of not needing to know their position are key features (He & Chan, 2016; Khalajmehrabadi et al., 2017; Xia et al., 2017). Of course, there are also known issues that still need to be addressed both from the research as well as the application perspective so as to make the method universally adoptable. The major challenge refers to the radio-map construction and maintenance, which tends to be a time-consuming task, mainly due to environmental changes that may force its repetition over time (Khalajmehrabadi et al., 2017; Mendoza-Silva et al., 2019; Pérez-Navarro et al., 2019). Other issues arise from the uneven propagation of signals (due to multipath effects resulting from obstacles, random noise, body attenuation, changes in the WiFi network, and device heterogeneity) (Mendoza-Silva et al., 2019; Torres-Sospedra & Moreira, 2017), that is exacerbated by the weak correlation between signal and position variations (Saccomanno, Brunello, & Montanari, 2021). Recently, the lack of interpretability of localization approaches has also been questioned (Chen, Wang, Lu, Trigoni, & Markham, 2020). This last point in particular lacks in-depth studies, although it is of great importance considering that indoor positioning is mostly tackled by means of machine learning (Bai, Luo, Yan, & Wan, 2021; Hernández et al., 2021; Lee, Kim, & Seo, 2022; Nabati & Ghorashi, 2023). Indeed, even if a very good model to address the localization task was available, understanding why it works, to gain scientific and operational insights that go beyond the original task, is compelling. To mention one possible impact, it may help to predict unanticipated failure cases of the system.

A possible workaround to deal with some of the above-described issues, not very much explored in the literature, involves relying on specific representations of fingerprints. For instance, the authors of Wu, Xu, Yang, Lane, and Yin (2017) propose *fingerprint spatial gradient*, that exploits spatial features of fingerprints from multiple adjacent locations to reduce spatial ambiguity and temporal instability of classical fingerprinting. Another alternative is to make use of ranking based fingerprints, where the RSS information is only exploited to sort the APs detected at each location in decreasing order (Cheng, Chawathe, LaMarca, & Krumm, 2005; Machaj, Brida, & Piché, 2011). As a result, sequences of AP identifiers, without any explicit information on their RSSs, are considered. This allows to better deal with problems such as device heterogeneity and signal perturbations, even though, in the past, the general performances brought by this kind of representation were lower than full-fledged fingerprint based approaches (Cheng et al., 2005; Laoudias, Piché, & Panayiotou, 2013; Lohan et al., 2017; Ma, Wu, & Poslad, 2019; Machaj et al., 2011; Saccomanno, Brunello, & Montanari, 2020; Tiku & Pasricha, 2019).

To the best of the authors' knowledge, there exists just a single previous attempt to combine single (ranking based) fingerprints with deep learning models meant to operate on sequential (in our case ordinal) input data (Saccomanno et al., 2020). The current work builds on the encouraging results presented in that study (i.e., a performance comparable to the approaches based on full-fledged fingerprints are achieved) by introducing and focusing on the contributions brought by the attention mechanism (Bahdanau, Cho, & Bengio, 2015), which we believe is the enabling element for interpretability in fingerprint based indoor positioning. Our choice is motivated by domain knowledge rather than mere empirical findings: it is inherently true that some access points are more significant than others when it comes to determining the most probable location of a user, however, we question whether such a role is always played by the most powerful ones. Knowing what the most relevant access points are, irrespective from their RSS, may help a user to understand the cause behind a model prediction, which adheres to a local definition of interpretability, such as the one provided by Miller (2019). In addition, from a global perspective, access point relevance patterns may characterize different areas of a building and, in turn, this could benefit a variety of tasks, ranging from radio-map maintenance, to the identification of wrong predictions, and to the overall improvement of localization accuracy.

The main contributions of our work can be summarized as follows:

- we propose a concept of interpretability for the fingerprint based indoor positioning domain that links access point relevance with position estimation, and we discuss its practical implications in supporting several positioning-related tasks;
- we show that access point relevance patterns obtained through the attention scores of a sequence-to-sequence deep learning model for ranking based fingerprints have a strong spatial characterization;
- we perform a series of qualitative and quantitative experiments to quantify the extent to which different attention compatibility functions convey our notion of interpretability, identifying a clear superiority of the *additive* one;
- as a by-product of interpretability, we show how, assuming an optimal strategy to combine attention scores and deep learning model likelihoods, it is possible to improve the overall performance in positioning tasks;
- we base our analyses on a large multi-building multi-floor well-recognized indoor positioning dataset, allowing for a fair comparison with other works and for an overall reproducibility of the achieved results.

Table 1 provides a concise, although complete account of the research questions, analysis methodologies and results of the article.

The rest of the paper is organized as follows. In Section 2, an overview of approaches for positioning, especially deep learning based, is presented, followed by an account of interpretability in machine learning. In Section 3, we discuss our idea of interpretability for fingerprinting, and in Section 4 we detail the deep learning model we designed to achieve it. In Section 5, the devised experiments and their results are reported. Finally, in Section 6 a thorough discussion of the overall outcomes, their possible practical applications, and directions for future work is presented.

2. Related work

In this section, first, an account of related work on fingerprinting is provided. An overview of interpretability in machine learning and its applications then follows.

2.1. Fingerprinting

In the literature, many surveys review the approaches, the advancements, and the challenges related to the research on fingerprint based indoor positioning, e.g., He and Chan (2016), Khalajmehrabadi et al. (2017), Roy and Chowdhury (2021) and Xia et al. (2017). From an historical perspective, fingerprinting solutions can be partitioned into two different classes (He & Chan, 2016; Pérez-Navarro et al., 2019). The first is represented by deterministic algorithms, which are the most simple ones. Generally, they are easy to implement and require few computational resources (He & Chan, 2016). Many of these solutions, such as RADAR (Bahl & Padmanabhan, 2000), are based on comparing a given radio-map fingerprint with those collected during the offline phase, a task that heavily depends on the chosen similarity measure. An exhaustive comparison of 50 combinations of K-Nearest-Neighbour (K-NN) and similarity measures is provided by Torres-Sospedra, Montoliu, Trilles, Belmonte, and Huerta (2015). The other class of methods is the probabilistic one. These model the positioning problem with an arbitrary complex formulation $\hat{l} = \arg\max_l P(l|o)$, where o is the observed fingerprint and $\hat{l}$ is the most likely location. Horus (Youssef, Agrawala, & Shankar, 2003) is a representative of this class.

Table 1
Summary of the experiments and their results.

Experiment	Description	Method	Results
Quantitative	Are clustering results meaningful?	Analysis of the clustering dendrogram	*dot* is unsatisfactory due to poor partitioning; *general* and *deep* produce too many groups; *add* and *cat* generate balanced partitions that exhibit a clear compositionality
	Do clusters group instances sharing very specific attention patterns?	Statistical procedure based on the Kolmogorov–Smirnov test	*add* and *cat* have the best behaviour: they provide a high number of clusters but only few of them are pairwise similar
	Do different attention types exhibit *spatial characterization*?	Cluster compactness and separation based on the Hausdorff metric	*add* emerges as the best one, since it leads to (spatial) distance values for similar clusters that are typically lower than those for dissimilar ones (according to KS)
Qualitative	What do the different types of attention highlight?	Visual and comparative inspection	*general* and *deep* focus on the strongest access point(s); *dot* considers the last position in the padded rank; *add* and *cat* exhibit rather heterogeneous patterns
	Are attention based groups truly spatially compact?	Visual inspection	Yes, clusters obtained with *add* exhibit clear spatial locality properties
Positioning evaluation	Are plain model performances comparable with SOTA?	Success rate and positioning error	Yes, they are in par with or better than SOTA solutions
	Can we improve the results leveraging interpretability insights?	Positioning error distribution	Yes, by a large margin, assuming an optimal strategy to choose between attention scores and deep learning model likelihoods; regardless, a naive combination method still brings to an improvement

Many approaches, especially the most recent ones, are not classified into one of the two groups. Instead, very often the term "advanced techniques" is used to describe methods relying on more complex concepts, such as, for instance, sensor fusion, trajectories, and machine learning (Akram, Akbar, & Shafiq, 2018; Torres-Sospedra et al., 2016). Nevertheless, it is worth highlighting that, to some extent, even these solutions can be associated with the deterministic and probabilistic definitions. For instance, as we shall see, classification-oriented deep learning algorithms return probabilities, fitting the probabilistic framework.

Recently, several approaches have tackled the positioning problem through deep learning (Feng, Nguyen, & Luo, 2021). Nowicki and Wietrzykowski (2017) use a stacked autoencoder with a two-layer classifier to determine building and floor information. Kim, Lee, and Huang (2017), again, propose a stacked autoencoder combined with a feed-forward neural network. The goal is that of obtaining a scalable solution by jointly reducing the feature space dimension and classifying building, floor, and room in a hierarchical fashion. Soro and Lee (2018), with the aim of reducing RSS fluctuations, propose and compare the performance of a stacked autoencoder against an ensemble of neural networks, establishing the superiority of the latter. Song et al. (2019) consider three models for multi-floor localization. They all rely on stacked autoencoders, that are combined with either a simple classification layer or one-dimensional convolutions. Each model is devoted to the prediction of a single specific component of the position: building, floor, and room. Ibrahim, Torki, and ElNainay (2018) use several convolutional neural networks with RSS time-series encoded as images, so as to reduce the noise and randomness associated with single fingerprint measurements. As in other approaches, each submodel deals with a specific hierarchical level of the position, but taking also into account the prediction related the preceding level of the hierarchy. Shao et al. (2018) encode together WiFi and magnetic field as high-resolution images, so as to tackle the positioning problem, which is considered as a computer-vision one, by automatically learning the mapping between ground-truth positions and such generated data. In the last years, several works started investigating the usage of channel state information (CSI), that is, a particular data source related to the WiFi channels, yet not detectable by all WiFi-enabled devices, that showed promising and accurate results in combination with deep learning approaches (Foliadis, García, Stirling-Gallacher, & Thomä, 2021; Wang, Gao, Mao, & Pandey, 2015, 2017). CSI is exploited also by Li, Chen, Wang, Wu, and Liu (2022) where, although focusing on device-free fingerprint positioning, they propose a deep learning approach to deal with fingerprint inconsistency issues, formulating the problem as a domain adaptation task (e.g., to adapt to changes that over time affected a given scenario). Some other methods analyse fingerprints employing recurrent models, such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks. This allows one to manage sequential information that might arise in positioning, for instance, when dealing with trajectories (i.e., sequences of position/fingerprint) or odometry. Hoang et al. (2019), through the comparison of multiple RNN models applied to full-fledged fingerprint trajectories, show how to address problems such as spatial ambiguity and RSS instability. Bai et al. (2019) propose a cascade of two recurrent models jointly trained to estimate and subsequently filter trajectories of fingerprints and corresponding locations. Hsieh, Prakosa, and Leu (2018) perform classification (for floor identification) and regression (for latitude/longitude positioning) using two different recurrent models. Particular attention is devoted to determining whether the enrichment of the models (i.e., stacked/deep RNN/LSTM) provided any advantage. The result, similar to the one achieved by Sahar and Han (2018), shows that this is not the case.

Deterministic-like ranking based fingerprints combined with K-NN are proposed by Machaj et al. (2011). Among multiple considered metrics, Spearman's footrule emerges as the most performing one. A further extension that considers a more comprehensive set of distance functions is proposed by Machaj et al. (2011), concluding the superiority of the Lorentzian metric. Yang, Dessai, Verma, and Gerla (2013), instead of considering raw signals, devise an encoding of fingerprints based on RSS relationships, so as to manage device heterogeneity, with no calibration efforts, in a crowdsourced context.

Finally, Saccomanno et al. (2020) propose an LSTM based approach that considers a single ranking based fingerprint as input. They achieve a good accuracy performance, comparable with those observed for approaches based on full-fledged fingerprints, on three public datasets, showing that robustness to signal perturbations is also granted by the model to some extent.

Our work differs from all the previous ones in many respects. First, the main focus is on the attention contribution, above all from the interpretability perspective, but also evaluating its capability of improving the performance of positioning-related tasks. In addition, we rely on recurrent deep learning models applied to ranking based fingerprints, which have been considered in just a single work till now, and we improve with respect to it, considering both the theoretical aspects as well as the positioning accuracy.

2.2. Interpretability

Interpretability in machine and deep learning refers to the ability of an algorithm or a model to provide clear and understandable explanations for its predictions. There exist multiple approaches to achieve interpretability, tied to the type of machine learning model and task at hand (Murphy, 2023). *Inherently interpretable models*, such as decision trees, logistic regression, and linear models, are often easy to interpret since their internal workings are transparent, and can be easily understood in terms of their parameters and decision rules. However, they may not always capture the complexity of the data, thus their performance may be insufficient in some applications. *Semi-inherently interpretable models*, also referred to as *example based methods*, use examples as the basis for their interpretation (e.g., K-NN). *Joint training interpretability* techniques enhance models with interpretability features, such as attention mechanisms or attribute importance scores. The latter include gradient-based methods, like integrated gradients and gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017). Finally, *post-hoc techniques* focus on explaining the decisions of models without modifying their internal workings. Notable approaches in this class are model-agnostic methods, such as LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, Singh, & Guestrin, 2016) and SHAP (Shapley Additive Explanations) (Lundberg & Lee, 2017).

Examples of applications of interpretability techniques are: in the detection of heart diseases (Wang, Tian et al., 2021) or fraudulent bank transactions (Psychoula et al., 2021), where SHAP can determine the features' contributions to the final output; scenarios such as medical imaging or object recognition, where Grad-CAM can highlight the regions of an image that are most important in making a prediction (Panwar et al., 2020); and, natural language processing tasks like sentiment analysis or text classification, where LIME can shed light on the decisions of black-box models (Mathews, 2019).

Overall, interpretability techniques are critical in ensuring the accountability and trustworthiness of machine learning models, particularly in high-stakes domains such as healthcare and finance. Achieving full interpretability remains a challenging task, and a trade-off between accuracy and interpretability is often required. Therefore, the development of transparent and interpretable machine learning models that can provide reliable and trustworthy explanations is an active area of research.

3. Interpretability in fingerprinting

As previously mentioned, according to Miller (2019), interpretability is the degree to which an observer can understand the cause of a decision. It is not about figuring out everything about a model, but it can be considered as a means to an end, which implies that the form taken by an interpretation depends on the needs of the specific application (Murphy, 2023). Our goal is to enhance the positioning process by gaining novel scientific knowledge and operational insights, without worsening the performance concerning the position estimation. In the context of WiFi fingerprinting, we hereby define the concept of local interpretability (i.e., an interpretation related to a specific prediction) as the relevance that each access point has to a given position estimate provided by the model. We name this as the *relevance pattern* for a (ranked) fingerprint.

In the past, several works either assumed or pointed out the very important role played by the most strongly (in terms of RSS) perceived AP. Indeed, this is one of the motivations that led to the development of the ranking based fingerprint representation. We believe, instead, the strongest AP(s) not being necessarily the most relevant for the positioning task. For instance, a set of very powerful APs might be detected more or less in the same way at several different places; in that case, the discriminative role could be played by some other less powerful APs, that are only seen at specific locations. Therefore, determining which APs are mostly used by the model to derive a given prediction is a natural way to interpret its behaviour, as well as to highlight some characteristics related to the considered scenario.

