

Self-Attention Capsule Networks for Image Classification

Assaf Hoogi^{1,2}, Brian Wilcox^{3*}, Yachee Gupta^{1*}, Daniel L. Rubin¹

¹ Dept. of Biomedical Data Science, Stanford University

² Dept. of Computer Science and Applied Math, The Weizmann Institute of Science

³ Dept. of Electrical Engineering, Stanford University

* Equal contributors

Abstract

We propose a novel architecture for image classification, called Self-Attention Capsule Networks (SACN). SACN is the first model that incorporates the Self-Attention mechanism as an integral layer within the Capsule Network (CapsNet). While the Self-Attention mechanism selects the more dominant image regions to focus on, the CapsNet analyzes the relevant features and their spatial correlations inside these regions only. The features are extracted in the convolutional layer. Then, the Self-Attention layer learns to suppress irrelevant regions based on features analysis, and highlights salient features useful for a specific task. The attention map is then fed into the CapsNet primary layer that is followed by a classification layer. The SACN proposed model was designed to use a relatively shallow CapsNet architecture to reduce computational load, and compensates for the absence of a deeper network by using the Self-Attention module to significantly improve the results. The proposed Self-Attention CapsNet architecture was extensively evaluated on five different datasets, mainly on three different medical sets, in addition to the natural MNIST and SVHN. The model was able to classify images and their patches with diverse and complex backgrounds better than the baseline CapsNet. As a result, the proposed Self-Attention CapsNet significantly improved classification performance within and across different datasets and outperformed the baseline CapsNet not only in classification accuracy but also in robustness.

1. Introduction

Image classification is a very challenging task, mostly because of the significant intra-class and inter-class variability, arising from different image acquisition conditions, rigid and non-rigid deformations, occlusions and corruptions. Handcrafted low-level fea-

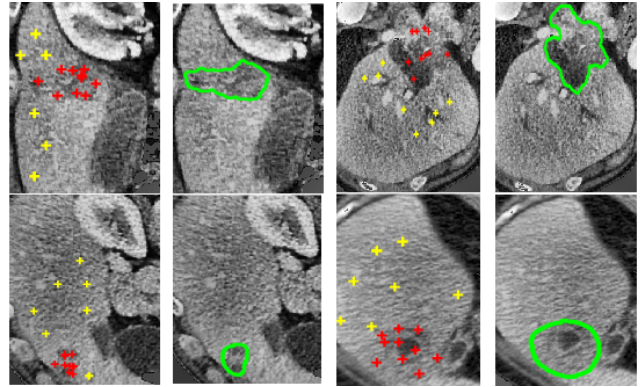


Figure 1. Classification of randomly selected patches - CT Liver lesions (LiTS public). Each pair of images (in the same row) represent the original image with the radiologist's lesion annotation (green) and the processed image. Red - classified lesion patches, Yellow - classified normal patches. The dataset contains difficult cases such as low contrast and highly heterogeneous lesions.

tures were proposed to handle these challenges, while unsupervised learning approaches are regularly developed to avoid the limitations of handcrafted features such as being user dependant.

Recent advances in computer vision highlight the capabilities of deep learning approaches to solve these challenges, achieving state of the art performances in many classification tasks. The main reason for the success of deep learning is the ability of Convolutional Neural Networks to learn a hierarchical representation of the input data. AlexNet, which was presented by Krizhevsky et al. [7] was one of the first and the simplest architectures for image classification. Later on, a deeper VGG16 model was introduced, dealing with nonlinear transformations [17]. ResNet was then developed to solve a common problem

in deep learning of increased test error rate while increasing the architecture depth [4]. DenseNet-40 [5] was recently developed and is currently considered as the best state of the art method for image classification tasks. It is similar to the ResNet architecture with the difference being significantly densely connected feature maps in the final layer of a dense block instead of a residual block.

Deep learning-based approaches became popular also in medical image domain due to the increasing computational power and availability of data. However, these methods still lack in robustness across different datasets and require a significant amount of annotated data. These limitations are even more substantial in the medical domain (relative to the natural domain) because the annotated data is highly heterogeneous and its size is relatively small. To tackle these challenges, the U-Net architecture was developed. It includes skip connections and was designed for medical images, wherein these additional connections can extract larger amounts of information from the limited data size [15] [22].

