ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse115862019.pdf
https://doi.org/10.30534/ijatcse/2019/115862019

Convolutional Neural Network based on Transfer Learning for Medical Forms Classification

Adil Alharthi[1], Nouf Alzahrani[2], Ikram Moalla[3]
[1] Department of Computer Science, Faculty of CSIT, Albaha University, Saudi Arabia, afalharthi@bu.edu.sa
[2]Department of Information Technology, Faculty of CSIT, Albaha University, Saudi Arabia, noufalzahrani@bu.edu.sa
[3]Department of Computer Science, Faculty of CSIT, Albaha University, Saudi Arabia, ikram.moalla@ieee.org

ABSTRACT

Guaranteeing a high accuracy across image documents' understanding technologies is very difficult in commercial or even in experimental field. Our aim was to work in known context on documents of interest from health organisms and to hospitals' users in KSA, to offer solutions that help with the automatic or semi-automatic processing of health forms. We treated; particularly, the problem of medical forms understanding based on Convolutional Neural Networks for deep features extraction. We used transfer learning with two pre-trained architectures as AlexNet and GoogLeNet. We evaluated these features for the classification problem in different setups, using several public datasets; together with, using our proposed database with a KNN classifier. Our results showed that deep features are of high performance by reaching 94.2% of recognition for our contextual dataset.

Key words: Logical form structure recognition, deep features learning, Convolutional Neural Networks, transfer learning.

1. INTRODUCTION

Automatic forms processing is a very active field in the industrial world. In fact, faced with the huge mass of forms to be processed, automatic analysis becomes a necessity. Nevertheless, the performance of current systems varies greatly depending on the types of documents processed. In medical domain, especially, in health organisms, several documents like patient's registrations, medical reports, sick leave reports, analysis reports, etc. can be find. Understanding and extracting information from such structured or semi-structured documents with image processing and AI techniques can assist and help health organisms to treat and automatically use all pertinent data from diverse documents. Many applications in document image understanding such as, document type classification is still a challenging task.

Document type classification is to assign a document image to a pre-stored template. The task is to assign a document to one or more classes or categories. The target document model can be expressed by an XML schema. This schema can be represented for example in the form of an XML Schema, a DTD or a RelaxNG schema.

In general, a document can be composed by two types of structures: physical and logical (see figure1). The physical or visual structure of a document represents the visual form in which it appears. The visual units are identified by indices of a typographic or dispositional nature. The logical structure of a document is defined as an abstract level ordering the document into elementary logical units and complex logical units. These units are said to be logical, because they participate in the comprehension of the text, independent of their propositional content.

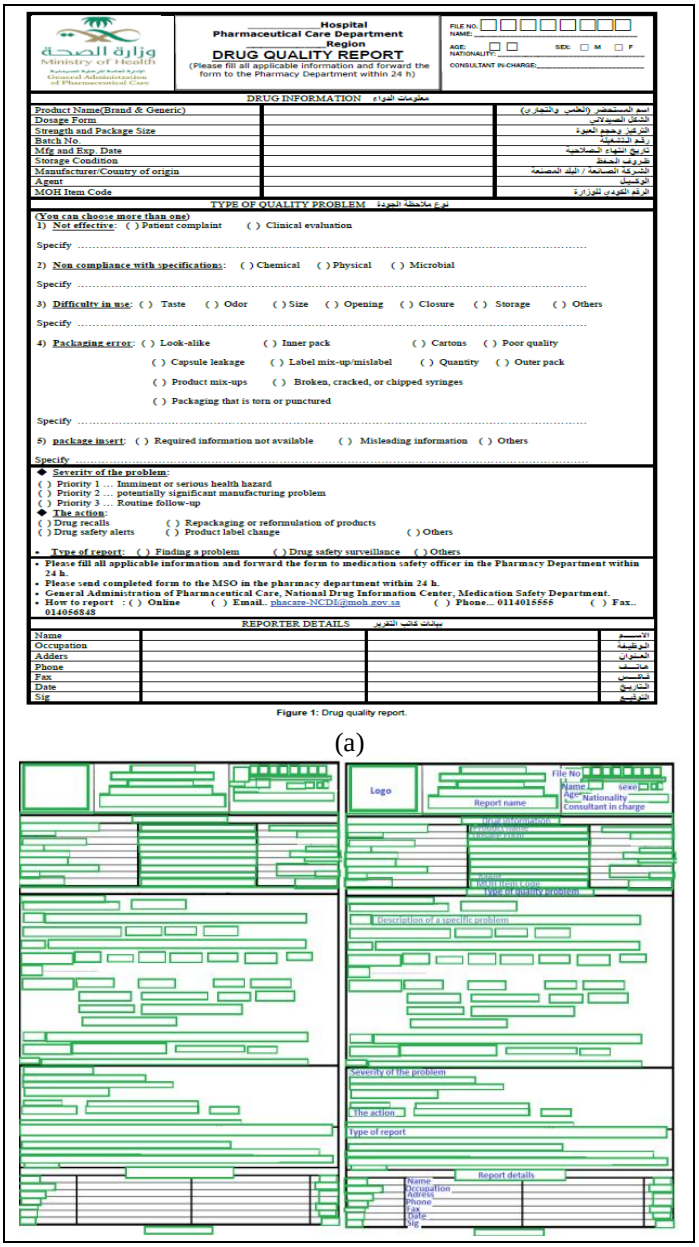
In our work, we will study hierarchical structures from their visual properties and which are inferred from logical and discursive elements.