In fact, other than local interpretability, it is worth asking ourselves whether a concept of global interpretability can also be defined, i.e., a general insight into the model behaviour based on a set of input data. Let us consider a set of predictions (i.e., location estimates) for which the relevance of the associated access points is known, and let us assume to group them together based on the similarity of their relevance patterns. We can derive such a global insight by studying whether the average relevance pattern associated to each different group is capable to (uniquely) characterize a delimited spatial area. In the remainder of the work, we are going to name such a property as *spatial characterization*, and we believe it would be of actual use in several positioning-related tasks (as we will see in Section 6.4).

To conclude the section, we now discuss how to obtain, in practice, a measure of relevance for the access points, which considers both the fingerprint representation and the chosen positioning algorithm. A possible approach to obtain interpretability is to train it jointly with the model (Murphy, 2023). In neural networks, the attention mechanism can be used to dynamically highlight relevant features of the input data (Galassi, Lippi, & Torroni, 2020; Mohankumar et al., 2020). At a high level, it works by assigning a weight to each input element, based on its relevance to the current task. These weights are computed using a learned function that takes into account the current state of the model and the input data. The weighted elements are then combined to produce a context vector that represents the most relevant information for the task at hand. Thus, in principle, relying on a deep learning model that can be easily extended with the attention mechanism (such as the one proposed by Saccomanno et al. (2020)) should allow us to reach our interpretability goal.

However, whether or not attention and its weights can be considered as a form of model interpretability is still an open debate, for instance, in the NLP community (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019). It follows that blindly applying attention is not enough to ensure interpretability. Thus, in the remainder of this work we show how and why in our case it provides plausible interpretations related to the positioning domain. Specifically, *(i)* attention is indeed capable to highlight relevant access points associated to a prediction; *(ii)* attention has a strong spatial characterization; *(iii)* different attention types have different behaviours and capabilities (i.e., not all formulations are equally good); and *(iv)*, attention based interpretations can contribute to downstream positioning tasks.

A summary of the overall experimental workflow (detailed in Section 5) is depicted in Fig. 1.

4. Fingerprinting with deep learning and attention

We start this section by formalizing the fingerprint representation employed in the work, discussing its main advantages and limitations. A description of the considered deep learning model then follows. Finally, we detail the attention mechanism and how to integrate in the model and, thus, in indoor positioning. A graphical overview of the overall framework and how information flows throughout it is reported in Fig. 2.

4.1. Ranking based fingerprinting

Ranking based fingerprinting is an effective approach to indoor positioning that has been first introduced by Cheng et al. (2005) and Machaj et al. (2011). It is based on the idea of transforming traditional fingerprints into a different representation. Specifically, each fingerprint is encoded by a vector of AP identifiers such that their position in the vector is determined by the RSS value of the corresponding AP, sorted from the strongest to the weakest signal (see Fig. 3). APs that

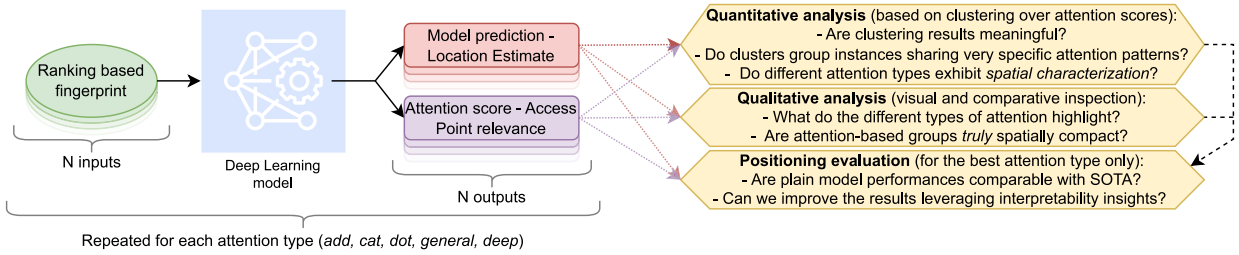


Fig. 1. Graphical high-level summary of the experimental evaluation pipeline.

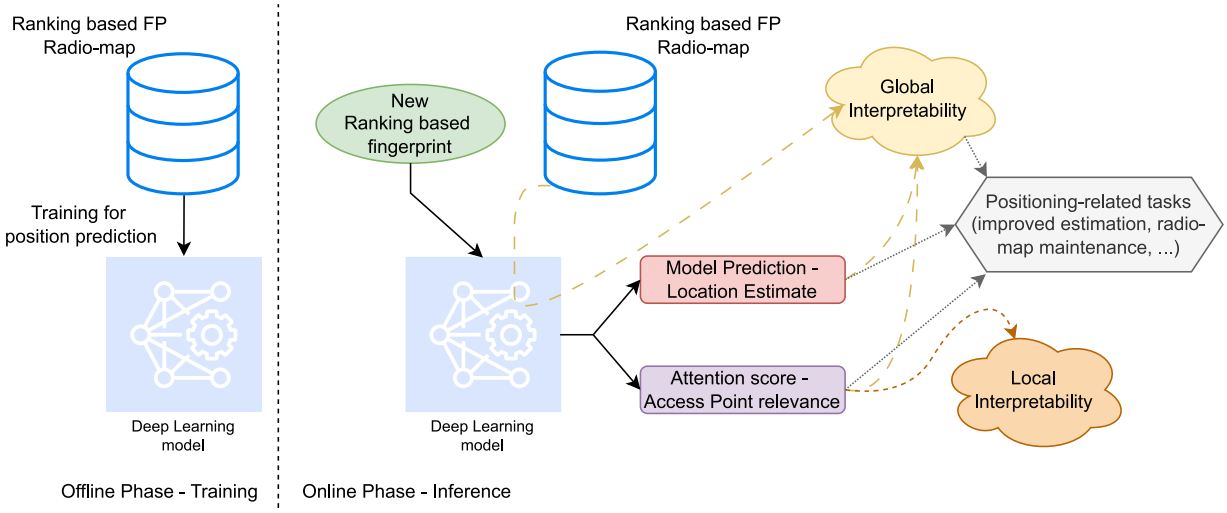


Fig. 2. Graphical representation of the elements composing the framework, their integration, and how information flows throughout it (i.e., from the radio-map and model training, to the generation of the interpretability outcomes, the positioning estimate, and the combination of such aspects for multiple tasks).

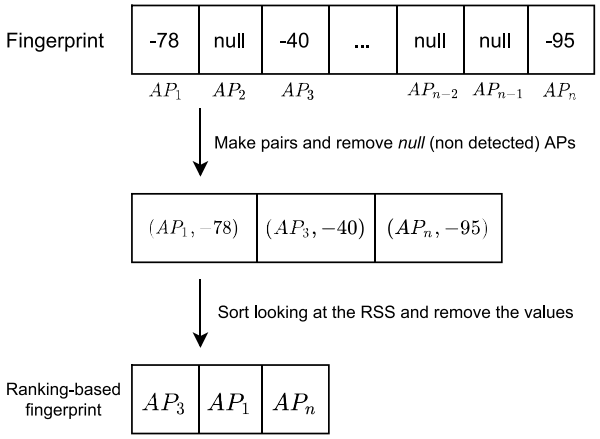


Fig. 3. Ranking based fingerprint representation construction process.

are not detected in a fingerprint are not present in its ranking based representation.

Formally, let $\mathcal{L} = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_o\}$ be the set of o discrete locations at which the desired scenario has been sampled. Each location $\mathbf{l}_i$, $1 \leq i \leq o$, is represented as a tuple that encodes position-related information, i.e., $(building_i, floor_i, room_i, x_i, y_i) \in \mathcal{B} \times \mathcal{F} \times \mathcal{R} \times \mathbb{R} \times \mathbb{R}$, with $\mathcal{B}$, $\mathcal{F}$, and $\mathcal{R}$ being respectively the sets that include categorical labels regarding the building, the floor, and the room/area where the locations are sampled, while x_i and y_i are the latitude and longitude data. The kth fingerprint associated with the location $\mathbf{l}_i$ is denoted as the vector $\mathbf{f}_{i_k} = [f_{i_k 1}, f_{i_k 2}, \ldots, f_{i_k q}] \in \mathbb{R}^q$, where q is the total number of APs appearing in the considered scenario (sensed all over the locations in $\mathcal{L}$) and $f_{i_k j}$, $1 \leq j \leq q$, is the RSS value related to the AP that is given the unique identifier j (or *null*, if such AP is not detected).

A raking based fingerprint can now be defined as an (ordered) vector $\mathbf{f}^{\mathbf{r}}_{i_k} = [a_1, \ldots, a_z]$ s.t. $f_{i_k a_h} \geq f_{i_k a_{h+1}}$, $1 \leq h < z$, where a_h is an AP identifier, and z, $1 \leq z \leq q$ (although generally $z \ll q$), is the number of detected APs (i.e., with RSS different from *null*) in fingerprint $\mathbf{f}_{i_k}$ (equivalently, $\mathbf{f}^{\mathbf{r}}_{i_k}$) at the location $\mathbf{l}_i$.

There are many advantages related to this design. First, the ranked representation is far more compact than the full-fledged fingerprint one, possibly reducing computational, storage and networking costs. In addition, ranked fingerprints are more robust to signal perturbations related to the heterogeneity of the devices: while two different devices might observe different RSSs from the same APs (e.g., due to diversities in their hardware), the corresponding ranked fingerprints will be much similar to each other, e.g., thanks to the fact that rankings are invariant to bias and scaling (Lohan et al., 2017; Machaj et al., 2011). Finally, classical fingerprints are sparse. At each location only a small subset of APs is visible with respect to the total q, which is often a large number, possibly leading to algorithmic level issues caused by the curse of dimensionality (Aggarwal, Hinneburg, & Keim, 2001).

On the negative side, the informative content associated with ranked fingerprints is reduced. Indeed, we shift from a rich continuous representation of the RSS values to far simpler sequences of discrete/categorical identifiers. While in the majority of previous work this caused a degradation of the positioning performance (Cheng et al., 2005; Laoudias et al., 2013; Lohan et al., 2017; Tiku & Pasricha, 2019), we hereby show that an approach based on this technique can be as accurate as those based on classical fingerprints.

4.2. Sequence-to-sequence modelling motivation

Ranking based fingerprints can be analysed by means of deep learning approaches that manage sequential data, such as RNN, LSTM, and

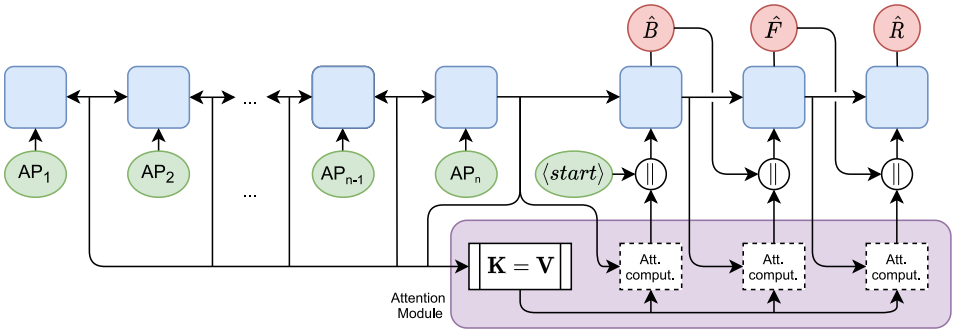


Fig. 4. Sequence-to-sequence LSTM model with attention for hierarchical position estimation.

Transformer architectures. Specifically, we are interested in sequence-to-sequence (seq2seq) models, which turn a sequence into another one. Considering them from a probabilistic perspective, it is possible to observe that (*i*) they perfectly fit in the probabilistic based indoor positioning paradigm, and (*ii*) the decoder can model and leverage the hierarchical structure of positions (e.g., buildings containing floors containing rooms) through its autoregressive nature.

Given an input sequence $\mathbf{x} = [x_1, \dots, x_n]$ and a target sequence $\mathbf{y} = [y_1, \dots y_m]$ we can define a probabilistic sequential model as follows:

$$P(\mathbf{y}|\mathbf{x}) = P(\mathbf{y}_{1:m}|\mathbf{x}_{1:n}) = \prod_{i=1}^{m} P(y_i|\mathbf{y}_{1:i-1}, \mathbf{x}_{1:n}). \tag{1}$$

Considering a position defined as, for instance, the triplet $(building, floor, room)$, and $\mathbf{x}$ as a (ranking based) fingerprint, it follows from Eq. (1) that the model adheres to the indoor positioning probabilistic framework:

$$\begin{aligned}(\widehat{building}, \widehat{floor}, \widehat{room}) = \hat{\mathbf{y}} &= \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \\ &= \arg\max_{\substack{building \in \mathcal{B}, \\ floor \in \mathcal{F} \\ room \in \mathcal{R}}} P(building, floor, room|\mathbf{x}) \,,\end{aligned} \tag{2}$$

where $\mathcal{B}$, $\mathcal{F}$ and $\mathcal{R}$ are respectively the set of all possible categorical labels for the buildings, the floors, and the rooms. Applying the chain rule of probability to Eq. (2), we obtain:

$$\begin{aligned}P(building, floor, room|\mathbf{x}) = \; & P(room|floor, building, \mathbf{x}) \\ & \cdot P(floor|building, \mathbf{x}) \\ & \cdot P(building|\mathbf{x}).\end{aligned} \tag{3}$$

Thus, the predictions for the higher hierarchical levels $\mathbf{y}_{1:i-1}$ (e.g., building and floor) are explicitly taken into account by the model to predict the lower levels $\mathbf{y}_{i:m}$ (e.g., room) in a natural and generalizable way. This is in contrast with the majority of previous indoor positioning approaches which used to handle the inherent hierarchical relationships by a set of specialized models arranged in a cascade fashion.

4.3. The developed model

In this work, we consider a sequence-to-sequence unidirectional LSTM; the choice of neglecting bidirectionality in the encoder was based on an empirically observed overfitting phenomenon. To allow for the analysis of how attention contributes to the positioning task, the model is equipped with a Bandahu-style attention (Bahdanau et al., 2015) module. The set of keys/values is generated by the encoder hidden states, while the query employed at each decoder step is given by the output at the previous step. Fig. 4 depicts the overall architecture. As can be seen, and as done in NLP, to exploit the autoregressive behaviour of the decoder an initial input token $\langle START \rangle$ has to be provided. Teacher forcing (Williams & Zipser, 1989) has been employed at training time to always provide the correct prior information during the sequential decoding phase.

Without delving too much into details, let us now understand how an attention-enhanced sequence-to-sequence model differs from a vanilla one, like the architecture used by Saccomanno et al. (2020), which can be summarized as in Fig. 5. To do that, we refer to the probabilistic framework described in Eq. (1).