CapsNet is one of the most recent architectures that was developed by Hinton's group for image classification [16]. It is powerful and was designed to deal with small datasets, as is typical for the medical domain. It learns the spatial correlation between objects, while the capsules allow the model to recognize multiple objects in the image even if they overlap. However, the CapsNet does not involve local constraints for feature learning i.e. CapsNet does not choose the features locally. Learning the important local features can be done manually or by methods such as auto-encoders, both of which increase the number of steps/tasks and increase the computational cost.

Therefore, developing a new architecture that will help to solve the mentioned limitations is highly desired and can help in advancing the field of image classification, especially in the medical domain.

This paper presents a **significant improvement of the Capsule Networks architecture** and has several key contributions.

- We introduce a novel architecture, called **Self-Attention Capsule Networks (SACN)**. The architecture includes an integral Self-Attention layer between the convolutional and the primary CapsNet layers, which allows the model parameters, even in shallower layers, to be updated mostly based on image regions that are more relevant to a given task. The attention mechanism, which

is used as a non-local operation, solves the task of learning the important features while the CapsNet considers the positional / rotational spatial relation between these features.

- The proposed architecture is **designed to work well under the constraint of limited computational resources**. While the baseline CapsNet [16] is considered an expensive architecture in terms of computational cost, the proposed model was designed to use a relatively shallow CapsNet architecture and compensates the absence of a deeper network by using the Self-Attention module to significantly improve the feature extraction. The attention mechanism supplies better classification with lower computational cost.
- We are not familiar with other works that were tested on both medical and natural image domains, as these domains have substantially different image characteristics. Here we conducted extensive experiments to show the **generalization of the model**. We showed that the new model was able to supply **more accurate, robust and stable image classification** within and across different datasets and domains (compared with previous methods). In addition, we are not familiar with any other architectures that were **specifically designed to deal with the challenges diversity of medical data for classification tasks**.

2. Related work

2.1. CapsNet architecture

Recently developed Capsule networks represent a breakthrough in the field of neural networks. The CapsNet architecture contains three types of layers - the convolutional layer, the primary capsule layer and the classification capsule layer [16]. Capsule networks are powerful because of two key ideas that distinguish them from the traditional CNNs; 1) dynamic routing-by-agreement instead of max pooling, and 2) squashing, where scalar output feature detectors of CNNs are replaced with vector output capsules. Routing-by-agreement means that it is possible to selectively choose which parent in the layer above the capsule is sent to. For each optional parent, the capsule network can increase or decrease the connection strength. As a result, the CapsNet can keep the spatial correlations between objects within the image [16]. Squashing means that instead of having individual neurons sent through non-linearities as is done in CNNs, capsule networks have their output squashed as an entire vector. The squashing function enables a better repre-

sensation of the probability that an object is present in the input image.

CapsNet should also supply better results than state of the art techniques for a relatively small and sparse set of images, as is typical for medical imaging. Afshar et al. [1] incorporated CapsNets for brain tumor classification. The authors investigated the over-fitting problem of CapsNets based on a real set of MRI images. Their results show that CapsNet can successfully outperform CNNs for the brain tumor classification problem. However, CapsNet is an expensive architecture in terms of computational and memory loads. As a result, the commonly-used CapsNets are relatively shallow architectures, which were proved to be better mainly for simple datasets. They did not perform well for more complex data. Deliege et al. [2] introduced HitNet, a deep learning network characterized by the use of a Hit-or-Miss layer composed of capsules. The idea is that the capsule corresponding to the true class has to make a hit in its target space, and the other capsules have to make misses. The method converged faster than CapsNet but their results were not able to outperform CapsNet for complex datasets. In [20], the authors explored the effect of a variety of CapsNet modifications, ranging from stacking more capsule layers to trying out different parameters such as increasing the number of primary capsules or customizing an activation function. However, the best validation accuracy for a relatively complex dataset that their architecture reached was 71.55%, only comparable to CapsNet performance on the same dataset. They mentioned that computational resources limited their performance. Another architecture, Diverse Capsule Networks, introduced in [12], was able to supply only a 0.31% improvement over the baseline CapsNet accuracy.

