International Journal of Engineering

Journal Homepage: www.ije.ir

Learning Document Image Features With SqueezeNet Convolutional Neural Network

M. Hassanpour*, H. Malek

Department of Computer Science Engineering, Shahid Beheshti University, Tehran, Iran

PAPER INFO

Paper history:
Received 10 November 2017
Received in revised form 23 December 2017
Accepted 04 Januray 2018

Keywords:
Squeezenet
Convolutional Neural Network
Document Image Classification

ABSTRACT

The classification of various document image classes is considered an important step towards building a modern digital library or office automation system. Convolutional Neural Network (CNN) classifiers trained with backpropagation are considered to be the current state of the art model for this task. However, there are two major drawbacks for these classifiers: the huge computational power demand for training, and their very large number of weights. Previous successful attempts at learning document image features have been based on training very large CNNs. SqueezeNet is a CNN architecture that achieves accuracies comparable to other state of the art CNNs while containing up to 50 times less weights, but never before experimented on document image classification tasks. In this research we have taken a novel approach towards learning these document image features by training on a very small CNN network such as SqueezeNet. We show that an ImageNet pretrained SqueezeNet achieves an accuracy of approximately 75 percent over 10 classes on the Tobacco-3482 dataset, which is comparable to other state of the art CNN. We then visualize saliency maps of the gradient of our trained SqueezeNet's output to input, which shows that the network is able to learn meaningful features that are useful for document classification. Previous works in this field have made no emphasis on visualizing the learned document features. The importance of features such as the existence of handwritten text, document titles, text alignment and tabular structures in the extracted saliency maps, proves that the network does not overfit to redundant representations of the rather small Tobacco-3482 dataset, which contains only 3482 document images over 10 classes.

doi: *10.5829/ije.2020.33.07a.05*

1. INTRODUCTION

Archive offices usually contain a large corpus of various paper documents, and maintaining these documents can be a challenging issue. Although the trivial solution might be to convert these documents into digitally scanned document images and store them on disk with similar documents residing in the same folders on the file system, implementing this classification process through human labor can be extremely time consuming and frustrating. One good solution to this problem is to use of deep learning and Convolutional neural networks (CNN) for document image classification [1] since these networks have recently created a breakthrough in the field of image classification.

A CNN classifier trained with back propagation can be very powerful at learning rather complex visual concepts, but its often very large number of weights and depth means that it requires lots of computational power to converge on the learning task, and a considerable amount of memory for storing the weights. The large number of weights may not seem a problem when deploying on powerful machines with large memory, but can become a serious drawback when deploying on embedded systems that use much smaller memories. While one approach to this problem is to use techniques such as network pruning, quantization and huffman coding to reduce the model's size for inference [2], the strategy proposed by SqueezeNet is to reduce the number of convolution weights by squeezing convolutional feature maps using 1x1 filters while maintaining accuracy as high as possible by stacking 1x1 and 3x3 feature maps [3]. By doing so, the baseline SqueezeNet network reduces its size to less than 1 million weights while maintaining AlexNet [4] level accuracy on the ImageNet [5] dataset.

To justify the use of SqueezeNet for document image classification, we shall first make three important

*Corresponding Author Email: *mo.hassanpour@mail.sbu.ac.ir* (M. Hassanpour)

Please cite this article as: M. Hassanpour, H. Malek, Learning Document Image Features with SqueezeNet Convolutional Neural Network, International Journal of Engineering (IJE), IJE TRANSACTIONS A: Basics Vol. 33, No. 7, (July 2020) 1201-1207

notes here. First, features extracted from document images are indeed robust to compression [6], therefore the SqueezeNet architecture may be applied to squeeze feature maps harmlessly. Second, features learned from ImageNet are transferrable to the document image domain and SqueezeNet performs well on the ImageNet dataset [6]. Third, state of the art CNN architectures experimented by Afzal et al. [7] (networks such as Resnet [8], GoogLeNet [9], AlexNet [4] and VGG-19 [10]), all achieve similar accuracy rates on the Tobacco-3482 dataset. Considering the previous three notes mentioned, along with the experience that SqueezeNet achieves AlexNet level accuracy on ImageNet, reaches us to the hypothesis that it is possible for SqueezeNet to also achieve state of the art level accuracy on document image classification, while using much less weights. We will be further experimenting this hypothesis for the first time in this work.