We shall see that the main difference consists of how the context vector $\mathbf{c}$ is constructed and carried on during the computation, allowing the model to interact with the input information in different manners. The encoder component of both models can be represented as a function f_{enc} defined as:

$$f_{enc} : \mathbf{x}_{1:n} \rightarrow [\mathbf{h}_1, \dots, \mathbf{h}_n], \text{ with } \mathbf{c}_0 = \mathbf{h}_n \,, \tag{4}$$

i.e., a function that produces a sequence of new partial input representations $\mathbf{h}_i$ (referred to as encoder hidden states), the latter of which is a new (squeezed) representation (fixed-size vector) of the overall input sequence, also referred to as the (initial) context vector $\mathbf{c}_0$. It follows that the decoder, in a probabilistic fashion, can be defined as:

$$P_{dec}(\mathbf{y}_{1:m}|\mathbf{c}_0) = \prod_{i=1}^{m} P_{dec}(y_i|\mathbf{y}_{0:i-1}, \mathbf{c}_0) \,, \tag{5}$$

where y_0 is the $\langle START \rangle$ token. Each decoder step i, considering both the previous output and the previous context vector, is thus modelled as follows by the recurrent architecture:

$$\begin{aligned}P_{dec}(y_i|\mathbf{y}_{0:i-1}, \mathbf{c}_0) = P_{dec}(y_i|\mathbf{s}_i) = \text{softmax}(\mathbf{s}_i), \\ \text{with } (\mathbf{s}_i, \mathbf{c}_i) = f_{dec}(y_{i-1}, \mathbf{c}_{i-1}).\end{aligned} \tag{6}$$

While in the vanilla recurrent model the decoder hidden state $\mathbf{s}_i$ matches the context vector $\mathbf{c}_i$, in the attentive model $\mathbf{c}_i$ is computed by the attention mechanism, by explicitly and jointly considering all $\mathbf{x}_{1:n}$ and $\mathbf{s}_i$. This implies that, while the attention model can directly access and leverage the input information (more details later), the vanilla one only can do that through the condensed array $\mathbf{c}_0$ obtained as output from the encoder, which is further modified at each step of the decoder[2].

Given that the considered positioning problem is framed as a multi-class classification one, and considering L_i as the negative log-likelihood loss applied to each output element, $1 \leq i \leq |\mathbf{y}|$, the overall loss function L is defined as follows:

$$L(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{y}|} w_i \cdot L_i(\hat{y}_i, y_i) \,, \tag{7}$$

[2] What we have just explained is a *greedy search* approach, that at each time step considers the element with the highest conditional probability. *Beam search* is a more sophisticated approach, in which at each time step the k elements with the highest conditional probabilities are considered. At each subsequent time step, based on the k candidates at the previous step, it selects other k output elements with the highest conditional probabilities, intuitively generating a tree of possible output sequences (Zhang, Lipton, Li, & Smola, 2021). In our work, we relied on beam search with $k = 5$.

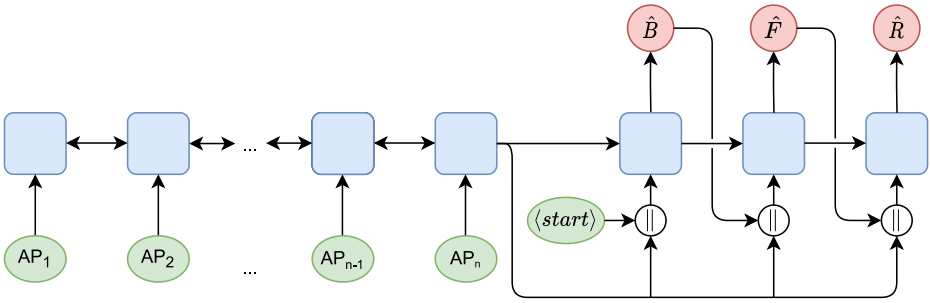


Fig. 5. Vanilla sequence-to-sequence LSTM model for hierarchical position estimation.

where w_i is an optional weight of the loss related to the hierarchical level i.

As a final remark, it is worth pointing out the reasons behind the choice of considering only recurrent models rather than the far more (nowadays) popular Transformer (Pavlopoulos, Malakasiotis, & Androutsopoulos, 2017) architectures. First, based on preliminary analysis, the performance of LSTM models was observed to be superior in our task with respect to Transformer, the latter perhaps being hindered by the relatively small quantity of training data at its disposal. The second point pertains to the interpretability of the overall approach. Indeed, Transformer makes full use of the attention mechanism, providing different representations at each encoder and decoder layer. Nevertheless, involving multiple layers would have made it more complex to obtain a representation both interpretable and exploitable for our purposes, even considering just the decoder. Conversely, the LSTM model provides just a single attention vector for each sequence.

4.4. Attention mechanism

In the case of sequences, like those considered in NLP and our scenario, the core principle behind attention is that of computing a weight distribution on the sequential input, assigning a higher score to those elements that are more relevant for the task at hand.

The attention mechanism can be defined as a weighted average of values that operates on three elements often referred to as keys (K), queries (Q), and values (V). This notation has been first proposed by Vaswani et al. (2017), but it is also possible to think of these elements from a database point of view: we want to determine how much the tuples (K) in a table are relevant (weak definition of matching) for a given input ($q \in Q$), returning as output a value (V).

Formally, let us define $\mathbf{K} \in \mathbb{R}^{d_k \times n_k}, \mathbf{Q} \in \mathbb{R}^{d_k \times n_q}$, and $\mathbf{V} \in \mathbb{R}^{d_k \times n_v}$ as matrices, each composed of a variable number of column vectors (respectively n_k, n_q, and n_v) having the same size (d_k). For our purposes, we will consider the query to be just a single column vector $\mathbf{q} \in \mathbb{R}^{d_k}$ rather than a matrix. The relevance of each column vector $\mathbf{k}_i \in \mathbf{K}$ with respect to $\mathbf{q}$ is evaluated by a compatibility function f, whose output is a vector $\mathbf{e} \in \mathbb{R}^{n_k}$ of energy scores:

$$\mathbf{e} = f(\mathbf{q}, \mathbf{K}). \tag{8}$$

Energy scores are then transformed to a vector $\mathbf{a} \in \mathbb{R}^{n_k}$ of attention weights, applying a distribution function. In our case, such a function will be the *softmax*, leading to the following definition:

$$\mathbf{a} = \mathrm{softmax}(\mathbf{e}) = \mathrm{softmax}([e_1, \ldots, e_{n_k}]), \tag{9}$$

$$\mathrm{softmax}(e_i) = \frac{\exp(e_i)}{\sum_{j=1}^{n_k} \exp(e_j)}. \tag{10}$$

Attention weights are further combined with matrix $\mathbf{V}$, to obtain a weighted representation of $\mathbf{V}$ itself. Note that for each element $\mathbf{k_i} \in \mathbf{K}$ there is a corresponding element $\mathbf{v_i} \in \mathbf{V}$ (i.e., $n_k = n_v$). The final outcome of the attention submodel is a context vector $\mathbf{c} \in \mathbb{R}^{d_k}$, that is typically employed by other components of the model where the attention architecture is integrated; it is defined as follows:

$$\mathbf{c} = \sum_{i=1}^{n_v} a_i \mathbf{v_i}. \tag{11}$$

Bringing together the sequence-to-sequence model and the attention mechanism, in order to compute the attention score vector for each output element t, $1 \leq t \leq m$, $\mathbf{q}$ must correspond to the previous hidden state of the decoder $\mathbf{s}_{t-1}$, $\mathbf{K}$ is the column based matrix obtained combining all the encoder hidden states $\mathbf{h}_i, 1 \leq i \leq n$, and $\mathbf{V} = \mathbf{K}$. Thus, considering our indoor positioning scenario, in which three hierarchical levels (building, floor, and room) are present, $t = 3$. As a result, for each ranking based fingerprint three attention score vectors are obtained, which can be seen as a matrix $\mathbf{A} \in \mathbb{R}^{3 \times n_k}$, where n_k will be set equal to the median value of the number of non-null APs seen among all (ranked) fingerprints (those ranked fingerprints with $z > n_k$ are simply truncated).

It is worth noticing that many different attention formulations have been proposed in the literature. Among them, we cannot rely on today's widely used self-attention mechanism (where $\mathbf{Q} = \mathbf{K} = \mathbf{V}$), since for our interpretability purposes we are interested in (cross) attending input and output elements. Nevertheless, many different compatibility functions can be used within the general cross-attention framework (i.e., the one employed by our model). We take into account the following ones:

$$dot = f(\mathbf{q}, \mathbf{K}) = \frac{\mathbf{q}^T \mathbf{K}}{\sqrt{d_k}} \tag{12}$$

$$general = f(\mathbf{q}, \mathbf{K}) = \mathbf{q}^T \mathbf{W} \mathbf{K} \tag{13}$$

$$cat = f(\mathbf{q}, \mathbf{K}) = \mathbf{w}^T \tanh(\mathbf{W}[\mathbf{K} \parallel \mathbf{q}]) \tag{14}$$

$$add = f(\mathbf{q}, \mathbf{K}) = \mathbf{w}^T (\mathbf{W}_1 \mathbf{K} + \mathbf{W}_2 \mathbf{q}) \tag{15}$$

$$deep = f(\mathbf{q}, \mathbf{K}) = \mathbf{w}^T \mathbf{E}^{(L-1)}, \tag{16}$$

$$\text{with} \begin{cases} \mathbf{E}^{(l)} = \mathrm{ReLU}(\mathbf{W}_l \mathbf{E}^{(l-1)}) & \text{if } 1 < l < L \\ \mathbf{E}^{(1)} = \mathrm{ReLU}(\mathbf{W}_1 \mathbf{K} + \mathbf{W}_0 \mathbf{q}) & \text{if } l = 1 \end{cases}$$

where $\mathbf{w}$ is a learnable vector and $\mathbf{W_{(i)}}$ are learnable matrices. The functions follow two main approaches. Attentions *dot* (Vaswani et al., 2017) and *general* (Luong, Pham, & Manning, 2015) are based on matching and comparing $\mathbf{K}$ and $\mathbf{q}$, with the main difference that *general*, thanks to the weight matrix $\mathbf{W}$, makes it possible to deal with queries and keys employing different representations. On the other hand, *cat* (Vaswani et al., 2017), *add* (Bahdanau et al., 2015), and *deep* (Pavlopoulos et al., 2017) follow a different strategy, and combine the keys and the query in a joint representation that is further weighted by an importance vector $\mathbf{w}$, which conveys the notion of relevance. This makes them a good choice when keys and queries are encoded in significantly different ways. They differ from each other in how they combine $\mathbf{K}$ and $\mathbf{q}$: *cat* concatenates them; *add* computes the contributions separately and then sums them; and, *deep* extends *add* using multiple layers allowing to build richer representations.

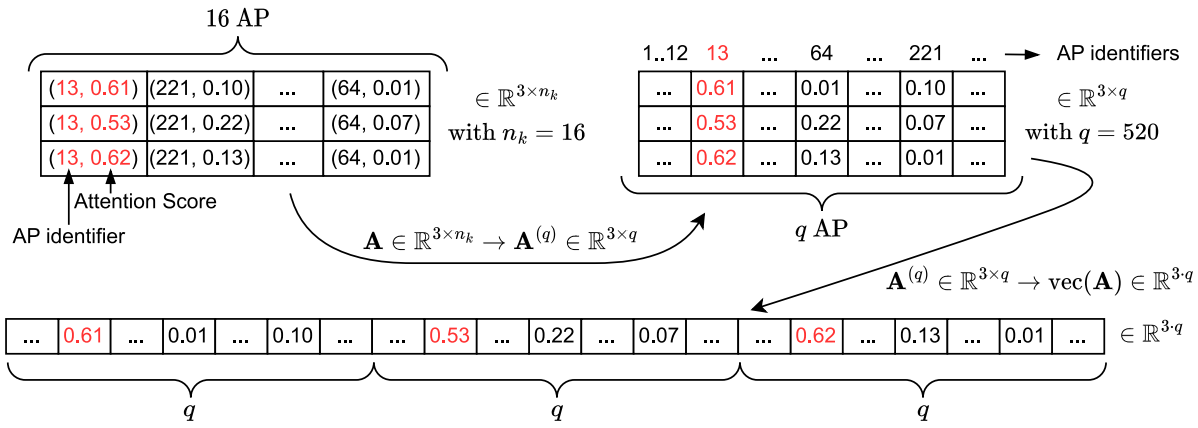


Fig. 6. Graphical account of the transformation of an attention matrix $\mathbf{A} \in \mathbb{R}^{3\times n_k}$ to its vectorized representation $\text{vec}(\mathbf{A}) \in \mathbb{R}^{3\cdot q}$.

In specific works in the deep learning literature (Galassi et al., 2020), attention has been shown to provide for each input sequence element (i.e., $x_i \in \mathbf{x}_{1:n}$ in Eq. (1), that will be the keys) a form of relevance to the prediction of each specific output sequence component (i.e., $y_j \in \mathbf{y}_{1:m}$ in Eq. (1), each of which will be a query). Nevertheless, there is still no unanimous consensus about its capability to actually fulfil such a task in a general setting, as witnessed for example by the many open debates within the NLP community (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019).

Therefore, it is worth studying whether attention, when applied to the previously described ranking based fingerprint representation, allows one to assess the relevance of the APs in the prediction of a given position (more precisely, of each hierarchical component describing a position), adhering to our concept of interpretability described in Section 3. Moreover, given that multiple types of attention (compatibility functions) exist, an investigation about their behaviours also becomes necessary. In the following section, thanks to a set of carefully designed qualitative and quantitative experiments, we will shed light on such matters.

5. Experimental evaluation

In this section, we first give an account of the considered indoor positioning dataset and the hyperparameter optimization process. Next, we provide a thorough description of the designed experiments and their results, aimed at evaluating the behaviour of the attention mechanism when applied to the indoor positioning context.

Recalling that Fig. 1 depicts the overall experimental workflow, we begin with an analysis of attention. We start with a quantitative assessment carried out by means of clustering, that will allow us to formally determine whether attention is capable of eliciting general spatial relationships in the data. Next, a qualitative analysis is performed, which will reveal some interesting attention patterns and will relate them spatially with the specific real-world premises considered. Finally, results are established regarding the prediction of the position of a given instance, leveraging the probabilities and the attention values provided by the model when taken separately as well as in conjunction.

In the remainder of the section, we will consider the following representation of the attention matrix $\mathbf{A} \in \mathbb{R}^{3\times n_k}$ described in Section 4. First, we transform each row into a vector of size q (i.e., the number of distinct access points), in such a way that the attention value associated with the AP j is mapped into the jth position of the vector. In the resulting matrix (of size $3 \times q$) all elements not associated with an attention value are set to zero. Finally, we consider a vectorized (i.e., flattened) representation of the latter matrix, defined as $\text{vec}(\mathbf{A}) \in \mathbb{R}^{3\cdot q}$. A summary of the process is reported in Fig. 6. This allows us to compare the attention values associated with different instances irrespective of their AP ranking.

5.1. Experimental setting

5.1.1. Dataset

Concerning the overall performance evaluation, for the sake of a fair and comprehensive comparison with previous literature, a publicly available and well-recognized dataset for indoor positioning has been taken into account.

UJIIndoorLoc (Torres-Sospedra et al., 2014) is likely the most well-known and exploited dataset for fingerprint based indoor positioning. It describes a multi-building and multi-floor setting spanning an area of 108 703 m^2, with the intent of mimicking the difficulties that could arise in the everyday usage of indoor positioning systems. It includes 19 938 training and 1111 test fingerprints. Test fingerprints have been collected four months later than training ones, also using different devices. The temporal aspect is crucial since it allows to test the performance of a model on a scenario affected by a variety of dynamics, such as the introduction or removal of APs compared to training time and other environmental changes.

Overall, 520 APs have been detected at multiple locations; around 50 of them are not identified during the training phase, becoming available only after the deployment. The median value of the number of non-null APs seen among all fingerprints is 16. Thus, following the rule of thumb described by Saccomanno et al. (2020) and mentioned also in Section 4.4, this will be the length of the ranked fingerprints considered. The soundness of such a choice is also confirmed by Fig. 7, where it can be seen that the number of visible access points and their RSS rapidly decreases with the rank length.