2.2. Self-Attention mechanism

The Self-Attention mechanism can help the model focus on more relevant regions inside the image and gain better performance for classification tasks with fewer data samples [13] or more complex image backgrounds. Attention mechanism allows models to learn deeper correlations between objects [9] and helps discover interesting new patterns within the data [6] [11]. Additionally, it helps in modeling long-range, multi-level dependencies across image regions. Wang et al. [19] address the specific problem of CNNs processing information too locally by introducing a Self-Attention mechanism, where the output of each activation is modulated by a subset of other activations. This helps the CNN to consider smaller parts of the image if necessary. Larochelle and Hinton [8] pro-

posed using Boltzmann Machines that choose where to look next to find locations of the most informative intra-class objects, even if they are far away in the image. Reichert et al. proposed a hierarchical model to show that certain aspects of attention can be modeled by Deep Boltzmann Machines [14].

Attention-based models were also proposed for generative models. In [18], the authors introduce a framework to infer the region of interest for generative models of faces. Their framework is able to pass the relevant information only, through the generative model. Recent technique that focuses on generative adversarial models is called SAGAN [21]. The authors proposed Self-Attention Generative Adversarial Networks (SAGAN) that achieve state-of the art generative results on the ImageNet dataset.

Recent work that deals specifically with medical data can be found in [10]. This work presents an attention mechanism that is incorporated in the U-Net architecture for tissue/organ identification and localization. However, U-Net was mainly developed for segmentation tasks in the medical domain, rather than for image classification.

Our SACN model plays a key role in advancing the medical imaging, as most classification tasks in this domain need positional relationships between features to perform optimally. By using our architecture, we can focus the attention on relevant locations in the image and better analyze the spatial relationships between their features by taking advantage of the CapsNet structure.

3. The proposed model

Our proposed model is illustrated in Figure 2. The information extracted from the initial convolutional layers of the CapsNet is fed into a Self-Attention mechanism to disambiguate irrelevant and noisy responses. According to the dominant features, we pass only relevant activations through the Primary Capsule layer. This allows the model parameters, even in shallower layers, to be updated mostly based on image regions that are more relevant to a given task. Let $x \in \mathbb{R}^{C \times N}$ be the output feature matrix that is extracted from the input image. This set of features is obtained by the initial convolutional layer. Let $f(x_i)$ and $g(x_j)$ be position modules that are used to calculate attention. $f(x_i)$ and $g(x_j)$ take input feature maps at the i^{th} and j^{th} positions. Then, by using 1×1 convolution kernels, they output new feature maps. In order to model long-range dependencies and to calculate the attention map, we use a non-local approach [19]. It helps the CapsNet model relationships between spatial regions that are far from each other while the Self-

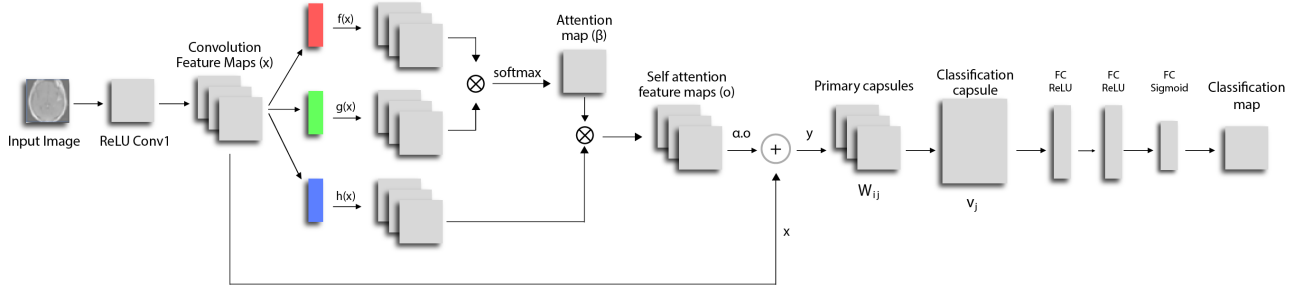


Figure 2. Our proposed SACN architecture.

