An Exploration of the Interaction Between capsules with ResNetCaps models

Rita Pucci
University of Udine
Udine, Italy
rita.pucci@uniud.it

Christian Micheloni
University of Udine
Udine, Italy
christian.micheloni@uniud.it

Vito Roberto
University of Udine
Udine, Italy
vito.roberto@uniud.it

Gian Luca Foresti
University of Udine
Udine, Italy
gianluca.foresti@uniud.it

Niki Martinel
University of Udine
Udine, Italy
niki.martinel@uniud.it

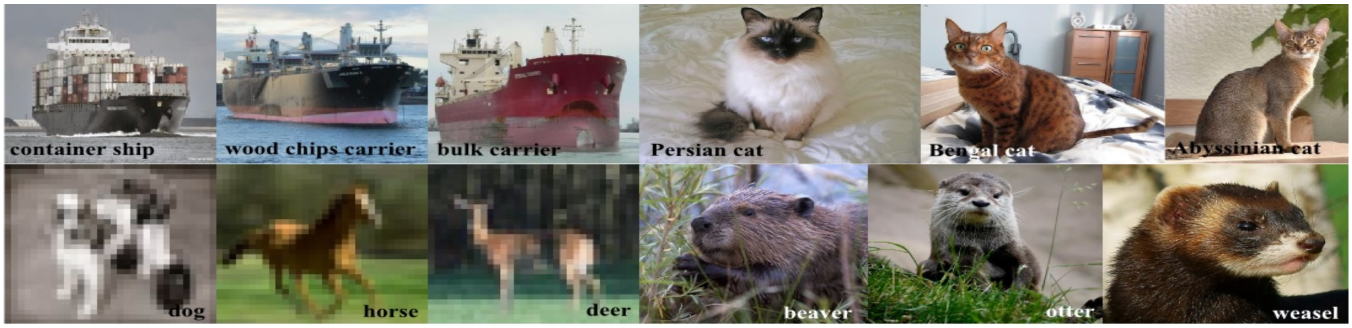


Figure 1: An example of the different challenges carried by the existing image recognition benchmark datasets: CIFAR10 (low image quality), MARVEL (very low inter-class variations), PETS (very low inter-class variations), and ANIMALS (challenging backgrounds and occlusions).

ABSTRACT

Image recognition is an open challenge in computer vision since its early stages. The application of deep neural networks yielded significant improvements towards its solution. Despite their classification abilities, deep networks need datasets with thousands of labelled images and prohibitive computational capabilities to achieve good performance. To address some of these challenges, the CapsNet neural architecture has been recently proposed as a promising machine learning model for image classification based on the idea of capsules. A capsule is a group of neurons whose output represents the presence of features of the same entity. In this paper, we start from the CapsNet architecture to explore and analyse the interaction between the presence of features within certain, similar classes. This is achieved by means of techniques for the features interaction, working on the outputs of two independent capsule-based models. To understand the importance of the interaction between capsules, extensive experiments have been carried out on four challenging dataset. Results show that the exploitation of capsules interaction yields to performance improvements.

KEYWORDS

Machine learning, Capsule network, image classification, bilinear function, feature interaction

ACM Reference Format:
Rita Pucci, Christian Micheloni, Vito Roberto, Gian Luca Foresti, and Niki Martinel. 2019. An Exploration of the Interaction Between capsules with ResNetCaps models. In *13th International Conference on Distributed Smart Cameras (ICDSC 2019), September 9–11, 2019, Trento, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3349801.3349804

1 INTRODUCTION

Recognising images is one of the key features to understand what is around us. Images are not just used by humans to understand an environment but also represent the main input source to tackle a wide plethora of research studies as ethology where images are used to classify animals or their behaviour [32], medicine where images can be used for cancer recognition [4], surveillance [19, 20, 22] and multimedia [23, 25, 26].