Each location is identified by four categorical variables: building, floor, room, and relative position (the latter two are typically combined together), as well as by latitude and longitude. Globally, 904 distinct locations (reference positions) have been considered for the fingerprint sampling stage (training). The sampling has been done according to a room based fashion and not by grid-partitioning the building structure, thus the distances between the training locations as well as the fingerprints' density coverages are rather variable.

Regarding the test set, room and relative position identifiers are not provided. This is a reasonable choice since, in the test phase, which aims to mimic the actual usage of the system, users may be located at an arbitrary position that may not match those used for the radio-map construction.

To reduce the Out-of-Distribution (OOD) generalization problem (Shen et al., 2021) caused by a divergence between the scenarios modelled by the training and test sets (both in terms of collecting devices and considered locations), we built a validation set so as to replicate as much as possible such a difference. Specifically, the latter is a subset of the training set instances collected only by the devices with id 7 and 10. This choice allowed us to obtain a (partially) different distribution, without removing any location from the ones observed in

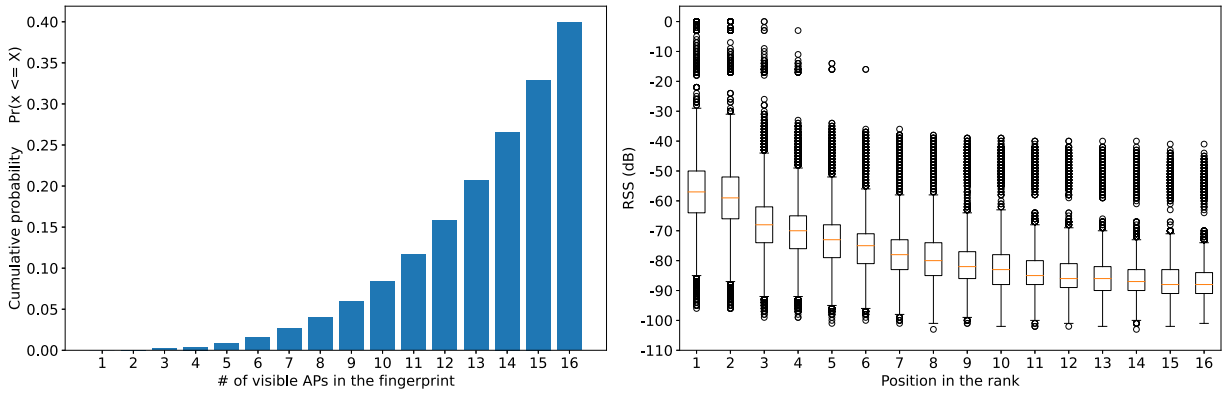


Fig. 7. Cumulative probability of the number of visible APs in a fingerprint (left) and RSS distribution over the rank positions (right).

Table 2
Characteristics of the considered UJIIndoorLoc dataset splits.

Split	# B	# F	FP × RP	# FP	# AP	# $\widetilde{AP}$	# RP	FPρ	Valid APs
Train	3	5	1 to 84	17 825	460	–	904	0.65 ± 0.38	17.7 ± 7.1
Valid.	2	4	2 to 37	2036	164	5	164	0.35 ± 0.22	21.7 ± 7.3
Test	3	5	–	1111	367	55	–	–	16.5 ± 6.9

B = Number of buildings; #F = Number of floors; FP × RP = fingerprints sampled per reference position; # FP = number of fingerprints; # AP = number of APs seen at least one time; # $\widetilde{AP}$ = number of seen APs not present in the training split; # RP = number of reference positions; FP_ρ = average FP density within a 5 m radius from each RP; Valid APs = average number of detected APs per fingerprint.

the training set (so as not to limit the spatial knowledge that can be learned by the algorithm). Table 2 summarizes the main features of the considered UJIIndoorLoc dataset splits. Observe that information regarding reference positions is not available for the test split.

5.1.2. Hyperparameters optimization

Model hyperparameters have been tuned relying on iterative grid search, based on the identified training and validation splits. We did not make use of embeddings for our input data, mapping the APs following a one-hot-encoding fashion. This determined the dimension of the encoder input (here, 520). We also observed that keeping the encoder hidden representation size similar to the dimension of the one-hot encoded input was rather effective. The other hyperparameters are as follows: we chose Stochastic Gradient Descent with warm Restarts (SGDR) (Loshchilov & Hutter, 2017) as the optimizer, with a learning rate of 0.05; batch size was set to 32; variational dropout was set to 0.5, while attention dropout was 0.2 (when applicable).

5.2. Quantitative analysis

In this section we present a procedure to evaluate which, among the considered attention compatibility functions, is the most capable of capturing spatial information. The overall idea is that instances sharing a similar attention pattern should also have in common some locality properties, e.g., they should be close to each other in the real-world space where they were collected, irrespective from their specific positions.

5.2.1. Experiments

To such an extent, we relied on a clustering analysis performed over training set attention vectors. Specifically, we focused on hierarchical clustering (Scikit-learn's `AgglomerativeClustering` (Pedregosa et al., 2011) method), as it allowed us to investigate compositional relationships between clusters through the associated dendrogram representation.

Hierarchical clustering requires the specification of two fundamental parameters, i.e., the metric used to compute the linkage, and the linkage criterion itself. As for the metric, which is calculated among vectors vec(**A**), we relied on *cosine* distance, as it is typically done in embedding-related tasks (Li et al., 2020). Turning to the second parameter, we evaluated the following linkage criteria: *average*, *complete*, *single*, *ward*. Based on a qualitative analysis of the generated dendrograms, we finally chose to rely on methodology *average*, since it showed the advantage of producing well-balanced splits among the clusters.

For each attention compatibility function, we then proceeded as follows. First, hierarchical clustering was run over the vectors vec(**A**) associated with training set instances, obtaining five dendrograms. Then, based on dendrogram analysis, and evaluating silhouette scores at different dendrogram cut points, we determined the most appropriate number of clusters for each function.

At this point, a question arises about whether the identified clusters are genuinely different from each other, i.e., they group together instances sharing a very specific attention pattern. To formally determine that, we devised a statistical procedure based on the Kolmogorov–Smirnov (KS) test (Massey, 1951). KS is a non-parametric test of the equality of continuous distributions, that can be used to quantify a distance between the empirical distribution functions of two samples.

The overall approach is as follows. Given two clusters, we construct the union set of their AP identifiers according to **A**[3]. Then, for each AP, within each cluster, we compute its empirical univariate distribution of attention values. Observe that an AP may be detected in just one cluster; in that case, we define a zero-valued dummy distribution for the other one. We now apply the KS test to the distribution pairs of each AP, to determine if they differ. The final cluster similarity is obtained by combining the single AP results by means of the Benjamini–Hochberg procedure for multiple comparisons (Benjamini & Hochberg, 1995).

The procedure is depicted in Algorithm 1. Given the input distributions pairs, the procedure evaluates each of them independently (Lines 1–16). For a pair X, Y, the KS test is evaluated, obtaining the corresponding KS statistic, i.e., the distance between the empirical distribution functions of two samples (Line 4). Then, a common continuous distribution, which will be used for null hypothesis testing, is

[3] This allows us to consider just the APs that are detected by instances belonging to the two clusters, leading to computational savings in the following steps.

Algorithm 1: Cluster similarity assessment

Input : $dist\_list = [(X_1, Y_1), \dots (X_n, Y_n)]$ list of distributions of pairs related to $cluster_1$ and $cluster_2$
Input : bs_size number of bootstrap samples
Output: `True` if clusters are different, `False` otherwise
Output: The ratio of statistically different distributions

1 $p\_values \leftarrow []$
2 **for** $i \in \{1, \dots, n\}$ **do**
3 $X, Y \leftarrow dist\_list[i]$
4 $KS\_res \leftarrow \text{KolmogorovSmirnovTest}(X, Y)$
5 $combined\_distribution \leftarrow X \| Y$
6 $KS\_boot \leftarrow []$
7 **for** $j \in \{1, \dots, bs\_size\}$ **do**
8 $S_1 \leftarrow \text{DrawSample}(combined\_distribution)$
9 $S_2 \leftarrow \text{DrawSample}(combined\_distribution)$
10 $S_1 \leftarrow S_1 + \mathcal{N}(0, \text{Cov}(S_1, S_2))$
11 $S_2 \leftarrow S_2 + \mathcal{N}(0, \text{Cov}(S_1, S_2))$
12 $KS\_boot[j] \leftarrow \text{KolmogorovSmirnovTest}(S_1, S_2)$
13 **end**
14 $KS\_boot[bsSize + 1] \leftarrow KS\_res$
15 $p\_values[i] \leftarrow \text{Mean}(KS\_boot \geq KS\_res)$
16 **end**
17 $reject\_null\_hypothesis \leftarrow \text{BenjaminiHochberg}(p\_values, 0.05)$
18 $diff\_clusts \leftarrow \text{CountTrue}(reject\_null\_hypothesis) == n$
19 $fract\_diff \leftarrow \text{CountTrue}(reject\_null\_hypothesis)/n$
20 **return** $diff\_clusts, fract\_diff$

generated by concatenating X and Y (Line 5). Observe that, for our purposes, the null hypothesis is that the batches of data are independent simple random samples taken from the common continuous distribution. At this point, as it is typically done when a measure of difference has to be estimated, we apply bootstrapping by resampling from the common distribution (Lines 7–13). For each bootstrap iteration, two random samples are independently extracted from the combined distribution, and random noise is applied to them (Lines 8–11). The KS statistic is then evaluated (Line 12). Then, by comparing bootstrapped and reference KS statistics, the p-value encoding the extent to which the KS distance between X and Y is statistically relevant is determined (Lines 14–15). Note that the p-value corresponds to the average of the boolean vector, interpreted as an integer one, generated by comparing the single KS bootstrap results with the reference one. Intuitively, if many KS bootstrap values are smaller than the reference one, then the two distributions are more likely to be different. At this point (Line 17), we need to combine the p-values related to the single APs in order to determine if the two input clusters are different. To such an extent, we rely on Benjamini–Hochberg procedure for multiple comparisons, computing the adjusted p-values based on a false discovery rate (FDR) level of 0.05, and returning *true* for those hypotheses that can be rejected for the given FDR level. If all null hypotheses can be rejected (Line 18), then the two clusters are deemed to be different. In addition (Line 19), we also compute the fraction of hypotheses that could be rejected by the method, in order to assess, when the clusters are considered to be alike, the ratio of non-similar APs.

Finally, it is also worth evaluating the clustering result from a spatial perspective. The idea is that clusters considered to be similar by the KS test should also be relatively close. Instead, statistically different clusters are intrinsically well-behaved, since they may possibly suggest different APs relevance (and, thus, strategies) followed by the neural network to derive the predictions even within small areas. To determine the distance between two clusters in the spatial domain, we relied on the Hausdorff metric. Informally, two clusters are deemed to be close by such a metric if every point of either cluster is close to some point of the other cluster. The Hausdorff distance is the longest distance one can be forced to travel by an adversary who chooses a point in one of the two clusters, from where he then must travel to the other cluster. In other words, it is the greatest of all the distances from a point in one cluster to the closest point in the other cluster. Formally, we define the Hausdorff distance between two spatial clusters $\mathcal{G}_1, \mathcal{G}_2 \subseteq \mathbb{R}^3$ as:

$$H(\mathcal{G}_1, \mathcal{G}_2) = \max_{\mathbf{g}_1 \in \mathcal{G}_1} \min_{\mathbf{g}_2 \in \mathcal{G}_2} \|\mathbf{g}_1 - \mathbf{g}_2\|_2. \tag{17}$$

Nevertheless, considering the maximum within the Hausdorff distance computation can lead to misleading results in the presence of spatial outliers that are likely to be generated during the clustering process. For this reason, as witnessed in literature (see, e.g., Isensee et al. (2019)), we calculate the 95th percentile of the minimum distances instead.

For each attention compatibility function, we evaluated the Hausdorff distance for all pairs of clusters considered to be similar by the previous KS test, as well as for all pairs of different ones. Intuitively, a well-performing compatibility function should lead to Hausdorff distances among similar clusters that are small, and overall smaller than those among dissimilar clusters (although, as already mentioned, it can be the case that some dissimilar clusters are actually close to each other).

5.2.2. Results

Fig. 8 shows the dendrograms obtained from the hierarchical clustering tasks for each attention compatibility function. As can be seen, despite relying on the same clustering approach, the results are quite heterogeneous[4]. Specifically, *dot* attention generates a very coarse clustering at the chosen cutting point; performing the cut at a lower level would result in a very large number of small clusters, while still not being able to break the largest (orange) group. Such a discrepancy suggests a bad behaviour of *dot* with respect to the clustering task. As for *general* and *deep* attentions, they share similar behaviour. Most of the clusters are aggregated at a very high point on the cluster distance axis, suggesting the presence of (too) many groups, typically quite different between them, and with no clear hierarchical structure. Finally, *add* and *cat* attentions show a more variegate situation as for cluster distances. Here, cluster compositionality can be clearly noticed and choosing different cutting points would always lead to balanced partitions, although of different sizes. Thus, they exhibit the best behaviour when it comes to capturing hierarchical relationships among clusters.

We now turn to study whether the identified clusters are genuinely different from each other, i.e., they group together instances sharing a very specific attention pattern. Figs. 9 and 10 report the result of the KS procedure applied to the previously discussed clusterings. Fig. 9 shows, for each pair of clusters, whether they are judged to be similar (dark colour) or not (bright colour) by the test. Intuitively, the ideal case is represented by an entirely bright picture, except for the diagonal (that compares each cluster with itself). Thus, the two best attention functions are *add* and *cat*, with a slight preference towards the latter[5]. To determine the extent of similarity between clusters judged to be similar by the procedure, let us consider Fig. 10. Here, a darker colour corresponds to pairs of clusters whose instances, i.e. attention patterns, highlight a larger number of access points with the same intensity. Intuitively, the optimal case is thus represented by a scenario in which the clusters judged to be similar are still characterized by a low fraction

[4] Note that the depicted cutting points have not been cherry-picked, but automatically selected through a silhouette based approach: given a dendrogram, a set of possible thresholds was considered, each leading to a different clustering. For each clustering, the *silhouette score* (Rousseeuw, 1987) was calculated, obtaining a list of values onto which an elbow criterion finally allowed us to determine the best threshold.

[5] Even though *dot* appears to share a similar behaviour, it has a poor results because of its severely unbalanced clustering.