Attention mechanism helps to create a balance between efficiency and long-range dependencies (large receptive fields) by supplying a weighted sum of the features at all image locations. We define the non-local operation as:

$$\eta_{ij}(x) = f(x_i)^T g(x_j) \quad (1)$$

$f(x_i) = W_f x_i$, $g(x_i) = W_g x_i$, where $W_g \in \mathbb{R}^{C \times C}$, $W_f \in \mathbb{R}^{C \times C}$ are the learned weight matrices, which are implemented as 1×1 convolutions. We then compute the softmax of η_{ij} to get an output attention map β_{ij} ,

$$\beta_{ij} = \frac{\exp(\eta_{ij})}{\sum_{i=1}^N \exp(\eta_{ij})} \quad (2)$$

To obtain the final Self-Attention map, $o \in \mathbb{R}^{C \times N}$, which will be the input of the primary CapsNet capsule, we apply matrix multiplication between the attention map β_{ij} and $h(x)$,

$$o_j = \sum_{i=1}^N \beta_{ij} h(x_i) \quad (3)$$

where $h(x) = W_h x_i$ is another input feature channel (see Figure 2) and similarly to W_f and W_g , W_h is also a learned weight matrix. Therefore, the final output of the layer of the Self-Attention mechanism is

$$y_i = \alpha o_i + x_i \quad (4)$$

In our model, α is initialized to 0. As a result, the model can explore the local spatial information first, before automatically refining it with the Self-Attention and analyzing higher data complexity by considering further regions in the image. Then, the network gradually learns to assign higher weight to the non-local regions. By initializing α to 0 and with no requirement of other pre-defined parameters, we are not dependent on the user input, contrary to

common attention mechanisms.

The final output of the Self-Attention layer, y_i , is then fed into the CapsNet primary layer. Let v_j be the output vector of capsule j . The length of the vector, which represents the probability of whether or not a specific object is located in that given location in the image, should be between 0 and 1. To ensure that, we apply a squashing function that keeps the positional information of the object. Short vectors are shrunk to almost 0 length and long vectors are brought to a length slightly below 1. The squashing function is defined as

$$v_j = \frac{\|\sum_i c_{ij} W_{ij} y_i\|^2}{(1 + \|\sum_i c_{ij} W_{ij} y_i\|^2)} \frac{\sum_i c_{ij} W_{ij} y_i}{\|\sum_i c_{ij} W_{ij} y_i\|} \quad (5)$$

where W_{ij} is a weight matrix and c_{ij} are the coupling coefficients between capsule i and all the capsules in the layer above j that are determined by the iterative dynamic routing process

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_j \exp(b_{ij})} \quad (6)$$

b_{ij} are the log prior probabilities that i^{th} capsule should be coupled to j^{th} capsule.

To obtain a reconstructed image during training, we use the vector v_j that supplies the highest coupling coefficient, c_{ij} . Then, we feed the correct v_j through two fully connected ReLU layers. The reconstruction loss $L_R(I, \hat{I})$ of the architecture is defined as,

$$L_R(I, \hat{I}) = \|I - \hat{I}\|_2^2 \quad (7)$$

where I is the original input image and \hat{I} is the reconstructed image. $L_R(I, \hat{I})$ is used as a regularizer that takes the output of the chosen v_j and learns to reconstruct an image, with the loss function being the sum

of squared differences between the outputs of the logistic units and the pixel intensities (L2-Norm). This forces capsules to learn features that are useful for the reconstruction procedure which inherently allows for the model to learn features at near-pixel precision. Therefore, the better the reconstruction loss the prediction. The reconstruction loss is then added to the following margin loss function, L_M ,

$$L_M = \sum_k T_k \max(0, m^+ ||v_k||)^2 + \sum_k \lambda(1 - T_k) \max(0, ||v_k|| - m^-)^2 \quad (8)$$

$T_k = 1$ if an instance of class k is present. $m^+ = 0.9$ and $m^- = 0.1$ were selected as was suggested in [16]. The end to end SACN architecture is evaluated and its weights are trained by using the total loss function, L_T , which is the total of all losses over all classes k ,

$$L_T = L_M + \zeta I_{size} L_R \quad (9)$$

$\zeta = 0.0005$ is a regularization factor per channel pixel value that ensures that the reconstruction loss does not dominate over L_M during training. $I_{size} = H * W * C$ is the number of input values, based on the height, width and number of channels in the input.