Nowadays, the image classification [24, 36] research field is dominated by Convolutional Neural Network (CNN) [14]. CNN achieved

ICDSC 2019, September 9–11, 2019, Trento, Italy

ACM ISBN 978-1-4503-7189-6/19/09…$15.00
https://doi.org/10.1145/3349801.3349804

important milestones in this era when CNN has been applied over a variety of vision-based problems such as person recognition [21], medical analysis [27], and animal behaviour understanding [42]. Despite their success, CNNs are easily fooled by adversarial perturbation [31], they only look for elements being present in the image without information about their relative location, and they need datasets that consist of thousands of images [5]. More recently, in [38] a new model for image recognition based on the idea of capsules [9] has been proposed: CapsNet. Since the CapsNet performances over MNIST are aligned with the state of the art [38], the performances obtained with CIFAR10 are lower than the state of the art [39]. Unlike traditional CNNs, CapsNet preserves the location and orientation of each component within an image. This important feature is implemented by the means of capsules that provide a different level of abstraction for the classification of each entity. The existing literature has not satisfactorily investigated the application of transfer learning techniques with CapsNet and more interestingly there is currently no work exploring the interaction between the features generated by capsules. **Contributions.** In this paper, we contribute to this field by *conducting an exploration of empowerment solutions for CapsNet architecture in order to improve their performances through capsules interactions*. In particular, we focus on transfer learning techniques to improve the features extraction ability of CapsNet and on interaction methods to investigate the use of two parallel CapsNet for analysing the interaction between the features present in the two outputs within certain similar classes. To properly evaluate the proposed approach, we conducted thorough evaluations on four image recognition benchmark datasets (*i.e.*, CIFAR10, MARVEL, PETS, and ANIMALS). We compared the results obtained by CapsNet, CapsNet with transfer learning with and without interaction methods. Results show that despite the different performances achieved by varying the number of training images, the complexity of the dataset, and the number of classes, the application of transfer learning in combination of interaction techniques, provides a push forward for better classification accuracy.

2 RELATED WORKS

2.1 CNN for image analysis

The CNN is designed to cope with data composed by multiple matrices. This innovative algorithm was presented in the early 90s [15] applied over document reading system. Since the beginning, this algorithm has been applied for object detection in natural images [33], and for face recognition [13]. It shortly became the primary method for the analysis of almost all sort of images [16]. The CNNs models evolved to many different architecture to deal with the wide range of inputs and hardware resources, in this paper we focus on the Residual Neural Network (ResNet) presented in [7]. Now, CNNs models, and in particular ResNet, are widely employed in challenging problems of image analysis for different research fields: object recognition [12, 40], object detection [12, 37, 41], medical application for pre-diagnosis [2, 18].

2.2 CapsNet for image analysis

Alongside the growth of CNNs, the new idea of capsules was presented in [9] where capsules were conceived as an approach to capture the representation of an entity. A capsule is a set of neurons that collectively produce an activity vector with one element for each neuron to hold that neuron's instantiation value (e.g., position). The capsule is the base idea of the CapsNet model presented in [38] that consists of layers of capsules. Authors formalise a training procedure based on routing-by-agreement where each capsule makes prediction over the parent capsule and computes a coupling coefficient between the actual capsule and the parent capsule outputs. Capsule outputs are vectors indicating the presence of an entity within the processed input, while their norms represent the confidence of the indication. Recent works present the application of CapsNet structure to cope with image classification using hyperspectral images [3] and medical images [1, 10, 28]. More than the application, we are interested in new architecture to empower CapsNet results such as in DenseNet [35], stressing out the CapsNet structure to better understand the performances [29, 30, 45]. Differently from previous works [35], we select a residual neural networks (ResNet) after an initial analysis among different pretrained CNNs models. We increase the depth of the CapsNet by transfer learning with ResNet [7]. Since we are interested in the interaction between capsules we focus our study on two popular methods of aggregation already used for image analysis with CNNs the bilinear [17] and the multimodal circulant fusion [43] methods. For this purpose we apply the bilinear and the multi-modal circulant fusion on the digit matrices obtained from two independent CapsNet models with ResNet. Since the digits matrices provide information for each class by a vector of probabilities of presence and features, by doing so we investigate how the information provided by the capsules for each class rely on the other classes.