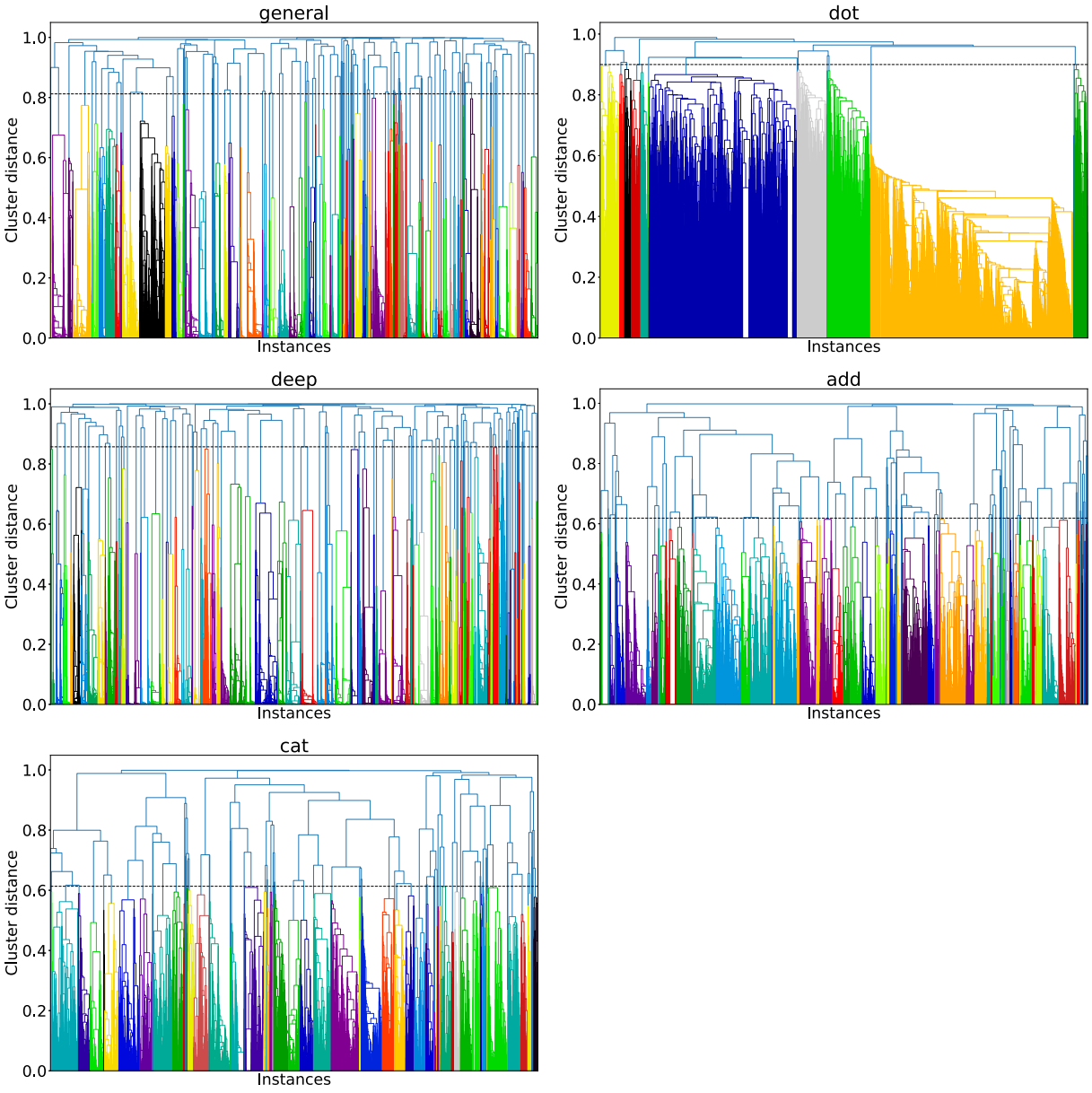


Fig. 8. Dendrograms obtained by clustering the attention vectors relying on different compatibility functions. The height of the tree branches indicates the distance between the clusters being hierarchically merged. The horizontal axis shows the individual data points being clustered. The black dashed line depicts the chosen cutting point, that leads to the specific sets of clusters identified by the colours. The result provided by of *dot* is unsatisfactory due to poor partitioning; *general* and *deep* produce too many groups; *add* and *cat* generate more balanced partitions where cluster compositionality can be clearly noticed.

of similar access points. This is precisely what happens with *add* and *cat*, confirming their good behaviour.

Finally, to assess the spatial characterization of the clusters, let us consider Fig. 11, which reports the distribution of Hausdorff distances among similar and dissimilar clusters (according to KS). Here, the ideal scenario is characterized by distance values for similar clusters typically lower than those for dissimilar ones. Of course, as already mentioned, it can be the case that dissimilar clusters still have a low Hausdorff distance, since two spatially close observations may rightfully consider different access points for the location prediction. Observe that, with the exception of *dot* attention, the median distance among similar clusters is always lower than that between dissimilar ones. Overall, *add* emerges as the best attention compatibility function, given the disjoint interquartile ranges and the low dispersion of distances calculated among similar clusters, leading us to choose it for the experiments that take into account also spatial information.

5.3. Qualitative analysis

Other than from a quantitative point of view, it is also worth investigating our findings from a visual perspective. Here, we also relate the attention patterns of the instances with the specific positions in the premises where they were collected.

5.3.1. Experiments

To such an extent, we first graphically compare the attention patterns exhibited by each compatibility function on a set of meaningful samples. Specifically, we want to assess, at different granularity levels (building, floor, and room), the relevance of the access points appearing in a fingerprint. This may allow us to investigate some of the phenomena mentioned in the previous sections, such as: does the deep learning model always consider the strongest access points to derive its output? Are the same access points exploited to perform the building, floor,

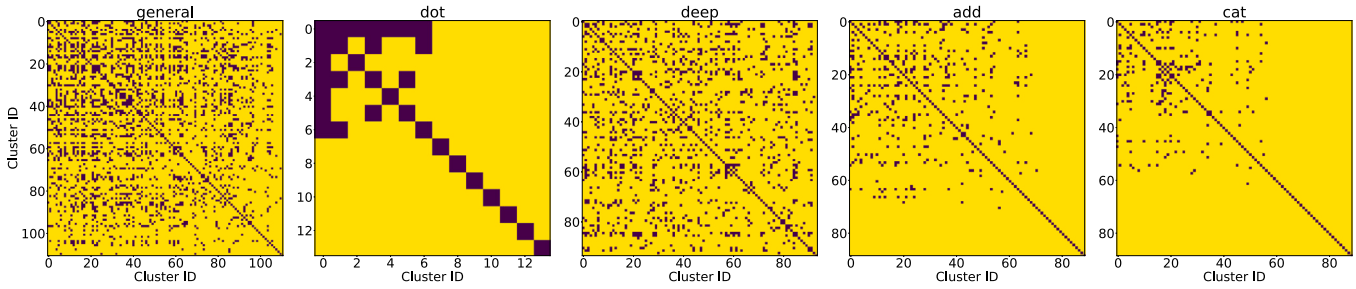


Fig. 9. KS based cluster similarity test result. Each cell denotes the comparison between a pair of clusters, which can be judged to be similar (dark colour) or not (bright colour). *add* and *cat* show the best behaviour, since they have a high number of clusters but only few of them are pairwise similar.

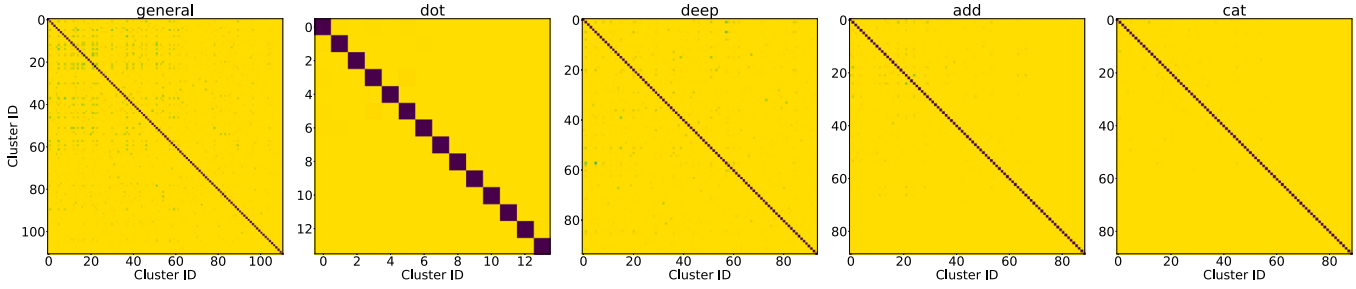


Fig. 10. KS based AP similarity ratio test result. Each cell denotes the comparison between a pair of clusters, which can share a large number of equally distributed APs (dark colour) or not (bright colour). *add* and *cat* show the best behaviour, since, in them, the clusters that were considered to be similar have a low number of equally distributed APs, a fact that still justifies their existence.

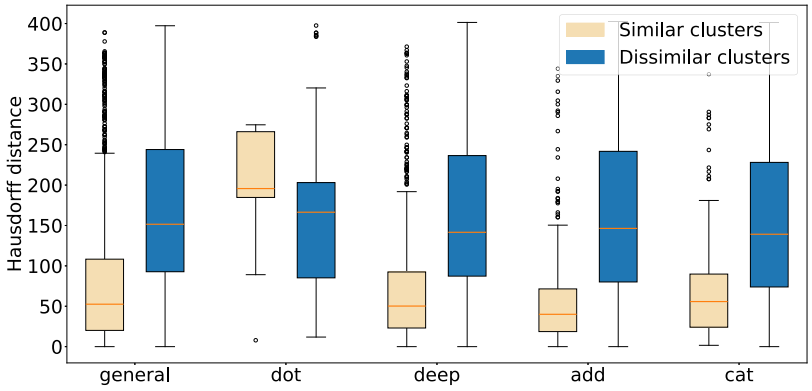


Fig. 11. Distributions of Hausdorff distances among similar and dissimilar clusters. Each box extends from the first to the third quartile values of the data, with a line at the median. Whiskers extend to the smallest and largest observations which are not outliers (considering 1.5 times the interquartile range). The optimal case is that of disjoint boxes, with distant medians, and distance values for similar clusters lower than those for the dissimilar ones. *dot* exhibits a wrong behaviour, while *add* emerges as the best.

and room predictions? Do the compatibility functions exhibit different attention patterns?

We finally evaluate the spatial arrangement of the attention based clusters. Intuitively, if attention values were to characterize specific areas of the indoor scenario, clusters should determine compact and well-separated regions in the considered premises.

5.3.2. Results

Fig. 12 shows the matrices **A** generated by different attentions (arranged into columns) for some representative instances (one for each row). We can immediately notice that the compatibility functions exhibit radical different behaviours. Interestingly, *general* and *deep* seem to focus just on the strongest access point(s), as opposed to *dot* that considers the last columns of the matrix. The latter is an incorrect behaviour since it also applies to the first case, where the last three columns refer to dummy APs (labelled with zero), that are used to pad the ranked fingerprint when less than n_k APs are detected. While these first attention functions appear to focus on specific columns rather than on the actual APs, this is not the case with *add* and *cat*, that exhibit rather heterogeneous patterns: typically, several access points are considered to be relevant; attention values may be distributed in different manners among them; and, the strongest access point is not always the most important. Finally, notice how, according to all attentions, the model seems to always exploit the same APs to derive the building, floor, and room predictions.

As a final qualitative assessment, let us now take into account also the spatial dimension. To do that, we focus on *add*, which emerged as the best attention compatibility function according to our quantitative analysis. Fig. 13 shows the assignments of training set locations to clusters. As can be noticed, the latter display evident spatial relationships, with instances belonging to the same cluster being placed relatively close to one another in the spatial domain. Observe that some dispersion or overlap phenomena can occur. This is the case, for instance, of cluster number 6 that, on floor 2, has its instances spread all over the central part of the map, interleaved with observations belonging to several other clusters. While dispersion may be attributed to an under-splitting of the clusters along the dendrogram, the spatial

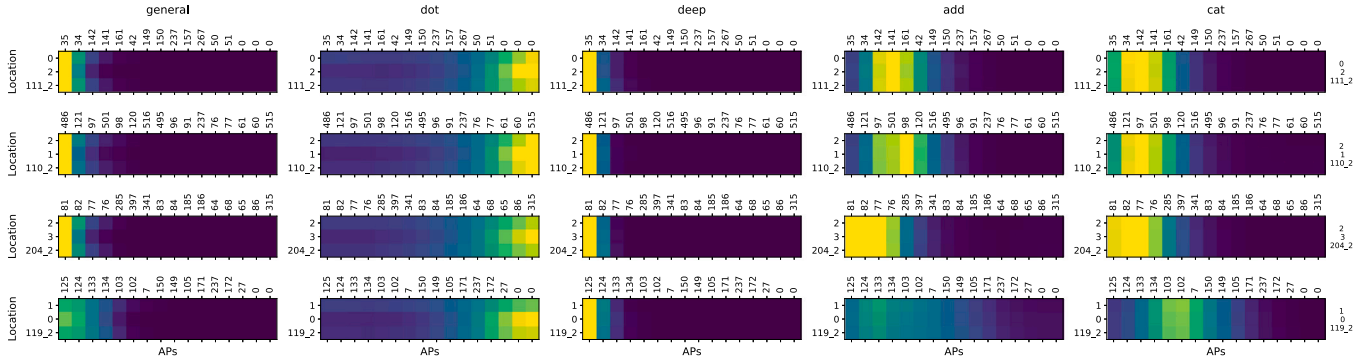


Fig. 12. Attention matrices **A** of four instances (one for each row) generated by different compatibility functions. A brighter colour denotes a higher relevance. The predicted location is reported on the left side of each matrix, while the ground truth is on the extreme right (for each matrix, the first row label is for the building, the second for the floor, and the third for the room hierarchical level). Access points identifiers are shown on the top of the matrices (0 = dummy AP). It is possible to observe that different types of attention focus on different parts of the ranked fingerprints.

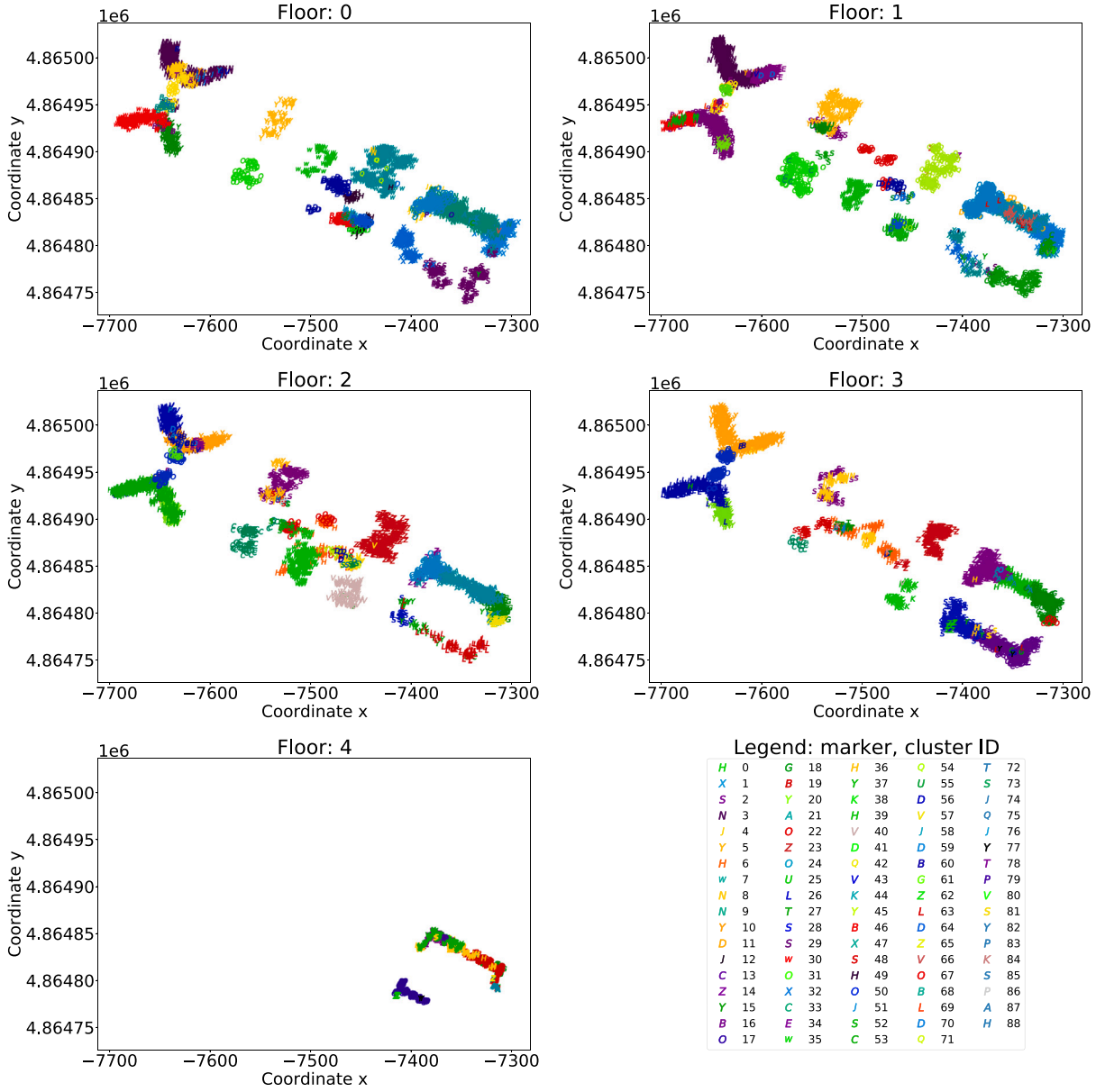


Fig. 13. Cluster assignments of instances on different floors for *add* attention compatibility function. It is possible to observe that clusters exhibit some spatial locality properties.

overlap between clusters may be due to either the presence of several, genuinely different attention patterns in the same areas, or to an over-splitting of the clusters. As a matter of fact, these noise phenomena are intrinsic to cluster based analysis, and do not hinder our conclusions.