Document image datasets share important structural similarities with generic image datasets such as ImageNet, while also having fundamental differences. The similarity between domains can easily be proven by showing that pretraining a CNN on ImageNet and then training on a document image classification dataset results in higher accuracy rates compared to random initialization of the weights [6, 7]. One important result of the differences between the two image domains is that augmentation techniques are not applicable to document images. While techniques such as image translation, rotation and scaling are successfully used to expand a dataset and reduce chances of overfitting, they cannot be applied to document images as these translations often disturb original document image features. We experimented this by training a Spatial Transformer Network (STN) which tries to learn the optimal linear augmentation on input images [11], but the final accuracy of the system degraded significantly.

The outline of the paper is as follows: in the second section we investigate related works done on the subject of document image classification. In the third section we first discuss a number of important strategies of the SqueezeNet architecture, and then describe the complete training procedure along with the choice of hyperparameters. In the fourth section, evaluation of the models classification accuracy is reported and saliency maps for the network input are analyzed. In the last section, the paper is concluded and possible future strategies accuracy are noted to improve.

2. RELATED WORKS

Over the years, various improvements and optimizations have been made to improve accuracy rates on the document image classification problem, most of which rely on CNN feature extractors. Here we will discuss a number of these efforts.

Kumar et al. [12] used a codebook of Speeded Up Robust Features (SURF) descriptors along with a random forest classifier with Support Vector Machine (SVM) to classify document images. This approach achieved a final accuracy of %43.27 on Tobacco-3482 [12], a dataset which was also first introduced in this paper and later converted into being one of the de-facto standards for benchmarking document image classification models.

Kang et al. [13] introduced one of the first attempts to train a document image classifier with CNN. They implemented a CNN with two convolution layers, max pooling and fully connected layers each, used the ReLU (Rectified Linear Unit) activation function along with dropout [14] to enhance the training process and then trained their network on two separate datasets: Tobacco-3482 and NIST tax-form. This method achieved better accuracy rates compared to previous state of the art approaches such as the hidden tree Markov model [15] and random forest classifier with SURF descriptors [12]. The main shortcoming of their method was that the designed CNN was too simple and therefore had a limited learning capacity. This problem was later overcome by proposing much deeper and complex CNNs such as Alexnet [4].

In [6], Harley et al. used an ensemble of five AlexNet networks with Principal Component Analysis (PCA) to present a new state of the art for document image classification. Their work made three main contributions to the field. First they showed that features extracted from document images were robust to compression. Second they showed that using an ensemble of networks does not greatly improve the classification results, and therefore it is unnecessary to enforce region specific feature learning for the task, under the consumption that enough training data is provided. Third, they showed that features extracted from other image classification tasks can be well transferred to the document image classification task. Pretraining a network on ImageNet and then training on the document image classification task improved the final accuracy rate. They also introduced the large scale RVL-CDIP dataset which contains 400 thousand document images over 16 classes.

In 2017, Afzal et al. performed an exhaustive investigation on various CNN architectures for document image classification [7]. The results showed that using architectures more complex than AlexNet for the task does not result in a noticeable increase of accuracy rate on the Tobacco-3482 and RVL-CDIP datasets. They also reduced the error rate on the Tobacco-3482 dataset by more than half by pretraining on the very large RVL-CDIP dataset. A drawback of their methods was that the trained networks still contained too many weights.

Tensmeyer and Martinez applied an unsupervised clustering based approach to cluster visually similar noise images [16]. They employed a 3-stage scalable clustering approach which first clusters a subset of the data, then these clusters are further split to create purer subclusters, and at last a classifier is trained on top to recreate the subclusters. Their method showed promising results on five various document datasets.