3 METHODS

The proposed ResNetCaps architecture is shown in Figure 2. From the left, the input image is re-sized to $3 \times 224 \times 224$ for being consistent with the input dimension of ResNet. The architecture begins with the first three layers of the original structure of ResNet, they are presented in the window labelled ResNet in Fig. 2. Conv1 layer is composed of 64 filters with dimension 7×7. The second convolutional layer, Conv2, consists of 64 filters implemented by residual blocks. In these blocks, each layer feeds into the next layer and directly into the layers about $2-3$ hops away, the structure is explained in the residual block window in Figure 2. The last convolutional layer from the residual network is Conv3 and follows the same structure of Conv2 with 128 filters. The output of these first three layers is a matrix $128 \times 28 \times 28$, the output of the Conv3 does not need any resize for the CapsNet avoiding lost of information. This is the input to the CapsNet structure [38]. The architecture of CapsNet consists of two convolutional layers and one fully-connected layer. `Conv1 CapsNet` has 256 filters with dimension 9×9, this layer provides the activities of local feature detectors to the primary capsules. The `PrimaryCapsules` layer is the second convolutional layer with 32 channels of 8D convolutional capsules. In the lowest level, there are *primary capsules* that receive small regions of an image and detect presence and pose of an entity. In the higher level, there are routing capsules apt to identify complex entities. The final layer is a DigitCaps, it receives as input the output of all the capsules and it consists of 16D capsules per digit class.

For the analysis of the interaction between capsules, we take into consideration two independent ResNetCaps executed over the same input (*i.e.*, no shared weights). Each digit matrices describes the input with the probabilities computed by all the capsules in the ResNetCaps.

3.1 Bilinear interaction method

Bilinear model is presented in [17] for image classification. It is a quadruple B = (f_a, f_b, P, C) where f_a and f_b are digit matrices, P is implemented with sum-pooling to aggregate digits, and C is a classification function to interpret the aggregation of digits. The functions f_a and f_b are implemented with two ResNetCaps that output two digits matrices with dimension $N \times 16$ each and where N is the number of classes. To combine the two digits matrices we apply the Euclidean Matrix Product AB^T, the result is a matrix $N \times N$ that let us be able to investigate the interaction between all the classes. The matrix is squeezed in a vector by sum-by-row, and the resulting bilinear vector x is then passed through signed square root step $y = sign(x)\sqrt{|x|}$. In the final step, y is normalised by $z = y/||y||_2$ and z is the input of the final softmax classification function. The model is shown in Fig.3.

3.2 MultiModal interaction method

The multimodal fusion model is an interesting idea first presented in [43] for image classification. Given two digits matrices A and B, we sum each other by row to obtain two vectors of N values (N is the class number) a and b respectively. Each vector is multiplied by a weight matrix $N \times N$ initialised with random values $V = aW_a$ and $C = bW_b$. We use V and C to construct circulant matrices $X = circ(V)$ and $Y = circ(C)$. To develop an intense interaction between the two digits matrices, we multiply the circulant matrices by the vectors: $F = XC$ and $G = YV$. We end the multimodal interaction with a summation of the two matrices and we use the output as the input of a final softmax layer for classification. For better visualisation, all the steps are visualised in Fig. 4