5.4. Positioning performance evaluation

5.4.1. Experiments

As witnessed by the recent literature (Mendoza-Silva et al., 2020), there is still no unanimous consensus on how to evaluate the performance of an indoor positioning system. This stems from the fact that, in a multi-building and multi-floor scenario, the error of being located into a wrong building or floor is far more serious than that of being assigned to a wrong position within the correct floor. Finding a composite evaluation metric is not trivial. In this regard, observe how the 3D distance between the estimated and real position coordinates is not satisfactory. As a consequence, in this work, we rely on two different metrics, which should be jointly considered. First, the so-called *success rate* measures the fraction of instances for which both building and floor have been correctly predicted. Second, among such correctly predicted instances, the 2D positioning error is determined by looking at the Euclidean distance $E(\mathbf{p}, \hat{\mathbf{p}})$ between the predicted $\hat{\mathbf{p}} = (\hat{x}, \hat{y})$ and the ground truth $p = (x, y)$ coordinates:

$$E(\mathbf{p}, \hat{\mathbf{p}}) = \|\mathbf{p} - \hat{\mathbf{p}}\|_2 = \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2}. \tag{18}$$

Observe how the model developed in Section 4 allows us do determine a location triplet $(\widehat{building}, \widehat{floor}, \widehat{room})$ leading to the highest likelihood for a given instance. Then, the probability distribution $P(room | floor = \widehat{floor}, building = \widehat{building})$, can be also retrieved. In addition, thanks to the attention module, each hierarchical component has an associated attention weight vector which, intuitively, should represent the most relevant access points used by the model to derive the single level prediction.

While the success rate can be estimated comparing $(\widehat{building}, \widehat{floor})$ with the ground truth, for the 2D position estimation we proceed relying on a K-Nearest-Neighbour-like approach. Specifically, considering different similarity approaches to determine the neighbours, we obtain two strategies.

Probability based. The coordinates for an entry are determined by computing the weighted centroid over the coordinates of the K most likely rooms identified according to the probability distribution, where the weights ($\boldsymbol{\pi} \in \mathbb{R}^K$) correspond to the probabilities. The parameter K can be either considered as fixed, as in the work of Saccomanno et al. (2020), or tuned according to the best performance exhibited on a training+validation split of the training data.

Attention based. As a preprocessing step, we determine the average attention vector $\text{vec}(\mathbf{A}_\mu) \in \mathbb{R}^{3 \cdot q}$ for every training set location, element-wise aggregating the attention vectors obtained running the model over the training set instances collected at the location. Such a representation is then compared to the attention vector $\text{vec}(\mathbf{A}_{new}) \in \mathbb{R}^{3 \cdot q}$ generated by the model for a *new* instance, employing the *cosine similarity*. The obtained similarity attention scores ($\boldsymbol{\alpha} \in \mathbb{R}^K$) are then used within the K-NN framework following the same procedure as done for the probability based approach. Again, the parameter K can be fixed or tuned according to an evaluation performed on a training+validation split of the training data. The net result is that training set locations with an attention pattern similar to the given instance are assigned a higher weight during the centroid computation phase.

For comparison purposes, in the next section we evaluate the performances provided by the two approaches against some recent relevant papers in the area that apply their solution to UJIIndoorLoc. Bear in mind that several works did not use the test set as it is provided, but they restrict to some specific floor or building. This is a conceptually wrong approach that might lead to biased results, since achieving a good positioning performance in certain sub-regions of the data dataset is far harder than in other parts.

5.4.2. Results

We now turn to evaluate the positioning performance provided by the model described in Section 4.

Table 3 shows the test set results provided by the two K-NN approaches that employ probability- and attention based similarities. Also, it reports the respective K (the number of neighbours) determined by an optimization procedure based on the training and validation splits[6]. As can be seen, the achieved success rate is on par with previous works from the literature. Note that, being such a value related to the (categorical) output of the model, it is independent of the specific approach we considered for position estimation. Focusing on the positioning error in meters, probability emerges as the best variant, largely surpassing the attention based one as well as being on par with one of the best solutions available from the literature (Saccomanno et al., 2020) (although performing better than the latter from the success rate perspective).

5.4.3. Combining probability and attention

Position estimation can also be performed by considering weights obtained from a suitable combination of probabilities and attention values.

Despite the much worse performance exhibited by the attention based approach with respect to the probability based one, it may still be interesting to compare their behaviours. Fig. 14 shows the distribution of the differences between the errors provided by the two solutions. As can be seen, the histogram is roughly centred on zero, meaning that on a large portion of test set instances the two variants behave similarly, while the superiority of the probability based one is justified by the larger right tail. Nevertheless, the left tail informs us about the presence of a significant amount of cases in which attention provides the best performance. This suggests that the two approaches convey different, complementary information, which could be used jointly to devise a better positioning method.

Let us assume now to have at our disposal a strategy capable of correctly suggesting, for each instance, the best variant to rely on for the position estimation, between attention and probability. Such an *oracle powered* approach would achieve a test set average positioning error of 5.26 meters (see Table 3). The difference between the oracle and the single variants becomes even more evident in Fig. 15, which compares the empirical cumulative distribution functions (ECDFs) of the different approaches, confirming the extent of the potentially achievable improvement. Note that the oracle based selection of attention and probability does not necessarily represent the optimal solution: it can be the case that an even better performance could be achieved by suitably combining the two basic methodologies, for instance, by means of a mathematical formula. Providing an optimal positioning framework is out of scope of this paper, nevertheless, in the remaining part of the section we show that, even with a straightforward approach, it is possible to exploit the aforementioned insights and improve over the baselines.

Our methodology to combine probabilities with similarity attention scores is rather simple. Given the most likely location $(\widehat{building}, \widehat{floor}, \widehat{room})$ for an instance, we extract the set of all reference positions belonging to the same building and floor, i.e., $\hat{\mathcal{L}} = \{\mathbf{l}_i \,|\, building_i = \widehat{building} \wedge floor_i = \widehat{floor}\}$. Following the probability- and attention based approaches, we also obtain the vector $\boldsymbol{\pi} \in \mathbb{R}^{|\hat{\mathcal{L}}|}$ of probabilities and the vector $\boldsymbol{\alpha} \in \mathbb{R}^{|\hat{\mathcal{L}}|}$ of cosine similarity attention scores associated to the RPs. We now compute the vector

$$\boldsymbol{\zeta} = \frac{\boldsymbol{\alpha}^4 \odot \boldsymbol{\pi}}{\sum_{i=1}^{|\hat{\mathcal{L}}|} \alpha_i^4}, \tag{19}$$

[6] We calculated the mean positioning error on the validation set for every possible value of K and selected the parameter that produced the lowest error as the final choice. This process was repeated independently for both K-NN models, one using probability-based similarities and the other using attention-based similarities.

Table 3
Comparison of the positioning estimation results.

Approach	Success rate [%]	Positioning error [m]				
		Mean	Median	75th percentile	95th percentile	RMSE
Torres-Sospedra et al. (2014)	89.9	7.9	–	–	–	–
Song et al. (2019)	96.0	11.8	–	–	–	–
Saccomanno et al. (2020)	94.8	6.57	4.58	–	–	9.52
Laska and Blankenbach (2021)	92.6	9.07	6.32	–	–	–
Wang, Tiku and Pasricha (2021)	93.9	6.95	–	–	–	–
prob (K = 9)	95.23	6.56	4.78	8.97	18.7	9.53
att (K = 1)		9.82	7.07	14.3	28.7	13.9
oracle	95.23	**5.26**	**3.43**	**7.21**	**15.7**	**8.00**
prob ⊙ att		**6.40**	**4.58**	**8.82**	**18.2**	**9.23**

Note: The results of our best solution and those of the oracle based approach (see Section 5.4.3) are in bold.

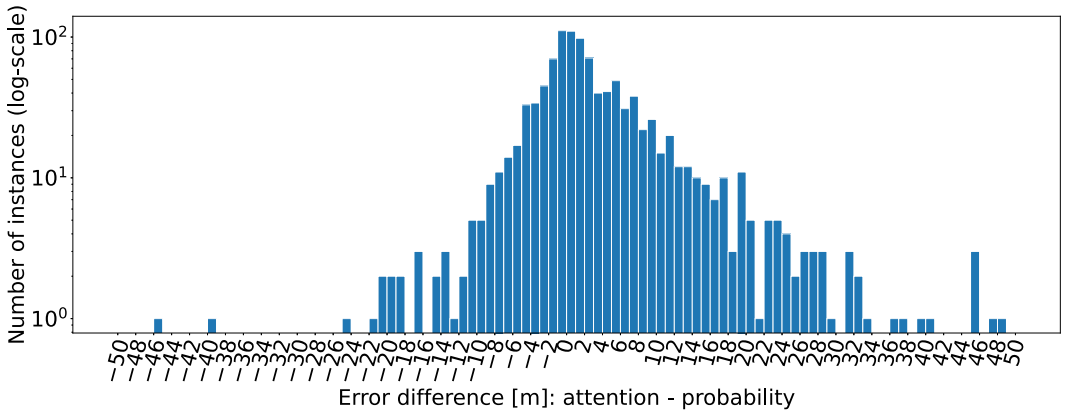


Fig. 14. Distribution of the differences between the errors of attention based and probability based approaches. The larger right tail shows the superiority of the probability based strategy, although it is possible to observe that for a significant amount of cases the attention based approach behaves better.

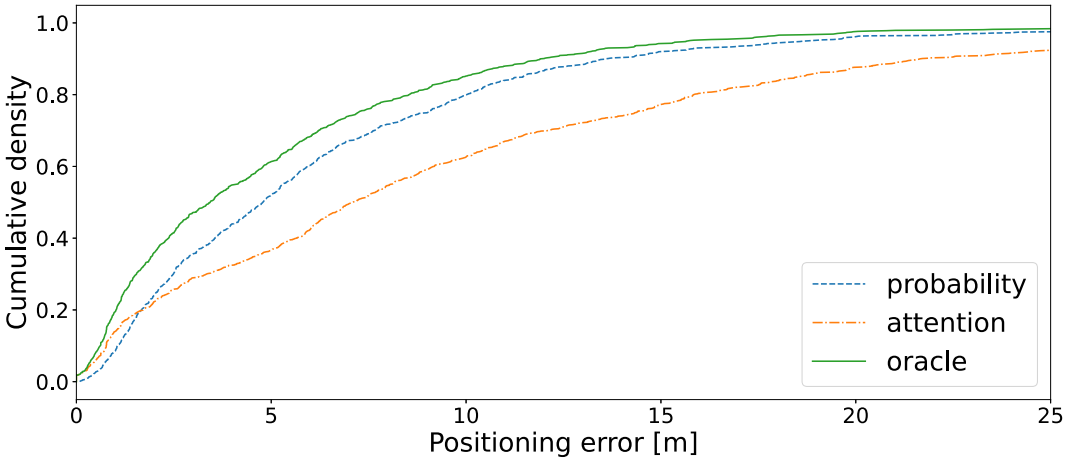


Fig. 15. ECDF showing the performances of probability and attention based approaches in comparison with the oracle. Being able to always choose the best approach between the two solutions would lead to a consistent boost of positioning performance.

which is used to calculate the final weighted centroid (⊙ is the Hadamard, i.e., element-wise, product).

The overall idea is to employ the attention based similarity scores as a filter for the probabilities. Intuitively, based on the previously discussed spatial characterization of attention patterns, the filtering aims to reduce the weights associated with regions that are not considered to be relevant by the attention mechanism. Eq. (19) indeed achieves this, since α^4 maps similarity attention scores to a (double) quadratic scale, and $\sum_{i=1}^{|\hat{\mathcal{L}}|}(\alpha_i)^4$ acts as a normalization factor. A clear advantage of the proposed solution is that it does not require any setting of K, since it considers all the locations belonging to a given floor. A graphical account of a typical scenario in which our method positively affects the result (error reduction from 18.7 to 4.07 m) is reported in Fig. 16. Notice how the area around the ground truth is characterized by higher similarity attention scores. This allows to strongly reduce the overall weight assigned to farther RPs that are associated with a relatively high probability, for instance, the one highlighted with a box. Results reported in Table 3 confirm the goodness of the approach, that achieves the lowest positioning error.

6. Discussion and conclusions

The main goal of this work was to understand whether a form of interpretability within the fingerprint based indoor positioning domain

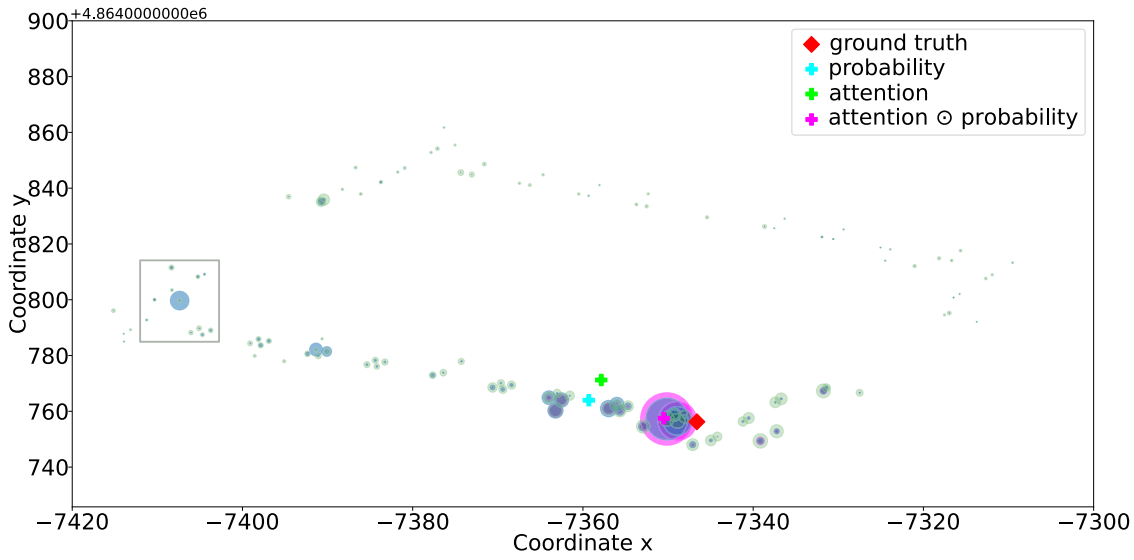


Fig. 16. A typical positioning scenario. RP weights (larger bubble = higher value/relevance) and position estimation results according to the different approaches are considered. The grey box on the left highlights a situation where probability emphasizes a RP poorly considered by the similarity attention score. As it can be seen, our simple approach to combine probability and attention values leads to a significant improvement.

could be defined. Specifically, we proposed a local notion of interpretation based on determining which access points are the most relevant to a given position estimate. Then, we extended it to a global perspective, conjecturing that access point relevance patterns may characterize specific areas of a building. In practice, such a measure of relevance was obtained by employing a sequence-to-sequence deep learning model equipped with an attention module, designed to operate on a ranking based representation of fingerprints.