3. METHODOLOGY

This section will go through the architectural details of SqueezeNet, along with the detailed procedure of training this network on the Tobacco-3482 dataset. The relatively small number of weights in SqueezeNet simplifies tasks such as implementing it on embedded systems and downloading its weights over the Internet. It also considerably speeds up the training process.

3. 1. SqueezeNet Architecture

This architecture was proposed by Iandula et al. in 2016 [3]. The SqueezeNet network itself consists of building blocks named Fire modules, as shown in Figure 1. Each Fire module is basically a Squeeze Layer followed by an Expand layer, where the Squeeze layer is simply a layer of 1x1 convolution maps, and the expand layer is a combination of 1x1 and 3x3 maps. The number of feature maps in the Squeeze layer is made less than or equal to the number of expand layer feature maps, therefore performing some kind of compression on the extracted feature maps while also reducing the number of network weights. These Fire modules are then eventually stacked together to build the microarchitecture of the SqueezeNet model, as can be seen in Figure 2. An important hyperparameter of the Fire module is the Squeeze ratio, the number of Squeeze layer feature maps divided by the number of expand layer feature maps. Increasing this ratio up to $\frac{1}{2}$ generally increases the networks accuracy rates at the cost of increasing the network's size.

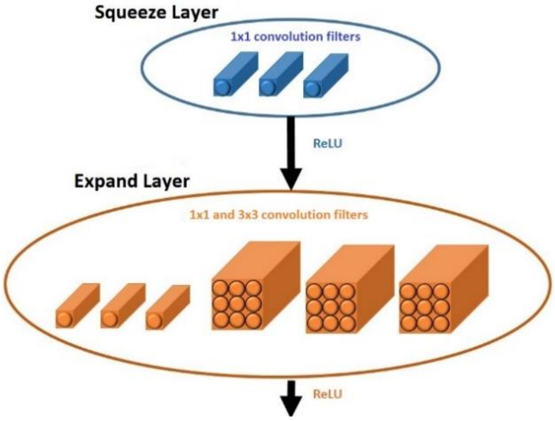

Figure 1. Overall structure of a Fire module in SqueezeNet model [3]

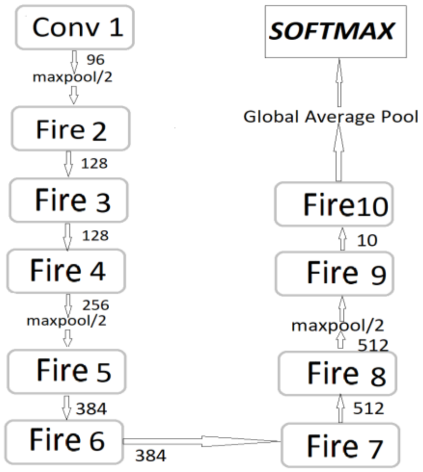

Figure 2. The macro architecture of a baseline SqueezeNet model [3]

SqueezeNet takes on three main strategies to improve the performance of traditional CNN networks. First, the majority of filters used in the network are 1x1 instead of 3x3; this greatly reduces the number of network weights. Second, it decreases the number of input channels to 3x3 filters. This approach also greatly reduces the number of network weights. Third, downsampling is performed later in the network on larger activation maps. The idea proposed is that a direct relationship exists between the size of activation maps of which downsampling is performed upon, and final classification accuracy results.

Another important strategy which greatly reduces the number of network weights, is the removal of the fully connected dense layers often used at the end of the network. This layer is replaced with a convolutional layer in which the number of output channels is equal to the number of data classes, and followed by a dropout layer and softmax activation function.

3. 2. SqueezeNet for Document Image Classification

According to classification results reported by [7], most deep CNN architectures achieve similar scores on the document image classification task both on small and large scale document image datasets separately. This gives us an intuitive understanding as of how much network complexity is correlated with classification accuracy on document image classification. It seems that raising the network's complexity higher than AlexNet level, does not have a significant effect on final classification accuracy.