4 DATASETS

In this work, we take into consideration four datasets of images that are available on open-source repositories. Figures 5a and 5b show samples respectively from MARVEL and CIFAR10 datasets. On the left side of the image there are samples from MARVEL dataset, three different classes (container ship, bulk carrier, and wood chips carrier) for four samples each. On the right side, we have samples of aeroplanes, birds and truck from CIFAR10. **MARVEL** [6] dataset consists of images of vessels with different types, dimensions, and colours organised in 26 classes. The entire dataset counts 237339 images of the vessels on the sea with a suggested split into 210954 images for training and 26385 images for test. The images in this dataset have $3 \times 256 \times 256$ resolution with a vast range of details but in this case, the classes are not far apart from each other indeed they are all part of the family of vessels and they are distinguished by models. **CIFAR10** [11] is a well known standard dataset for image recognition experimentation, it consists of 60000 images from 10 classes of objects from different contexts. We maintain the dataset split in training and test suggested by the dataset authors: 50000 images in training set and 10000 images in test set. All the images have resolution $3 \times 32 \times 32$, three times smaller than MARVEL. The last two datasets are focussed on animals: wild animals with ANIMALS and domestic animals with PETS. Figures 5c and 5d show samples respectively from ANIMALS and PETS datasets. On the left side, it is possible to observe samples of bobcat, otter, and skunk from ANIMALS. On the right side of the image, there are samples of basset-hound, Abyssinian cat, and American bulldog from PETS. In **ANIMALS** [44], the dataset consists of 50 different species of wild animals in a natural environment (free animals) and in captivity (animals in the zoo). In total there are 37324 images split into 26147 images in training set and 11177 images in test set. The last dataset is **PETS** [34], in this dataset, there are collected images of domestic animals in particular cats and dogs organised in 37 classes distinguished by breed. With this dataset we have 7391 images split in 5175 images for training set and 2216 images for test set (the split is made with ratio 70/30). The resolution of images is not homogeneous for all the images.

Table 1: Datasets summary

Dataset	#total	#trainset	#testset	#classes
MARVEL	237339	210954	26385	26
CIFAR10	60000	50000	10000	10
ANIMALS	37324	26147	11177	50
PETS	7391	5175	2216	37

We select these four datasets of images to perform a comparison between different digit aggregators and to analyse how they cope with classes strongly related as in MARVEL dataset, classes loosely related as in CIFAR10, with different resolutions, and different amount of data per class as in ANIMALS and PETS. Table 1 presents a summary of the datasets' features.

5 EXPERIMENTS

We demonstrate the importance of an interaction between multiple digits in order to improve the performance of transfer learning models with CapsNets. We compare the results obtained with CapsNet with results obtained with ResNetCaps with and without bilinear method and multimodal method. For this experiments we use an NVIDIA Titan Xp with single GPU with 12GB GDDR5X, and we develop the models in PyTorch 0.4 and CUDA 10.0. We use an initial public available code for CapsNet [8], and we develop all the other models [1]. In ResNetCaps, the first three layers from a ResNet model are pre-trained over ImageNet, the CapsNet layers are trained with the training datasets considered in this work. All the models are trained by the means of an Adam optimiser for $[10, 30, 100]$ epochs with initial learning rate of $[10e-3, 10e-6]$. The four datasets are fed in batch of 32 images with no pre-processing procedure applied. The test errors are computed once for each model and we do not apply any data augmentation scheme. Following, we present the results obtained in experimentation phase and organised by dataset. In all the models used during experiments the hyperparameters are not optimal and we use default parameter from CapsNet for normalisation and for training.