3.1. Architecture and Hyper-parameters

Since the CapsNet is a relatively expensive architecture in terms of computational load, we designed our architecture to work well under the constraint of limited computational resources and boost the performance by adding the Self-Attention module. Therefore, our CapsNet architecture contains one convolutional layer with 5×5 filters, one Capsule layer and one routing iteration. The attention module includes four additional 1×1 convolutional layers. We chose a batch size of 64 and a learning rate of $1e^{-3}$. Thirty epochs were used because this was sufficient to train the smaller dataset. Patches size of $16 \times 16 \times 1$ was chosen as it supplied the best classification results. A value 0.5 was chosen for the λ down-weighting of the loss, together with a weight variance of 0.15.

Algorithm 1 describes the training process of the proposed model.

Data: I, G : Pairs of image I and the ground truth G

Result: Y_{out} : Final instance classification

while *not converging* **do**

CapsNet Convolutional Layer: features are extracted and are divided into three output feature vectors (f, g, h).

Attention Layer: Self-Attention map y_i is generated based on the features vectors, attention map β_{ij} , learned weight matrix W_h and a specific image location x_i .

Primary and Classification Layers: the dominant features are then fed into the Primary CapsNet layer and from there to the Classification layer. Output classification Y_{out} is obtained.

Calculate the Attention-based CapsNet loss:

$L_T \leftarrow Loss(G, Y_{out})$

Back-propagate the loss and compute: $\frac{\partial L}{\partial W}$

Update the weights: matrices W are updated for both the Self-Attention layer and the CapsNet architecture.

end

Algorithm 1: Our SACN training process

4. Experiments

4.1. Datasets

We conducted extensive experiments on highly diverse medical data, and present initial results for natural data as well (as described in the "Natural datasets" subsection). The medical dataset is composed of three separate subsets of images, each contains cancer lesions that are located at different body organs and were screened by different imaging modalities. Two subsets were collected by radiologists at Stanford hospital (250 CT Lung images and 369 MR Brain tumors) and the third set of 1102 CT Liver lesions, is a public one (LiTS). In addition to the differences in the organs and the imaging modalities, these datasets are different from each other also regarding other acquisition criteria; 1) their spatial resolution is within the range of $0.78mm/pixel - 0.94mm/pixel$ and 2) their slice thickness ranges from $2.5mm$ to $5mm$. These differences affect the appearance of the cancer lesions, characterized by a different noise level, homogeneity or contrast relative to the surrounding normal tissue. Each subset has its major challenges but the CT Lung dataset is considered the more difficult dataset for patch classification, while CT Liver is the easiest one. The inter- and intra-variability between sets of images is shown in Figure 1 and in Figure 3. An external expert annotated two separate regions in each image -

normal tissue and cancer lesion. Thirty patches were extracted from each region, means that each training image supplied 60 samples to the whole training cohort. For all experiments, we used 80% of the dataset for training, 10% for testing and 10% percent for validation.

4.2. Performance evaluation

The performance of our method was measured as a patch-wise classification - normal or lesion patches. To evaluate the capabilities of our proposed method, we compared the developed architecture with 1) the baseline CapsNet that this work mainly aims to improve, and with 2) the state of the art ResNet and DenseNet-40 architectures. The DenseNet and ResNet were adjusted for smaller input image patches of 16×16 pixels. These architectures have been modified for image sizes much smaller than ImageNet and therefore, this is a fair comparison. We evaluated the effectiveness of these methods by calculating several statistics and statistical significance between the methods was calculated by using Wilcoxon paired test.

5. Results

5.1. Qualitative evaluation

Figure 3 shows the classification results of a subset of randomly chosen patches. For the purpose of visualization, only the colored patches have been classified into normal/lesion regions. The figure shows the substantial diversity of the image characteristics, within and across subsets (CT Lung, CT Liver and MR Brain). Our method shows its ability to handle small lesions, highly heterogeneous lesions and low contrast lesions. It can also distinguish very well between normal structures within the tissue (e.g. blood vessels in CT lung, normal structures in the MR Brain image) and cancer lesions. All these challenges, which usually fail common techniques, are dealt well by our proposed method.