The great diversity of hierarchical structures induces a great variety in physical and logical patterns to be considered in each application. For example, in scientific articles, the typical structure includes a title followed by an abstract, then an introduction and eventually some key words. For medical documents, the structure is in general coroneted by a header containing the logos of hospitals, health ministers, followed by the person info bloc and then by general health information, etc.

The essential idea of the proposed approach is to classify the image of each administrative medical document in a global way. It should be possible to classify the form of a request document image using low-level features. Recognition of the physical structure should infer the higher level semantic information and therefore lead to recognition of the logical structure of the document. This avoids the necessary passage by the heavy step of segmenting into physical blocks of the document. The success of this approach therefore requires the use of highly structured documents.

These could guarantee that the same relevant information should always appear in the same place on the page. For example, in a hospital application, the patient's name always appears in the same box and place of the document. In contrast, in unstructured documents, information can appear in unexpected places on the document.

The objective of this work is to propose an approach to determine the physical class of a structured document to deduce its important regions. These regions will be stored after recognition of their words, in centralized databases. This will allow avoid losing hundreds of hours of tedious and manual work to the various KSA health organizations. It will also have a significant gain in terms of medical follow-up according to accurate and quickly updated information.



(a)

(b) (c)

Figure 1: Example of medical form structures. (a) Medical form, (b) Physical structure, (c) Logical structure

2. STATE OF THE ART

Recently, the recognition of the form structure has become a simpler mission, in a known context, having a fixed number of standard document classes to identify. The disposition of the blocks of texts, images, graphs, digits, ... appearing in the same form remains the same by filling it by one person or another; the semantic content of an "address" field will always contain an address, the contents of a "blood analyzes" field are always characterized according to the same topics and therefore always appear in the same locations in the document, etc.

Indeed, the recognition of the document infers the recognition of the logical structure of the same class of document. In this context, the author in [1] used kNearest Neighbours (kNN) and MLP classifier for document images classifications.

In [2], the author used a directed classifier for treated classes and visual similarity of document form structure for their classification. In [3], the author proposed a generative classifier model to recognize the document classes and multi-scale runs length histograms for the physical representation step.

In [4], the author used Deep Convolutional Neural Networks to learn successfully the documents structures.

In [5], a DCNN structure was also used to classify the 10 classes of document images from Tobacco3482 dataset.

In [6],[7] and [8], in order to improve the recognition rate, authors used pre-training DCNN on ImageNet dataset before their test on Tobacco3482 dataset.

In [9], AlexNet architecture was used on RVLCDIP and ANDOC datasets. Many variants were tested like the images sizes, the protection of the aspect ratio, layout features and the size of the datasets. In [10], the author compare the performances of GoogLeNet and AlexNet models on RVL-CDIP and Tobacco3482 datasets by using of transfer Learning concept.

3. METHODOLOGY

We have chosen to use Convolutional Neural Networks (CNNs) to classify the different classes of forms of medical documents. Recently, the CNNs have successfully been implemented in the medical field to assist the diagnosis of the disease. In addition, we want to demonstrate their effectiveness in the understanding of medical documents. CNN learning is based on sufficient information to learn layered hierarchical characteristics of the images. Indeed, in the medical field, it's difficult to find large databases as well annotated as ImageNet.

For the problem of medical document classification, as in other kind of image understanding applications, we can train the CNN from scratch or use a pre-trained CNN. The construction of a deep CNN, requires a lot of data and requires wide computation resources for the learning stages, while a pre-trained classifier is done in an unsupervised way. Only fine-tuning is done with even a small [11, 12] database according to the context of the application. That's why we chose to use a trained CNN, and do our fine-tuning with our relatively small database which we will present later in this paper.

3.1 Convolution Neural Network

A few years back, traditional neural network used to learn and classify basic recognition tasks. However, recently with the availability of large datasets pushed the need to use deep Convolutional Neural Networks (CNNs) algorithm. Such an algorithm had always been the go-to model for various tasks classifications and recognitions.

The deep CNNs automatically generate multiple-representations from the original data by utilizing different feature extraction approaches, and then learning the multiple-representations at different hidden layers. The deep CNNs can be used for auto correlated data, image recognition, image classification, image segmentation etc.