As Harley et al. [6] showed, features extracted from document images are robust to compression. Therefore it is possible to effectively train SqueezeNet on this task, a network that continuously uses feature compression on every Squeeze layer to reduce the overall network size, while maintaining accuracy. This automatically makes SqueezeNet the superior choice for document image classification, as it achieves accuracy rates comparable to AlexNet while using as much as 50 times less weight [3].

3. 3. Training Procedure

To evaluate the performance of SqueezeNet on document image classification, we trained this model on the Tobacco-3482 dataset which contains 3482 high resolution document images over 10 classes. A number of sample documents from this dataset can be seen in Figure 3. In each class, 80 images were used for training, 20 for validation and the rest for testing model performance.

All of the dataset's grayscale images were repeated over three channels, resized to 224x224 and mean subtracted. After performing experiments, we found a minibatch size of 64 and learning rate of 10^{-4} to be the best option for training. The optimizer we used for learning network weights was Adam [17], with the hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. This optimizer has proven to perform well when training CNN networks. All network weights were pretrained on ImageNet, and training was performed over 150 epochs where each training epoch took approximately 6 seconds on a Tesla K80 GPU. Due to the small dataset size, training was performed five times with five different train/validate/test splits and the accuracies achieved from these splits were averaged to get the final accuracy.

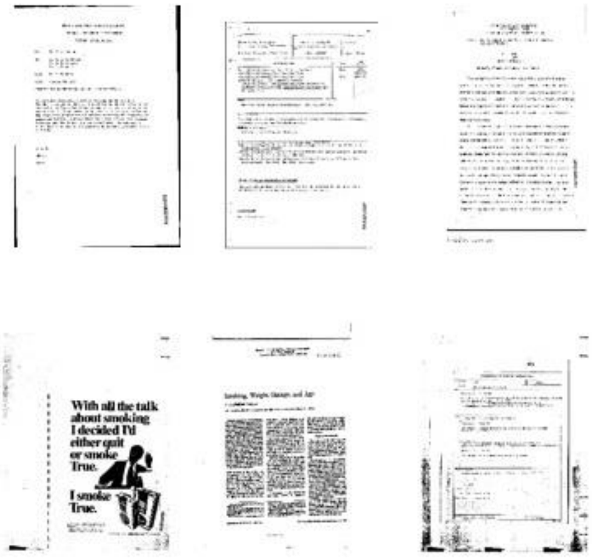
With all the talk about smoking I decided I'd either quit or smoke True.
I smoke True.

Figure 3. Sample documents from Tobacco-3482. The sample classes from top left to bottom right are memo, resume, note, advertisement, scientific and form

4. Evaluation

4. 1. Network Accuracy

Our results show that the SqueezeNet model performs well compared to other CNN models, and achieves an accuracy of %74.5 as can be seen in Table 1, which is only 1 percent less than the accuracy achieved by AlexNet. It is worth mentioning that the original SqueezeNet paper by Iandola et al. [3] introduced two strategies for further improving SqueezeNet classification rates at the cost of adding more network weights. The first approach is to increase the network squeeze ratio from $\frac{1}{4}$ up to $\frac{1}{2}$. This will make the network perform less compression on feature maps which in turn results in less data loss due to compression and higher accuracy rates on the data.

It is very likely that this little tweak will also boost accuracy rates for the document image classification task (we have not experimented this due to the unavailability of weights pretrained with ImageNet and the hardware limitations we had for training on ImageNet). The second approach is to add skip connections to each Fire module for increasing learning capacity. Due to the small size of the Tobacco-3482 dataset, these connections will likely harm accuracy results, as it was also shown by Afzal et al. that the classification rate achieved by a Resnet-50 network on Tobacco-3482 with no document pretraining, stands far behind other CNN models that do not contain these connections [7]. The only explanation for this phenomenon is that networks containing skip connections require larger amounts of training data to converge on a supervised image classification problem.