5.2. Quantitative evaluation and Comparison with other common techniques

Table 1 presents the classification accuracy of *all* patches in the testing set (contrary to the subset of patches that is visualized in figure 3). We set the optimal parameters for every architecture we compared our SACN with, to ensure that any difference between the performances is directly related to the novelty of our architecture. Table 1 shows that our method significantly outperforms all other methods for each subset that has been analyzed and for each of the following parameters that has been explored.

First, the performance accuracy within each specific subset (Liver, Lung, Brain) is consistently higher when using our proposed method. *Second*, the standard deviation (std) of the classification accuracy over different images within the same subset is lower than the equivalent values when using the baseline CapsNet, DenseNet-40 and ResNet. *Third*, the robustness and the stability across different subsets are also significantly higher when using our model.

It is worth mentioning that the performance difference between our technique and the other methods we compared with, becomes more significant and going along with the level of the data complexity. This key result enhances the strength of our method. For example the difference between our method and the others is larger for CT Lung and smaller for CT Liver.

To ensure that our architecture does not overfit to the training data, we explored the loss/error rates of the training and the validation sets for each individual tested subset (Figure 4). It can be clearly seen that the loss of the training and the validation sets are comparable, having the same trend and without a substantial differences between them.

5.3. Natural datasets

We showed *preliminary* results for natural data as well, exploring the generalization of the proposed technique to other domains, except for the medical one. The MNIST database includes a training set of 60,000 hand-written digits examples, and a test set of 10,000 examples. The Street View House Numbers (SVHN) is a real-world image dataset It contains 600,000 digit images that come from a significantly harder real world problem compared to MNIST. The images lack any contrast normalization as well as contain overlapping digits and distracting features which makes them a much more difficult classification dataset compared to MNIST. We chose a batch size of 64 for MNIST and 32 for SVHN, a learning rate of $2e^{-4}$ and Sixty epochs. Patch size of $24 \times 24 \times 3$ was chosen as it supplied the best classification results. A weight variance of 0.01 was used.

On contrary to the medical datasets, we used an image-wise classification for natural images considering the relevant classes in the dataset. For the MNIST dataset, we obtained a classification accuracy of 0.995, which is comparable to the state of the art methods and to the baseline CapsNet architecture. For the SVHN, we were able to improve the classification accuracy of the baseline CapsNet, which is already pretty high, by 2.4%. Typical SVHN examples that have been correctly classified by our method can be found in Figure 5.

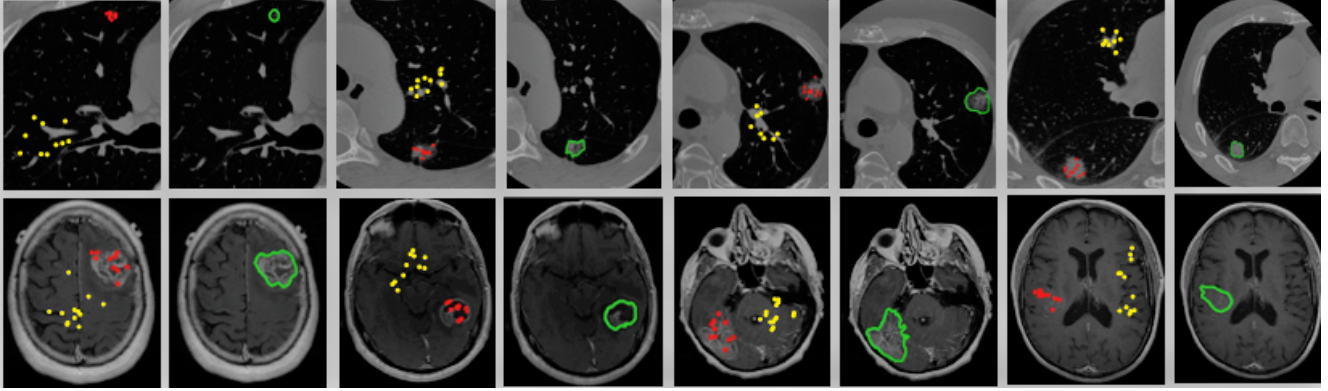


Figure 3. Classification of selected patches. Each pair of images (in the same row) represents the original image with the radiologist’s lesion annotation (green) and the processed image. Red - classified lesion patches, Yellow - classified normal patches. Upper row - CT Lung, Bottom row - MR Brain. In each image, we show classification results for the example colored patches.