[1] Code is available at https://github.com/Riretta/NewDatasets_w_CapsuleNet

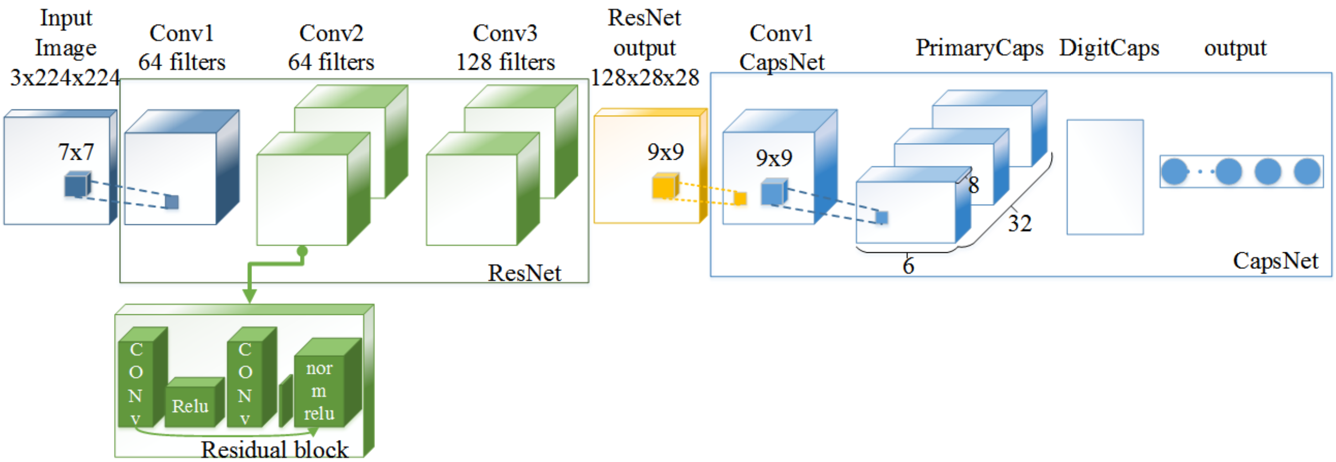


Figure 2: ResNetCaps Architecture: the two windows "ResNet" and "CapsNet" are presented for clarity: "ResNet" with residual blocks and "CapsNet" with the original structure of CapsNet.

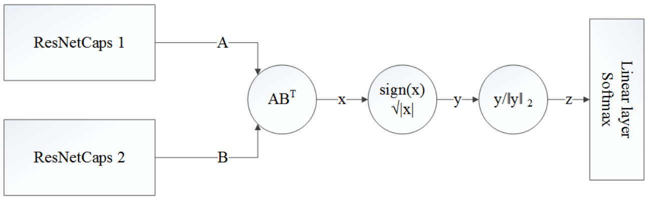


Figure 3: The bilinear layer first combines the digits from the Capsnets by the means of the matrix inner product with the Euclidean matrix product; second the matrix is normalised by the intensity of each digit; third the normalised vector is used to feed a softmax function.

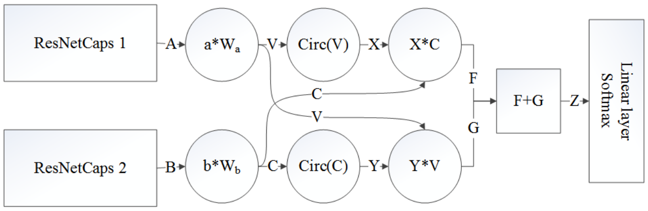


Figure 4: The multimodal layer first combines the digits in a deep interconnected relation between each other; the final vector is used to feed a softmax function.

Table 2: Results over MARVEL and CIFAR10 test set

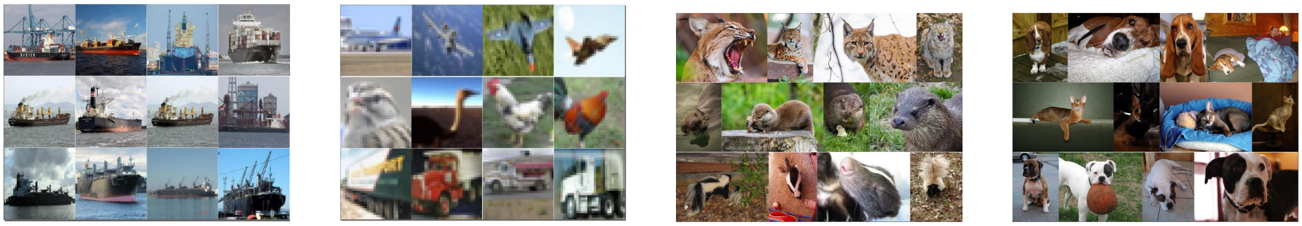
	CapsNet	ResNetCaps	Bilinear	Multimodal
		MARVEL test set		
Accuracy	46.0%	50.0%	**61.9%**	57.3%
Time execution	140s	110s	237s	400s
		CIFAR10 test set		
Accuracy	68.7%	78.0%	**78.6%**	78.5%
Time execution	37s	52s	26s	74s