4. 2. Saliency Map Visualization

To show the effectiveness of SqueezeNet at learning document image features, we used saliency maps to visualize gradients of the network's output layer with respect to its input, as proposed posed by Simonyan et al. [18]. Simply put, we are trying to compute the gradient $\frac{\delta\ \textbf{output}}{\delta\ \textbf{input}}$, where the output is the network's softmax layer and the input is the input image we feed to the network.

TABLE 1. Comparison of classification results on Tobacco3482 with ImageNet pretraining between SqueezeNet using a Squeeze ratio of ¼ as experimented by us and other CNN architectures as experimented by Afzal et al. [7]

Network	Accuracy (%)	Num. Parameters
Resnet-50	67.93	25.6 M
GoogLeNet	72.98	4 M
SqueezeNet	74.40	0.8 M
AlexNet	75.73	62.3
VGG-16	77.52	138

Visualizing this gradient for each input image results in a saliency map which shows how the softmax output changes with respect to changes in the input. Brighter regions in this saliency map for each image indicate parts of the input that create higher activations on the output softmax neurons. This will let us know which features the network is paying more attention to, and whether it is learning a meaningful substructure or simply just overfitting on outlier features specific to the training data. This can become an issue especially when the dataset being trained on is small in size.

A number of the resulting saliency maps can be seen in Figure 4. Our visualization shows that the network is paying attention to a number of important features in documents, which shall be mentioned below.

- Document headers such as titles,
- The alignment of text paragraphs. Different document classes use different alignment methods for text paragraphs,
- Tables in documents. Particular document classes such as forms can be classified from other classes using this particular feature,

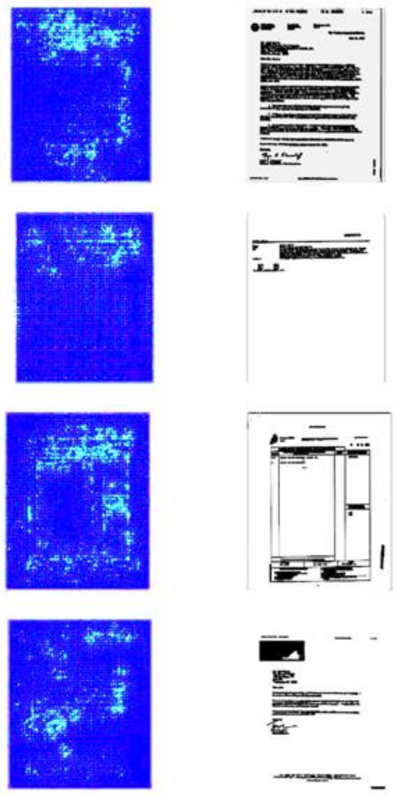

Figure 4. Four samples of extracted saliency maps (left column) and sample documents from the associated class (right column). The classes from top to bottom are Letter, Email, Form and Letter

- Handwriting on the document is also a crucial feature. Notes and Letters containing handwritten signatures could be classified from other classes through this feature.

The visualized maps can also help us understand which document features are more important with respect to each document class.

5. DISCUSSION

Most attempts in this field have been focused on the use of Convolutional Neural Networks for document image classification. Although most of these networks are very large and expensive to train, the SqueezeNet CNN is able to achieve state of the art level accuracy in document image classification with only 800 thousand weights, and the relatively small size of this network makes it suitable for deployment on cheaper embedded devices. Although, one drawback of CNN networks in document image classification is that they are not able to exploit the sequential structure of a document image and the correlation between its elements (these are in fact important features in this context because a document image is somewhat sequential in nature [20], i.e the header, body and footer of a document image are very likely related to each other). The inability mentioned above is because convolutional architectures are not able to encode the position of features, and feature maps (even different regions of a single feature map) are computed independently so correlations cannot be exploited. Due to these shortcomings, future work in this field could possibly involve using recurrent architectures to exploit these attributes. In addition, image enhancement and binarization techniques can be used to enhance document images for a better classification result [19, 20].