Dataset	ResNet [3]	DenseNet [5]	Baseline CapsNet [16]	Our SACN
Liver (LiTS)	0.87 ± 0.02	0.87 ± 0.01	0.89 ± 0.03	0.9 ± 0.01
Brain	0.91 ± 0.03	0.91 ± 0.02	0.91 ± 0.02	0.94 ± 0.01
Lung	0.87 ± 0.08	0.88 ± 0.07	0.85 ± 0.12	0.92 ± 0.05

Table 1. Comparison (mean, std) of our proposed method with baseline CapsNet, DenseNet and ResNet architectures. The mean and the std values were calculated for different images in the subset. Wilcoxon paired test was calculated ($p < 0.001$ for all comparisons except for the comparison with the baseline CapsNet for the LiTS dataset). **The best results for each subset are bolded.**

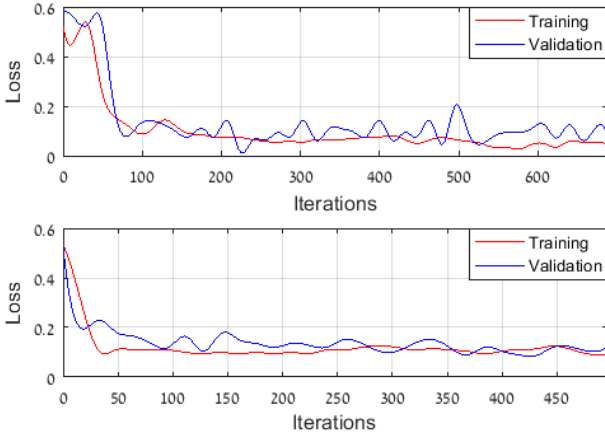


Figure 4. Calculated loss for training and for validation sets. Upper -MR Brain, Bottom - CT Lung.

6. Discussion and Conclusion

This paper introduces a novel architecture, called Self-Attention Capsule Networks (SACN), which was proposed to specifically improve the known

CapsNet architecture. The architecture utilizes the important key ideas of the CapsNet architecture, and boosts its performance by incorporating the Self-Attention mechanism as an integral layer within the CapsNet architecture. This architecture allows the model parameters, even in shallower layers, to be updated mostly based on image regions that are more relevant to a given task.

We conducted an extensive set of experiments, focusing on medical domain but presenting also preliminary analysis of natural images. For the medical subsets, which were part of highly diverse cohort, our proposed method significantly outperformed the baseline CapsNet. We also compared our technique with the advanced state of the art architectures - DenseNet-40 and ResNet. Our method was significantly better from these architectures as well. The better performance of our model is reflected in higher accuracy and lower standard deviation. Table 1 shows a key advantage of the proposed SACN over the baseline CapsNet, ResNet and DenseNet architectures - **when the cohort is more complex, the strength of the proposed method becomes more dominant.** This ob-