5.1 MARVEL

MARVEL dataset is the biggest dataset used to validate the methods. It is worth noting that the classes are from the same context, there is not real variability between the objects represented because the images are all vessels and the classes make an evidence of the type of vessel. What we observe in Table 2 is a strong improvement in applying the interaction methods. We start with the application of a single CapsNet, trained for 10 epochs with a learning rate of $= 10e - 5$. The accuracy obtained is 46% and it improves by 4% when we apply a ResNetCaps. This result is a promising initial result for a dataset that is challenging for classification. When we apply the interaction methods we observe an additional improvement in performances. In fact, with the multimodal method, the ability in classification is increased by 11.3% and with the bilinear method, we obtain an improvement of 16%.

5.2 CIFAR10

CIFAR10 dataset has the lowest number of classes used in this work. Table 2 shows the accuracy obtained with the test dataset. The single CapsNet with the original structure is applied with $learningrate = 0.001$ for a training period of 60 epochs. The trained model provides an accuracy of 69%, this result verified the performance at the state of the art. The ResNetCaps model is trained for 7 epochs with a similar learning rate of CapsNet and the results obtained are far way better. From 69% with CapsNet, we reach an accuracy of 78% with ResNetCaps. This improvement underlines the promising benefit obtained by transfer learning with this model. Also with this dataset, the application of the bilinear interaction method and the multimodal interaction methods provides an additional increase in accuracy by 0.6%.



a: MARVEL **b: CIFAR10** **c: ANIMALS** **d: PETS**

Figure 5: Datasets samples: each row is a class, each column is an instance of the class

Table 3: Results over ANIMALS and PETS test set

Dataset	CapsNet	ResNetCaps	Bilinear	Multimodal
		ANIMALS test set		
Accuracy	12.1%	**60.4%**	54.8%	58.2%
Time execution	454s	480s	483s	680s
		PETS test set		
Accuracy	9.5%	**48.6%**	23.6%	46.6%
Time execution	29s	34s	36s	60s

5.3 ANIMALS

ANIMALS is an interesting dataset of wild animals shot by camera traps, cameras, and mobiles and represents a challenging dataset due to his variability in classes and the images types with not a homogeneous. Even if the amount of images is almost half the amount of images in CIFAR10, we have in this case, five times the classes with an average amount of 750 images per class. We can observe the incidents of the dimension in dataset over the behaviour of CapsNet. In fact, the higher number of classes and a lower number of images affect the performance of the model that provides an accuracy of 12%. Here ResNetCaps makes an evident improvement providing an accuracy of 60% that 48% better than CapsNet. In this case, the interaction methods improve the accuracy by 42% with bilinear method and by 46% with multimodal method, these improvements are lower than ResNetCaps but still evident compared to CapsNet.

5.4 PETS

PETS is the smallest dataset taken into consideration. In this case, the dataset has less than 10000 images in total for 37 classes. This is an interesting challenge for a model that is thought of being able to cope with a small training dataset [38]. The results obtained with the four models are shown in Table 3. Starting with CapsNet, we obtain a low result of only 9% in test dataset after a training phase of 10 epochs. We decide for a brief training phase due to a stabilisation of the increase of accuracy. With ResNetCaps, we obtain the best result of 48% with this dataset, that makes evidence of a need for prior knowledge with a dataset with small dimension, and low variability. The bilinear method does not provide any improvement in accuracy compared to ResNetCaps but an improvement of 14.6% compare to CapsNet. With the multimodal method, we can observe a different behave, in fact with an accuracy of 46%, this interaction method obtains a similar result of ResNetCaps.