A more recent learning framework such as Contrastive Predictive Coding (CPC) [21] may also be employed in the future to learn document image representations in an unsupervised manner which requires much less labelled data compared to supervised methods. CPC learns representations by predicting the future in latent space using autoregresstive models. A probabilistic contrastive loss is used to induce this latent space, and negative sampling makes the model's training procedure tractable. The advantages of this method compared to CNN is that its future prediction in latent space could be able to exploit the correlation between various parts of a document image, and the accuracy achieved by this method on ImageNet is comparable to fully supervised methods, despite using 2 to 5 times less training labels. Still, implementing this method on small embedded hardware remains a challenge, while this is not the case for SqueezeNet CNN.

6. CONCLUSION

In this work we studied previous works done on document image classification, then proposed that SqueezeNet is a suitable CNN architecture for this task. This architecture was then trained on the Tobacco-3482 dataset. The accuracy achieved by a baseline SqueezeNet with only 800 thousand weights, was comparable to other state of the art CNN architectures with weights in the order of tens of millions. We then visualized our network's saliency maps and investigated document features which were learned by the network.

7. REFERENCES

1. Vincent, N. and Ogier, J.-M., "Shall deep learning be the mandatory future of document analysis problems?", ***Pattern Recognition***, Vol. 86, (2019), 281-289. https://doi.org/10.1016/j.patcog.2018.09.010
2. Han, S., Mao, H. and Dally, W.J., "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding", arXiv preprint arXiv:1510.00149, (2015).
3. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K., "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size", arXiv preprint arXiv:1602.07360, (2016).
4. Krizhevsky, A., Sutskever, I. and Hinton, G.E., "Imagenet classification with deep convolutional neural networks", in Advances in neural information processing systems., 1097-1105.
5. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database", in 2009 IEEE conference on computer vision and pattern recognition, Ieee., 248-255. DOI: 10.1109/CVPR.2009.5206848
6. Harley, A.W., Ufkes, A. and Derpanis, K.G., "Evaluation of deep convolutional nets for document image classification and retrieval", in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE. , 991-995. DOI: 10.1109/ICDAR.2015.7333910
7. Afzal, M.Z., Kölsch, A., Ahmed, S. and Liwicki, M., "Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification", in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE. Vol. 1, 883-888. DOI: 10.1109/ICDAR.2017.149
8. He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition., 770-778.
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going deeper with convolutions", in Proceedings of the IEEE conference on computer vision and pattern recognition., 1-9.
10. Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, (2014).
11. Jaderberg, M., Simonyan, K. and Zisserman, A., "Spatial transformer networks", in Advances in neural information processing systems., 2017-2025.
12. Kumar, J., Ye, P. and Doermann, D., "Structural similarity for document image classification and retrieval", ***Pattern Recognition Letters***, Vol. 43, No., (2014), 119-126. https://doi.org/10.1016/j.patrec.2013.10.030
13. Kang, L., Kumar, J., Ye, P., Li, Y. and Doermann, D., "Convolutional neural networks for document image classification", in 2014 22nd International Conference on Pattern Recognition, IEEE., 3168-3172. DOI: 10.1109/ICPR.2014.546
14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., "Dropout: A simple way to prevent neural networks from overfitting", ***The Journal of Machine Learning Research***, Vol. 15, No. 1, (2014), 1929-1958. DOI: 10.5555/2627435.2670313
15. Diligenti, M., Frasconi, P. and Gori, M., "Hidden tree markov models for document image classification", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 25, No. 4, (2003), 519-523. DOI: 10.1109/TPAMI.2003.1190578
16. Tensmeyer, C. and Martinez, T., "Confirm–clustering of noisy form images using robust matching", ***Pattern Recognition***, Vol. 87, (2019), 1-16. https://doi.org/10.1016/j.patcog.2018.10.004
17. Kingma, D.P. and Ba, J., "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, (2014).
18. Simonyan, K., Vedaldi, A. and Zisserman, A., "Deep inside convolutional networks: Visualising image classification models and saliency maps", arXiv preprint arXiv:1312.6034, (2013).
19. He, S. and Schomaker, L., "Deepotsu: Document enhancement and binarization using iterative deep learning", ***Pattern Recognition***, Vol. 91, (2019), 379-390. https://doi.org/10.1016/j.patcog.2019.01.025
20. Guo, J., He, C. and Wang, Y., "Fourth order indirect diffusion coupled with shock filter and source for text binarization", ***Signal Processing***, Vol. 171, (2020), 107478. https://doi.org/10.1016/j.sigpro.2020.107478
21. Oord, A.v.d., Li, Y. and Vinyals, O., "Representation learning with contrastive predictive coding", ***arXiv preprint*** arXiv:1807.03748, (2018).