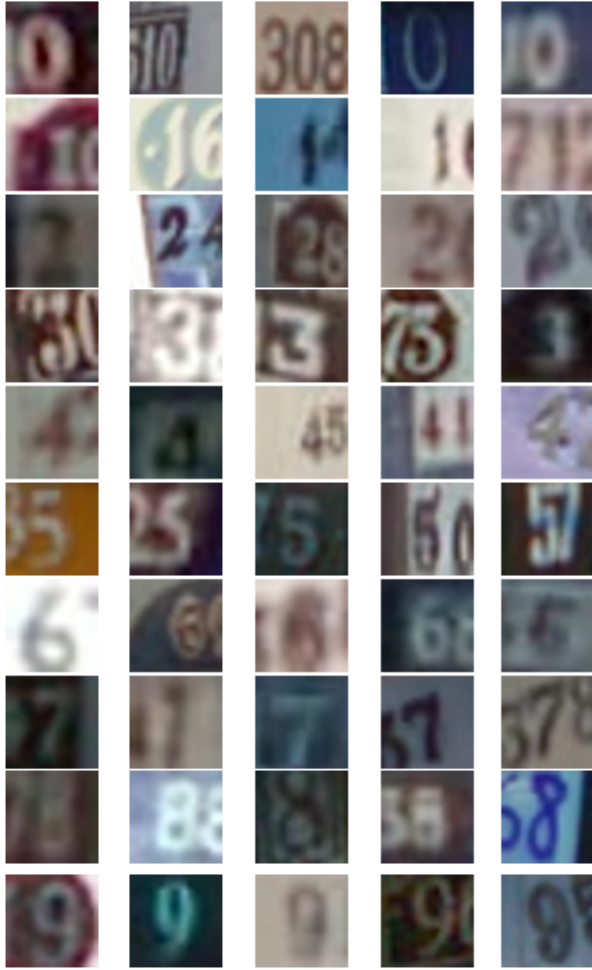


Figure 5. Representative examples for SVHN digits. Each column presents different variations of the same digit (0-9)

servation is well fitted to the known CapsNet limitation, which tries to account for everything present in an image, and for more complex images, where the background is too diverse, the CapsNet does not perform well in either case.

In regards to the public natural data that we analyzed, we were able to show classification accuracy that was comparable or better than the CapsNet or other state of the art methods that were reported in literature. **Implementing the model across substantially diverse domains/datasets shows its high generalization, robustness and classification capabilities.**

The baseline CapsNet is considered an expensive architecture in terms of computational load. For example, analyzing some of the datasets with the baseline CapsNet, resulted in Out of Memory errors on the GPU resources. In this work, we were able to sup-

ply classification accuracy that is significantly better than the baseline CapsNet architecture by using a relatively shallow CapsNet architecture and incorporating the attention module. **We were able to supply better results with less computational load that was reported in literature as a CapsNet cause for process shutdown.** Our architecture is powerful and has potential to be widely-used as it requires less computational resources.

Future work will include additional experiments, focusing on more complex natural and medical datasets. These experiments will be conducted for 2D and 3D data, using additional computational resources.

References

- [1] P. Afshar, A. Mohammadi, and K. N. Plataniotis. Brain tumor type classification via capsule networks. *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018.
- [2] A. Delige, A. Cioppa, and M. V. Droogenbroeck. Hit-net: a neural network with capsules embedded in a hit-or-miss layer, extended with hybrid data augmentation and ghost capsules. *arXiv preprint arXiv:1806.06519*, 2018.
- [3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [5] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.
- [6] Y. Jo, H. Cho, S. Y. Lee, G. Choi, G. D. Kim, H. L. Min, and Y. Park. Quantitative phase imaging and artificial intelligence: A review. *IEEE Journal of Selected Topics in Quantum Electronics*, 25:1–14, 2019.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.
- [8] H. Larochelle and G. E. Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Advances in Neural Information Processing Systems 23*, pages 1243–1251, 2010.
- [9] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, volume 3, 2014.
- [10] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [11] B. A. Olshausen, C. H. Anderson, and D. C. V. Essen. A neurobiological model of visual attention and invariant

pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13:4700–4719, 1993.

- [12] S. S. R. Phaye, A. Sikka, A. Dhall, and D. Bathula. Dense and diverse capsule networks: Making the capsules learn better, 2018.
- [13] W. Qian, Z. Jiaxing, S. Sen, and Z. Zheng. Attentional neural network: Feature selection using cognitive feedback. In *Advances in Neural Information Processing Systems 27*, pages 2033–2041, 2014.
- [14] D. P. Reichert, P. Seriès, and A. J. Storkey. A hierarchical generative model of recurrent object-based attention in the visual cortex. In *ICANN*, 2011.
- [15] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, pages 234–241, 2015.
- [16] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829*, 2017.
- [17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [18] Y. Tang, N. Srivastava, and R. Salakhutdinov. Learning generative models with visual attention. *arXiv preprint arXiv:1312.6110*, 2013.
- [19] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. *arXiv preprint arXiv:1711.07971*, 2017.
- [20] E. Xi, S. Bing, and Y. Jin. Capsule network performance on complex data. *CoRR*, abs/1712.03480, 2017.
- [21] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
- [22] M. Zhang, X. Li, M. Xu, and Q. Li. Image segmentation and classification for sickle cell disease using deformable u-net. 10 2017.