In this section, we assess the work done under several perspectives. We start with a critical analysis of the results reported in Section 5. Then, we compare, from the interpretability standpoint, the RSS and ranking based fingerprinting, highlighting the superiority of the latter. Next, the main differences between our proposed framework and other approaches specifically designed to deal with RSS perturbations and AP selection are discussed. Afterwards, a series of applications that rely on our proposed notion of interpretability are presented. Finally, we outline directions for future research.

6.1. Critical analysis of the results

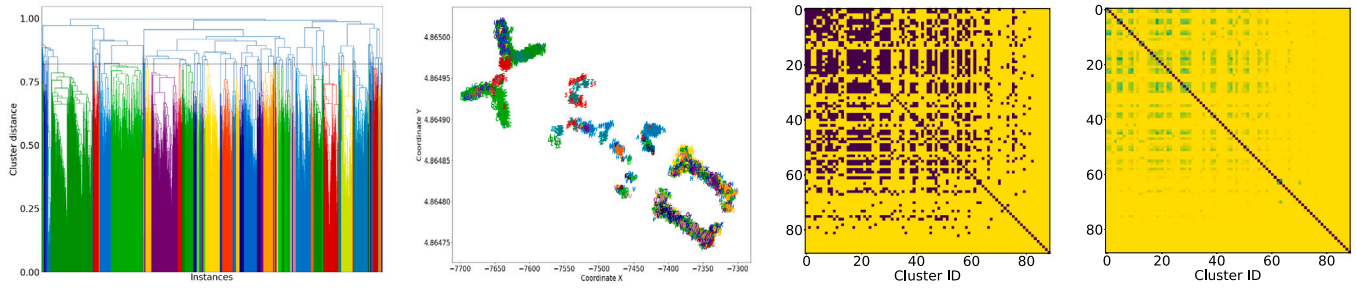
On the account of a series of thoroughly designed quantitative and qualitative experiments, that are summarized in Table 1, we showed that attention-derived patterns indeed characterize well-defined spatial regions, considering a large multi-building multi-floor well-recognized indoor positioning dataset. Nevertheless, our results also pointed out that not all the examined compatibility functions are equally capable of fulfilling such a role (*add* is clearly the best), highlighting the importance of choosing a suitable one. Notably, the spatial characterization of attention was observed even though our model exploits only categorical information about locations, neglecting any kind of proximity relationships among them. During such analyses it also emerged that deep learning models do not always exploit the strongest APs to perform a position estimation. This may suggest that the relevance of the APs depends, at least in part, on the model used. Thus, caution is advised in the design of fingerprint preprocessing and filtering strategies, as their validity could be model-dependent. In essence, attention indeed allowed us to get new scientific and operational insights about WiFi fingerprinting and deep learning.

Although our analysis confirms some very strong interpretability desiderata for our framework, i.e., plausible interpretations aligned with the domain (by means of a spatial characterization), one might argue: could the same results have been obtained just by considering the presence of specific access points in the ranked fingerprints, or do the attention values assigned to such access points indeed play a major role? To determine that, we shuffled within each ranked fingerprint the attention values assigned by *add* to the access points, repeating our quantitative and qualitative experiments. The rationale is that if the original attention values were stochastic, the outcomes should be similar to those for the shuffled case, suggesting that our interpretations are barely plausible and certainly not faithful nor meaningful. The dendrogram (Fig. 17(a)) begins with a coarse partitioning of the instances into three groups, roughly corresponding to the three buildings. This is not surprising, as in different buildings we can expect disjoint sets of access points to be present in the rankings. Nevertheless, at a finer granularity level, the partitioning is more chaotic, as confirmed by the inspection of the cluster assignments on Floor 3, where groups appear highly random (compare Fig. 17(b) with Fig. 13). Indeed, also the KS based similarity test shows some very different results (compare Figs. 17(c) and 17(d) with Figs. 9 and 10 respectively): shuffling the attention values we obtain substantially darker images, meaning that very large number of clusters are now considered to be similar by the KS procedure. Thus, although we remark that attention provides *a possible* interpretation and not *the* interpretation for the model behaviour, we provided evidence that in our case the former is meaningful and aligned with the domain, as the shuffling experiment led to fundamentally different results, showing, above all, no spatial characterization.

6.2. RSS vs. ranked fingerprinting for interpretability

In this work, for our interpretability framework we specifically focused on ranking based fingerprints and attention. Although in principle interpretability techniques can also be applied to full-fledged fingerprints, when pursuing such an approach there are some inherent difficulties that have to be taken into account.

To begin with, recurrent neural networks cannot be applied to RSS fingerprints, since they rely on a sequential inductive bias (Battaglia et al., 2018). Thus, other kinds of deep learning architectures should be exploited, like fully connected neural networks (FCNN) together with gradient based attribution methods (instead of attention) (Ancona, Ceolini, Öztireli, & Gross, 2019), to highlight the access points that are most relevant to generate a prediction. In any case, RSS fingerprints tend to be very long, having an element for each distinct access point present in the considered scenario (520 APs in the dataset UJIIndoorLoc). The result is that a large number of relevance values should be



(a) Dendrogram. (b) Cluster assignments of instances on Floor 3. (c) KS based cluster similarity. (d) KS based AP similarity ratio.

Fig. 17. Results of *add* compatibility function, shuffling the attention values within each fingerprint.

taken into account, hindering the *compactness* and *sparsity* requirements of the interpretation, as stated by Murphy (2023).

In addition, RSS fingerprints pose an issue from a practical point of view, as they are not transparent to the number of access points installed in the considered premises. Indeed, adding a new AP would require changing the architecture of and retraining the whole FCNN (increasing the dimension of the input layer by 1), which is not required with the recurrent approach as shown by Saccomanno et al. (2020).

Finally, note how, in our recurrent solution, the sequence-to-sequence modelling and the autoregressive nature of the decoder allow for a simple and effective extension to different indoor scenarios, both in terms of APs distribution and granularity of the hierarchical structure of positions. By relying on a FCNN, the hierarchical form of the prediction (building, floor, room), generated using a single model, would be lost, together with the adherence to the probabilistic framework described in Section 4.2. To recover it, two alternative directions can be followed. As a first solution, different models could be built for each distinct premise and hierarchical level. In the case of dataset UJIIndoorLoc, this translates to 1 model for the building prediction, 3 models for the floor prediction (one for each building), and around 12–15 models for the final room prediction (one for each floor of the different buildings). Alternatively, 3 models could be considered: 1 model for the building prediction taking in input the RSS fingerprint, 1 model from the floor prediction taking in input the RSS fingerprint and the predicted building, and 1 model for the room prediction taking in input the RSS fingerprint and the building and floor predictions. In both situations, the gradient based relevance values would have a completely different and more complex interpretation, as they would depend on more than one model and/or rely on heterogeneous features.

6.3. Other approaches for access point selection and RSS noise mitigation

Note that our proposed architecture allows us to ignore aspects that are often central to the development of positioning systems, such as those pertaining to the selection of APs (i.e., features), the normalization of their RSS, and the mitigation of RSS perturbation phenomena. Of course, in the literature other specific approaches are available to deal with such issues.

Considering AP selection, several solutions are available to determine the subset of APs that are more relevant for the prediction of the user location, i.e., those which are more likely to deliver a low localization error. Here, note how the strongest access points may not always provide the best positioning accuracy, as already shown in this work and by Chen, Yang, Yin, and Chai (2006). Among the selection approaches, either offline and online techniques can be found. The former selects a static subset of access points for the considered premises based on training data; such access points are then used for all downstream positioning tasks (Abed & Abdel-Qader, 2018; Laitinen & Lohan, 2015, 2016). The latter, given a new fingerprint observed at an unknown location, generates a dynamic subset of APs, based also on the fingerprint information, before performing the actual positioning (Cheng, Chou, & Chang, 2016; Kushki, Plataniotis, & Venetsanopoulos, 2007; Zou, Luo, Lu, Jiang, & Xie, 2015). Our solution is clearly different with respect to the offline approaches, since attention values are computed for every new prediction. In addition, as opposed to the other online methods, it allows us to seamlessly and jointly perform, within a single model, selection and prediction phases, without the risk of heuristically discarding useful information in the workflow. Specifically, the relevance values generated by the attention mechanism can be considered as fuzzy selectors for the access points. The fact that attention is applied over a limited number of stronger APs does not constitute an issue, as remarked in Section 5.1.1.

As for dealing with RSS perturbation, proposed techniques include hyperbolic location fingerprints (Kjærgaard, 2011), differential fingerprints based on signal strength difference (Hossain, Jin, Soh, & Van, 2011), and mean differential fingerprints (Laoudias, Kolios, & Panayiotou, 2014). Although these works are designed to cope with RSS perturbations caused by device heterogeneity, the same holds also for our framework. Indeed, while ranks are generated starting from RSS values, previous literature (Lohan et al., 2017; Ma et al., 2019; Machaj et al., 2011; Saccomanno et al., 2020) showed that ranked fingerprints are more stable than RSS fingerprints, as they are coarser. The downside is that ranks are also less informative; nevertheless, deep learning appears to be an effective manner to cope with this, as showed here and by Saccomanno et al. (2020). In addition, our solution has a much broader scope than the previously mentioned techniques, as it allows us also to develop a domain related notion of interpretability.

6.4. Applications

We believe our proposed *spatial characterization* of interpretability to be of actual use in several positioning-related tasks. To begin with, it could act as a guide for radio-map maintenance tasks. In the event of a known AP being faulty or removed, it becomes necessary to update the radio-map to reflect such changes. Instead of re-collecting fingerprints over a large area (note that an AP could be sensed even on multiple floors), relevance values could be exploited to limit the update operations to just the locations in which such AP was pertained to be useful by the model, leading to savings in terms of human time and computing resources. Of course, the usefulness of this application depends on the frequency of such AP replacement events.

As a second application, the access point relevance patterns could also help to identify unreliable predictions. One big issue in indoor positioning is that the environment is highly dynamic. Besides changes related to people crowding rooms and the intrinsic heterogeneity of devices, a major source of problems is represented by either the removal, the replacement, or the installation of APs. To avoid degradation of the positioning performance, the radio-map and the associated predictive

models need to be updated to reflect such changes. Nevertheless, determining when and where such updates are necessary is a difficult task, especially considering that it may not be known which of the APs have undergone changes. A possible solution could be that of determining what is the typical pattern associated with a specific location or region of the premises, and then measuring to which extent a newly given fingerprint, predicted to belong to that area, adheres to the shared pattern. If radical changes occur to the environment, they should be reflected in a divergence between the two relevance patterns, intuitively showing a change to the importance being assigned to the sensed access points. In turn, such a discrepancy could inform the positioning system regarding the possible unreliability of the prediction and, in that case, determine the region affected by the error, consequently suggesting possible corrective actions as suggested earlier in this section.

Finally, consider the general positioning procedure, followed by both the probabilistic and the deterministic approaches; it begins with determining the most probable locations as predicted by the model, leading to the retrieval of a set of points from the radio-map, for which latitude and longitude coordinates are known. Such coordinates are, in turn, exploited to derive the final position estimate. Spatial regions sharing the same access point relevance pattern could act as a constraint to reduce the number of candidates to consider when performing the radio-map lookup, contributing both to an improvement of the positioning performance, as well as to the overall speed of the process. To perform such a screening phase it may be sufficient to compare the relevance values of the given instance with the patterns characterizing the various areas, considering only the locations belonging to the most similar one. This may, for instance, complement or extend the simple position estimation approach that we presented in Section 5.4.3.

6.5. Current limitations and future work

The presented work still has some limitations. To begin with, the generalizability of our results is limited by deep learning in itself, given its strong dependence on training data and on the supervised paradigm. Despite that, the developed experimental workflow and its outcomes are based on a rigorous statistical procedure, which should allow to derive similar conclusions in other indoor scenarios, provided that a sufficient amount of training data is available. Further work is going to be done on the application of our technique to different datasets.

Also, note that our improvement in positioning performance, which is not the aim of the work but a way to show how interpretability can contribute in practice in indoor positioning, was based on an intuitive, although rather naive approach. More sophisticated techniques to jointly consider model likelihoods and attention scores should be investigated, however, preliminary analyses point out that such a task is not straightforward. The latter aspect is highly relevant since, as suggested by the oracle, the margin of refinement is considerably large.

Other future research includes the development of techniques to detect radio-map inconsistencies and the definition of an attention based metric to estimate the reliability of predictions.

Altogether, in this work we confirmed the observation made by Chaudhari, Mithal, Polatkan, and Ramanath (2021), namely, that the study of the relationships between attention weights and model interpretability is an active area of investigation that should be carefully considered by the research community.

CRediT authorship contribution statement

Andrea Brunello: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Angelo Montanari:** Writing – review & editing, Supervision, Funding acquisition. **Nicola Saccomanno:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nicola Saccomanno reports financial support was provided by u-blox AG. Angelo Montanari reports financial support was provided by u-blox AG. Andrea Brunello reports financial support was provided by Francesco Severi National Institute of Higher Mathematics National Group of Scientific Calculations.

Data availability

The data that has been used is already publicly available.

Acknowledgements

AM and NS acknowledge the support of *u-blox AG*, for the projects *Fingerprints and spatial knowledge in indoor positioning* and *Advanced solutions for indoor positioning*, respectively. The authors acknowledge the support of the Italian INdAM-GNCS project *Ragionamento Strategico e Sintesi Automatica di Sistemi Multi-Agente*.

References

Abed, A. K., & Abdel-Qader, I. (2018). Access point selection using particle swarm optimization in indoor positioning systems. In *Proc. 2018 IEEE conf. NAECON* (pp. 403–410). IEEE.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional spaces. In *Proc. 8th int. conf. ICDT, Vol. 1973* (pp. 420–434). Springer.

Akram, B. A., Akbar, A. H., & Shafiq, M. O. (2018). HybLoc: Hybrid indoor Wi-Fi localization using soft clustering-based random decision forest ensembles. *IEEE Access*, *6*, 38251–38272.

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2019). Gradient-based attribution methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 169–191.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proc. 3rd int. conf. ICLR* (p. 15). abs/1409.0473.

Bahl, P., & Padmanabhan, V. N. (2000). RADAR: an in-building RF-based user location and tracking system. In *Proc. 19th IEEE int. conf. INFOCOM* (pp. 775–784). IEEE Computer Society.