6 DISCUSSION

We were able to apply CapsNet and ResNetCaps to our four datasets to add results at the state of the art. The four datasets selected for this paper are from different subject areas, with a different distribution, and different number of classes. All four datasets consist of coloured images (RGB) shot with camera traps, cameras, and mobiles. We take into consideration these characteristics in the analysis of results. From [30] and from our experiments, we know that CapsNet, with no ensemble, can not achieve the state of the art with CIFAR10 getting 68% accuracy. On the other hand, CapsNet is still a new model and there are a low bibliography and application at the state of the art. The behaviour observed with CIFAR10 is due by intra-class variation and background noise, and in our experiments, this is true also with all the other three datasets of images. In [38], authors propose a solution to the lack in performances with CIFAR10, the ensemble between seven models. This strategy provided a tangible improvement in performances (from 69% to 89%) at the price of a high number of hyperparameters and the need of powerful computational resources. We want to avoid the ensemble strategy and, for this purpose, we implement ResNetCaps, that is based on the original CapsNet and empowered with three layers from a Residual neural network. We proved that ResNetCaps outperforms CapsNet in every one of the cases took into consideration. The improvement is evident in the range 4% with MARVEL dataset up to 39% with PETS dataset. The idea of interaction between more than one ResNetCaps in the classification tasks introduces the interest for interaction methods between independent CapsNet models over the same input. The comparison among ResNetCaps with and without the interaction methods does not show an evident increase in performance for CIFAR10 (0.5%). With PETS and ANIMALS the behaviour is slightly lower for the ResNetCaps. For PETS this is due to the low variability between classes that does not provide additional information when let interact with each other. ANIMALS has high variability in classes with a low amount of images compare to other datasets, that can limit the capsules ability in the identification of entity presence. With MARVEL dataset we observe a strong benefit in using interaction methods obtaining an improvement in accuracy by 12%. In this case, the interaction methods emphasise the characteristics that are crucial for the model to discern among vessels.

7 CONCLUSION

We proposed two methods for building interaction between digit matrices obtained from two independent ResNetCaps. We first build ResNetCaps by increasing the convolutional phase of CapsNet with the first two layers of a Residual neural network pre-trained over ImageNet. The addition of the two layers helps learn better the features in the image ending with an increase in performances with complex datasets. The use of transfer learning with CapsNet has a demonstrated effectiveness in performance compared to the original structure of CapsNet and this is emphasised by the application of interaction methods, in particular, we validate them over CIFAR10, MARVEL, PETS and ANIMALS datasets. In all these datasets we observed an improvement in performances by $1-10\%$ opening up to promising new applications. Proposing the use of interaction methods, we introduce an interesting possibility in using multiple entries to improve the ability of the CapsNet algorithms in recognise images. In future, we plan to apply new interactive methods on digit matrices obtained by different sources used as different inputs for each ResNetCaps. This might provide a better understanding of an interactive policy between digit matrices.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This research was partially supported by the PRESNET MoD project CIG68827500FB.

REFERENCES

[1] Parnian Afshar, Arash Mohammadi, and Konstantinos N Plataniotis. 2018. Brain tumor type classification via capsule networks. In *ICIP*. 3129–3133.

[2] Lei Bi, Jinman Kim, and Euijoon et al. Ahn. 2017. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. *arXiv:1703.04197* (2017).

[3] Fei Deng, Shengliang Pu, and Xuehong et al. Chen. 2018. Hyperspectral image classification with capsule network using limited training samples. *Sensors* 18, 9 (2018), 3153.

[4] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[5] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international Conf. on artificial intelligence and statistics*. 249–256.

[6] Yucesoy V. Koc A. Gundogdu E., Solmaz B. 2016. Marvel: A Large-Scale Image Dataset for Maritime Vessels. In *ACCV*.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[8] higgsfield. 2019. *Capsule Network Tutorial*. Technical Report. https://github.com/higgsfield/Capsule-Network-Tutorial.

[9] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In *International Conf. on Artificial Neural Networks*. 44–51.

[10] Amelia Jiménez-Sánchez, Shadi Albarqouni, and Diana Mateus. 2018. Capsule networks against medical imaging data challenges. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. 150–160.

[11] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.