Persian Abstract

چکیده

توانایی دسته‌بندی کردن اسناد اسکن شده از روی تصویر، قابلیتی است که می‌توان از آن در کتاب‌خانه‌های دیجیتال یا سیستم‌های اتوماسیون اداری به خوبی بهره برد. در همین راستا، شبکه‌های عصبی پیچشی آموزش داده شده با الگوریتم پس انتشار به عنوان روشی امروزی و قدرتمند برای دسته‌بندی تصاویر شناخته می‌شوند. امّا چنین شبکه‌هایی در حال حاضر دارای دو اشکال هستند: هزینه محاسباتی آموزش دادن آن‌ها بسیار سنگین است و معمولاً حافظه بسیار زیادی را به جهت داشتن تعداد بسیار زیادی پارامتر اشغال می‌کنند. با این وجود موفقیت‌های به دست آمده در مساله دسته‌بندی اسناد اسکن شده عموماً از طریق آموزش دادن شبکه‌های پیچشی بسیار بزرگ حاصل شده‌اند. شبکه پیچشی عصبی SqueezeNet، شبکه‌ای به نسبت کوچک امّا قدرتمند است که قادر است با وجود داشتن تعداد پنجاه برابر پارامتر کمتر نسبت به شبکه‌ای قدرتمند مانند AlexNet، در مساله دسته‌بندی تصاویر ImageNet، دقّتی هم‌اندازه با آن را کسب نماید، امّا تاکنون عملکرد آن در مساله دسته‌بندی اسناد تصویری ارزیابی نشده است. به همین جهت ما در این تحقیق تصمیم گرفته‌ایم تا عملکرد SqueezeNet را در دسته‌بندی اسناد اسکن شده مورد بررسی قرار دهیم. ما نشان می‌دهیم که یک شبکه SqueezeNet پیش‌آموزش یافته از روی مجموعه داده ImageNet دقتی تقریباً معادل با ۷۵ درصد را بر روی مجموعه داده Tobacco–۳٤۸۲ متشکل از ۱۰کلاس به دست می‌آورد، که دقتی قابل قیاس با سایر شبکه‌ای عصبی پیچشی می‌باشد. سپس گرادیان خروجی شبکه نسبت به تصاویر ورودی را با استفاده از نقشه برجستگی مورد بررسی قرار می‌دهیم و نشان می‌دهیم که شبکه ویژگی‌های سودمند و معنی‌داری را از روی تصاویر آموزش دیده است. این در حالیست که در تحقیقات گذشته تلاشی در راستای ظاهرسازی و یا تفسیر ویژگی‌های آموزش دیده به وسیله شبکه به چشم نمی‌خورد. ما با تحلیل این نقشه‌های برجستگی نشان می‌دهیم که شبکه SqueezeNet با وجود آموزش دیدن بر روی مجموعه داده‌ای به نسبت کوچک، ویژگی‌هایی مانند امضا، عنوان، جدول و نوع خاص هم‌ترازی متن را به خوبی تشخیص می‌دهد و از آن‌ها در راستای دسته‌بندی و تفکیک اسناد بهره می‌برد.