Bai, S., Luo, Y., Yan, M., & Wan, Q. (2021). Distance metric learning for radio fingerprinting localization. *Expert Systems with Applications*, *163*, Article 113747. http://dx.doi.org/10.1016/j.eswa.2020.113747.

Bai, S., Yan, M., Wan, Q., He, L., Wang, X., & Li, J. (2019). DL-RNN: An accurate indoor localization method via double RNNs. *IEEE Sensors Journal*, *20*(1), 286–295.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V. F., Malinowski, M., et al. (2018). Relational inductive biases, deep learning, and graph networks. CoRR abs/1806.01261. arXiv:1806.01261. URL: http://arxiv.org/abs/1806.01261.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Chaudhari, S., Mithal, V., Polatkan, G., & Ramanath, R. (2021). An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology*, *12*(5).

Chen, C., Wang, B., Lu, C. X., Trigoni, N., & Markham, A. (2020). A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. (p. 26). CoRR abs/2006.12567. arXiv:2006.12567. URL: https://arxiv.org/abs/2006.12567.

Chen, Y., Yang, Q., Yin, J., & Chai, X. (2006). Power-efficient access-point selection for indoor location estimation. *IEEE Transactions on Knowledge and Data Engineering*, *18*(7), 877–888.

Cheng, Y., Chawathe, Y., LaMarca, A., & Krumm, J. (2005). Accuracy characterization for metropolitan-scale Wi-Fi localization. In K. G. Shin, D. Kotz, & B. D. Noble (Eds.), *Proc. 3rd int. conf. MobiSys* (pp. 233–245). ACM, http://dx.doi.org/10.1145/1067170.1067195.

Cheng, Y.-K., Chou, H.-J., & Chang, R. Y. (2016). Machine-learning indoor localization with access point selection and signal strength reconstruction. In *Proc. 83rd IEEE conf. VTC Spring* (pp. 1–5). IEEE.

Feng, X., Nguyen, K. A., & Luo, Z. (2021). A survey of deep learning approaches for WiFi-based indoor positioning. *Journal of Information Technology*, 1–54.

Foliadis, A., García, M. H. C., Stirling-Gallacher, R. A., & Thomä, R. S. (2021). CSI-based localization with CNNs exploiting phase information. CoRR abs/2101.08983. arXiv:2101.08983. URL: https://arxiv.org/abs/2101.08983.

Galassi, A., Lippi, M., & Torroni, P. (2020). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(10), 1–18.

He, S., & Chan, S. G. (2016). Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons. *IEEE Communications Surveys and Tutorials*, *18*(1), 466–490.

Hernández, N., Parra, I., Sánchez, H. C., Izquierdo, R., Ballardini, A. L., Salinas, C., et al. (2021). WiFiNet: WiFi-based indoor localisation using CNNs. *Expert Systems with Applications*, *177*, Article 114906. http://dx.doi.org/10.1016/j.eswa.2021.114906.

Hoang, M. T., Yuen, B., Dong, X., Lu, T., Westendorp, R., & Reddy, K. (2019). Recurrent neural networks for accurate RSSI indoor localization. *IEEE Internet of Things Journal*, *6*(6), 10639–10651.

Hossain, A. M., Jin, Y., Soh, W.-S., & Van, H. N. (2011). SSD: A robust RF location fingerprint addressing mobile devices' heterogeneity. *IEEE Transactions on Mobile Computing*, *12*(1), 65–77.

Hsieh, H., Prakosa, S. W., & Leu, J. (2018). Towards the implementation of recurrent neural network schemes for WiFi fingerprint-based indoor positioning. In *Proc. 88th IEEE VTC* (pp. 1–5). IEEE.

Ibrahim, M., Torki, M., & ElNainay, M. (2018). CNN based indoor localization using RSS time-series. In *Proc. 23rd IEEE symp. ISCC* (pp. 1044–1049). IEEE.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., et al. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, *40*(17), 4952–4964.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proc. 2019 conf. NAACL-HLT* (pp. 3543–3556). ACL.

Khalajmehrabadi, A., Gatsis, N., & Akopian, D. (2017). Modern WLAN fingerprinting indoor positioning methods and deployment challenges. *IEEE Communications Surveys and Tutorials*, *19*(3), 1974–2002.

Kim, K. S., Lee, S., & Huang, K. (2017). A scalable deep neural network architecture for multi-building and multi-floor indoor localization based on Wi-Fi fingerprinting. (p. 9). CoRR abs/1712.01990.

Kjærgaard, M. B. (2011). Indoor location fingerprinting with heterogeneous clients. *Pervasive and Mobile Computing*, *7*(1), 31–43.

Kushki, A., Plataniotis, K. N., & Venetsanopoulos, A. N. (2007). Kernel-based positioning in Wireless Local Area networks. *IEEE Transactions on Mobile Computing*, *6*(6), 689–705.

Laitinen, E., & Lohan, E.-S. (2015). Are all the access points necessary in WLAN-based indoor positioning? In *Proc. 2015 int. conf. ICL-GNSS* (pp. 1–6). IEEE.

Laitinen, E., & Lohan, E. S. (2016). On the choice of access point selection criterion and other position estimation characteristics for WLAN-based indoor positioning. *Sensors*, *16*(5), 737.

Laoudias, C., Kolios, P., & Panayiotou, C. (2014). Differential signal strength fingerprinting revisited. In *Proc. 5th int. conf. IPIN)* (pp. 30–37). IEEE.

Laoudias, C., Piché, R., & Panayiotou, C. G. (2013). Device self-calibration in location systems using signal strength histograms. *Journal of Location Based Services*, *7*(3), 165–181. http://dx.doi.org/10.1080/17489725.2013.816792.

Laska, M., & Blankenbach, J. (2021). DeepLocBox: Reliable fingerprinting-Based Indoor Area localization. *Sensors*, *21*(6), 2000.

Lee, S., Kim, W., & Seo, D. (2022). Automatic self-reconstruction model for radio map in Wi-Fi fingerprinting. *Expert Systems with Applications*, *192*, Article 116455. http://dx.doi.org/10.1016/j.eswa.2021.116455.

Li, H., Chen, X., Wang, J., Wu, D., & Liu, X. (2022). DAFI: WiFi-based device-free indoor localization via domain adaptation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, *5*(4).

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020). On the sentence embeddings from pre-trained language models. (p. 12). arXiv preprint arXiv:2011.05864.

Lohan, E. S., Torres-Sospedra, J., Leppäkoski, H., Richter, P., Peng, Z., & Huerta, J. (2017). Wi-Fi crowdsourced fingerprinting dataset for indoor positioning. *Data*, *2*(4), 32.

Loshchilov, I., & Hutter, F. (2017). SGDR: stochastic gradient descent with warm restarts. In *Proc. 5th int. conf. ICLR* (p. 16). OpenReview.net.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Proc. 2017 NIPS* (pp. 4765–4774). URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proc. 2015 conf. EMNLP* (pp. 1412–1421). ACL.

Ma, Z., Wu, B., & Poslad, S. (2019). A WiFi RSSI ranking fingerprint positioning system and its application to indoor activities of daily living recognition. *International Journal of Distributed Sensor Networks*, *15*(4), Article 1550147719837916.

Machaj, J., Brida, P., & Piché, R. (2011). Rank based fingerprinting algorithm for indoor positioning. In *Proc. 2nd int. conf. IPIN* (pp. 1–6). IEEE.

Massey, F. J., Jr. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

Mathews, S. M. (2019). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In *Proc. 2019 computing conference* (pp. 1269–1292). Springer.

Mendoza-Silva, G. M., Torres-Sospedra, J., & Huerta, J. (2019). A meta-review of indoor positioning systems. *Sensors*, *19*(20), 4507.

Mendoza-Silva, G. M., Torres-Sospedra, J., Potortì, F., Moreira, A., Knauth, S., Berkvens, R., et al. (2020). Beyond euclidean distance for error measurement in pedestrian indoor location. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–11.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mohankumar, A. K., Nema, P., Narasimhan, S., Khapra, M. M., Srinivasan, B. V., & Ravindran, B. (2020). Towards transparent and explainable attention models. In *Proc. 58th conf. ACL* (pp. 4206–4216). ACL.

Murphy, K. P. (2023). *Probabilistic machine learning: advanced topics*. MIT Press, URL: probml.ai.

Nabati, M., & Ghorashi, S. A. (2023). A real-time fingerprint-based indoor positioning using deep learning and preceding states. *Expert Systems with Applications*, *213*, Article 118889. http://dx.doi.org/10.1016/j.eswa.2022.118889, URL: https://www.sciencedirect.com/science/article/pii/S0957417422019078.

Nowicki, M., & Wietrzykowski, J. (2017). Low-effort place recognition with WiFi fingerprints using deep learning. In *Proc. int. conf. auto., Vol. 550* (pp. 575–584). Springer.

Panwar, H., Gupta, P., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. (2020). A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. *Chaos, Solitons & Fractals*, *140*, Article 110190.

Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). Deeper attention to abusive user content moderation. In *Proc. 2017 conf. EMNLP* (pp. 1125–1135). ACL.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pérez-Navarro, A., Torres-Sospedra, J., Montoliu, R., Conesa, J., Berkvens, R., Caso, G., et al. (2019). Challenges of fingerprinting in indoor positioning and navigation. In *Geographical and fingerprinting data to create systems for indoor positioning and indoor/outdoor navigation* (pp. 1–20). Academic Press.

Potortì, F., Crivello, A., Palumbo, F., Girolami, M., & Barsocchi, P. (2021). Trends in smartphone-based indoor localisation. In *Proc. 21th int. conf. IPIN* (pp. 1–7). IEEE.

Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., & Petitcolas, F. (2021). Explainable machine learning for fraud detection. *Computer*, *54*(10), 49–59.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD* (pp. 1135–1144).

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Roy, P., & Chowdhury, C. (2021). A survey of machine learning techniques for indoor localization and navigation systems. *Journal of Intelligent & Robotic Systems*, *101*(3), 63.

Saccomanno, N., Brunello, A., & Montanari, A. (2020). Let's forget about exact signal strength: Indoor positioning based on access point ranking and recurrent neural networks. In *Proc. 17th EAI int. conf. mobiquitous* (pp. 215—224). ACM.

Saccomanno, N., Brunello, A., & Montanari, A. (2021). What you sense is not where you are: On the relationships between fingerprints and spatial knowledge in indoor positioning. *IEEE Sensors Journal*, *22*(6), 11. http://dx.doi.org/10.1109/JSEN.2021.3070098.

Sahar, A., & Han, D. (2018). An LSTM-based indoor positioning method using Wi-Fi signals. In *Proc. 2nd int. conf. ICVISP* (pp. 1–5). ACM.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. 2017 IEEE ICCV* (pp. 618–626).

Serrano, S., & Smith, N. A. (2019). Is attention interpretable? In *Proc. 57th conf. ACL* (pp. 2931–2951). ACL.

Shao, W., Luo, H., Zhao, F., Ma, Y., Zhao, Z., & Crivello, A. (2018). Indoor positioning based on fingerprint-image and deep learning. *IEEE Access*, *6*, 74699–74712.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., et al. (2021). Towards out-of-distribution generalization: A survey. (p. 22). CoRR abs/2108.13624. arXiv:2108.13624. URL: https://arxiv.org/abs/2108.13624.

Song, X., Fan, X., Xiang, C., Ye, Q., Liu, L., Wang, Z., et al. (2019). A novel convolutional neural network based indoor localization framework with WiFi fingerprinting. *IEEE Access*, *7*, 110698–110709.

Soro, B., & Lee, C. (2018). Performance comparison of indoor fingerprinting techniques based on artificial neural network. In *Proc. IEEE region 10 conf. TENCON* (pp. 56–61). IEEE.

Tiku, S., & Pasricha, S. (2019). PortLoc: A portable data-driven indoor localization framework for smartphones. *IEEE Design & Test of Computers*, *36*(5), 18–26. http://dx.doi.org/10.1109/MDAT.2019.2906105.

Torres-Sospedra, J., Mendoza-Silva, G. M., Montoliu, R., Belmonte, O., Benitez-Paez, F., & Huerta, J. (2016). Ensembles of indoor positioning systems based on fingerprinting: Simplifying parameter selection and obtaining robust systems. In *Proc. 7th int. conf. IPIN* (pp. 1–8). IEEE.

Torres-Sospedra, J., Montoliu, R., Trilles, S., Belmonte, Ó., & Huerta, J. (2015). Comprehensive analysis of distance and similarity measures for Wi-Fi fingerprinting indoor positioning systems. *Expert Systems with Applications*, *42*(23), 9263–9278.

Torres-Sospedra, J., Montoliu, R., Usó, A. M., Avariento, J. P., Arnau, T. J., Benedito-Bordonau, M., et al. (2014). UJIIndoorloc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems. In *Proc. 5th int. conf. IPIN* (pp. 261–270). IEEE.

Torres-Sospedra, J., & Moreira, A. J. C. (2017). Analysis of sources of large positioning errors in deterministic fingerprinting. *Sensors*, *17*(12), 2736.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Proc. 2017 conf. NeurIPS* (pp. 5998–6008). Curran Associates, Inc..

Wang, X., Gao, L., Mao, S., & Pandey, S. (2015). DeepFi: Deep learning for indoor fingerprinting using channel state information. In *Proc. 2015 conf. WCNC* (pp. 1666–1671). IEEE.

Wang, X., Gao, L., Mao, S., & Pandey, S. (2017). CSI-based fingerprinting for indoor localization: A deep learning approach. *IEEE Transactions on Vehicular Technology*, *66*(1), 763–776.

Wang, H., Guo, B., Wang, S., He, T., & Zhang, D. (2022). CSMC: Cellular signal map construction via mobile crowdsensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, *5*(4).

Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Liu, Y., et al. (2021). Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Computers in Biology and Medicine*, *137*, Article 104813.

Wang, L., Tiku, S., & Pasricha, S. (2021). CHISEL: Compression-aware high-accuracy embedded indoor localization with deep learning. (p. 4). CoRR abs/2107.01192. arXiv:2107.01192. URL: https://arxiv.org/abs/2107.01192.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In *Proc. 2019 conf. EMNLP-IJCNLP* (pp. 11–20). ACL.

Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, *1*(2), 270–280.

Wu, C., Xu, J., Yang, Z., Lane, N. D., & Yin, Z. (2017). Gain without pain: Accurate WiFi-based localization using fingerprint spatial gradient. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, *1*(2).

Xia, S., Liu, Y., Yuan, G., Zhu, M., & Wang, Z. (2017). Indoor fingerprint positioning based on Wi-Fi: An overview. *ISPRS International Journal of Geo-Information*, *6*(5), 135.

Yang, S., Dessai, P., Verma, M., & Gerla, M. (2013). FreeLoc: Calibration-free crowdsourced indoor localization. In *Proc. 32nd IEEE int. conf. INFOCOM* (pp. 2481–2489). IEEE.

Youssef, M. A., Agrawala, A. K., & Shankar, A. U. (2003). WLAN location determination via clustering and probability distributions. In *Proc. 1st IEEE int. conf. (PerCom)* (p. 143). IEEE Computer Society.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. (p. 839). arXiv preprint arXiv:2106.11342 2106.11342.

Zou, H., Luo, Y., Lu, X., Jiang, H., & Xie, L. (2015). A mutual information based online access point selection strategy for WiFi indoor localization. In *Proc. 2015 IEEE int. conf CASE* (pp. 180–185). IEEE.