[13] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. 1997. Face recognition: A convolutional neural-network approach. *IEEE Transactions on neural networks* 8, 1 (1997), 98–113.

[14] Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361, 10 (1995), 255–257.

[15] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. 1990. Handwritten digit recognition with a back-propagation network. In *NIPS*. 396–404.

[16] Yann LeCun, Koray Kavukcuoglu, Clément Farabet, et al. 2010. Convolutional networks and applications in vision.. In *ISCAS*, Vol. 2010. 253–256.

[17] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. 2015. Bilinear cnn models for fine-grained visual recognition. In *ICCV*. 1449–1457.

[18] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, and et al. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

[19] Niki Martinel. 2018. Accelerated low-rank sparse metric learning for person re-identification. *Pattern Recognition Letters* 112 (2018), 234–240.

[20] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2018. Unsupervised Hashing with Neural Trees for Image Retrieval and Person Re-Identification. In *ICDSC*.

[21] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2019. Aggregating Deep Pyramidal Representations for Person Re-Identification. In *CVPR*.

[22] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2019. Distributed person re-identification through network-wise rank fusion consensus. *Pattern Recognition Letters* (2019). https://doi.org/10.1016/j.patrec.2018.12.015

[23] Niki Martinel, Christian Micheloni, and Gian Luca Foresti. 2013. Robust Painting Recognition and Registration for Mobile Augmented Reality. *IEEE SPL* 20, 11 (2013), 1022–1025.

[24] Niki Martinel, Christian Micheloni, and Gian Luca Foresti. 2015. The Evolution of Neural Learning Systems: A Novel Architecture Combining the Strengths of NTs, CNNs, and ELMs. *IEEE Systems, Man, and Cybernetics Magazine* 1, 3 (2015), 17–26.

[25] Niki Martinel, Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. 2015. A Structured Committee for Food Recognition. In *ICCV*. 92–100.

[26] Niki Martinel, Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. 2015. On Filter Banks of Texture Features for Mobile Food Classification. In *ICDSC*. 11–16.

[27] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*. 565–571.

[28] Aryan Mobiny and Hien Van Nguyen. 2018. Fast capsnet for lung cancer screening. In *International Conf. on Medical Image Computing and Computer-Assisted Intervention*. 741–749.

[29] Rinat Mukhometzianov and Juan Carrillo. 2018. CapsNet comparative performance evaluation for image classification. *arXiv:1805.11195* (2018).

[30] Prem Nair, Rohan Doshi, and Stefan Keselj. 2018. Pushing the limits of capsule networks. *Technical note* (2018).

[31] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*. 427–436.

[32] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* 115, 25 (2018).

[33] Steven J Nowlan and John C Platt. 1995. A convolutional neural network hand tracker. *NIPS* (1995), 901–908.

[34] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. 2012. Cats and Dogs. In *CVPR*.

[35] Sai Samarth R Phaye, Apoorva Sikka, and Abhinav et al. Dhall. 2018. Dense and diverse capsule networks: Making the capsules learn better. *arXiv:1805.04001* (2018).

[36] Asha Rani, Gian Luca Foresti, and Christian Micheloni. 2015. A neural tree for classification using convex objective function. *Pattern Recognition Letters* (2015).

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 91–99.

[38] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *NIPS*. 3856–3866.

[39] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *NIPS*. 2377–2385.

[40] Christian Szegedy, Sergey Ioffe, and Vincent et al. Vanhoucke. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

[41] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. 2013. Deep neural networks for object detection. In *NIPS*. 2553–2561.

[42] Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas. 2017. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological informatics* 41 (2017), 24–32.

[43] Aming Wu and Yahong Han. 2018. Multi-modal Circulant Fusion for Video-to-Language and Backward.. In *IJCAI*, Vol. 3. 8.

[44] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE TPAMI* (2018).

[45] Canqun Xiang, Lu Zhang, Yi Tang, Wenbin Zou, and Chen Xu. 2018. MS-CapsNet: A novel multi-scale capsule network. *IEEE SPL* 25, 12 (2018), 1850–1854.