Deep CNNs consist of various layers' types including: (1) Convolutional Layer which detects local conjunctions of features from the previous layer and mapping their appearance to a feature map, (2) Non-Linearity Layer produces its output (namely activation map) by applying an activation function on the feature map previously generated by the convolutional layer, (3) Rectification Layer is responsible for performing element-wise absolute value transformation on summed weighted input volume from the node into the activation of the node, (4) Rectified Linear Units (ReLUs) are a special implementation piecewise linear function and rectification layers in CNNs to gain directly the output in positive case, otherwise, it will produce zero. (5) Pooling or Down sampling layer is responsible to address the sensitivity of feature location, referred to as "local translation invariance.", in the input tasks by reducing the size of the samples and summarizing the presence of features (via average pooling and max pooling) in activation maps, (6) fully connected layers are practically multilayer perceptron's that aims to recognize and classify the tasks by breaking them down into features, and analyzing them independently. The result of mapping the activation volume from the combination of previous different layers result in fully connected neural network and a class probability distribution to drive the final recognition and classification decision, (7) Dropout Layer is used to address the problem of overfitting in the fully connected layers, also it can be applied after pooling layer (e.g. max-pooling) to create noise augmentation in the input tasks. Moreover, it uses activation function (namely Softmax function) to compute the probabilities of each class over all possible target classes in the required tasks. Then the probabilities matrices will be used for determining the actual class for the given inputs in in the required tasks.

The essential advantage of applying CNNs for image processing and classification is their capability to identify certain of features by using smaller portion of the image rather than the entire image at once. In addition, the CNNs have depth architectures which can enhance the representational capacity and quality in comparison to shallow architectures [13].

For a complex problem, CNNs have a great capability to efficiently learn from different representations and abstraction by stacking the outputs of multiple linear and non-linear processing units in each layer.

Deep CNNs architecture has also tremendous performance capability to classify and recognize tasks within hundreds of categories or examples within hundreds of class labels [14]. In recent years, the utilization of the concept of Transfer Learning (TL) on low-level and high-level features of a task to construct generic recognition systems [14][15].

Several deep convolutional neural network architectures have been proposed such as AlexNet [16], ZFNet [17], VGG [18], GoogleNet [19], Residual Net [20], DenseNet [21] and FCN [22].

We chose to evaluate our work by using the two most popular CNNs Architectures: AlexNet and GoogLeNet. The main characteristics of these two networks are presented below.

3.2 AlexNet

AlexNet was the winner of the ILSVRC image classification competition in 2012. Alex Krizhevsky et al. [16], created a neural network architecture called 'AlexNet'. It was the first real "deep" network that has had a great impact on the field of artificial intelligent encompasses all industrial and non-industrial applications including: machine learning, deep learning, and machine vision.

AlexNet showed his powerful by the high accuracies achieving on very challenging datasets. The architecture of AlexNet contains eight learned layers, eight-layer convolutional neural network and 3 three-layer fully-connected neural network (see figure 2).

The main characteristics make AlexNet special are: (i) ReLU activations (Rectified Linear Units) are applied after every convolutional layers, (ii) Multiple GPUs to cut the taring time by dividing the training datasets and traing the bigger model on multi-GPU, and (iii) Overlapping Pooling to reduce the error rate and avoid overfitting. Nevertheless, to reduce the overfitting, which is caused by larger number of parameters, AlexNet employed two methods: (1) Data Augmentation to create different representations for same data (e.g image translations / horizontal reflections), and (2) Dropout is applied to turn-off neurons with 50% probability for the model's convergence and also acceleration speed.

The image size in the following architecture chart should be 227x227x3. (see figure 2). The first convolutional network incorporates 96 kernels of filter size 11×11 for the construction and with a stride 4 and padding 0 to halve the resolution of the grid. The second convolutional network incorporates 256 kernels of filer size 5×5 for the construction and with a stride 1 and padding 2. The third convolutional network incorporates 384 kernels of filer size 3×3 for the construction and with a stride 1 and padding 1. The fourth convolutional network has the same settings as the third convolutional network. The fifth convolutional network incorporates 256 kernels of filter size 3×3 for the construction and with a stride 1 and padding of 1. The max-pooling convolutional network has the same setting and structures of convolutional network #1, #2 and #5, where the average max-pooling size is 3×3 and the stride is 2. In particular, there are 4096 neurons in the fully connected convolutional network #1 and #2. The third last connected convolutional network has N outputs, where N represents the number of expected classes.

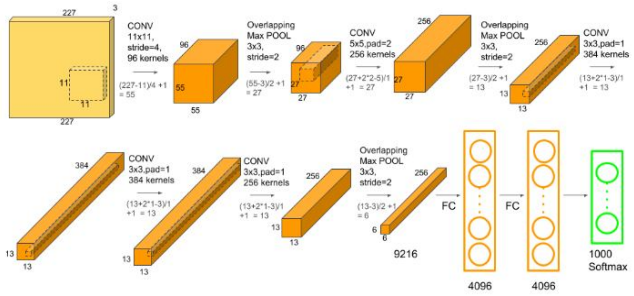


Figure 2: The AlexNet architecture [16]

3.3 GoogleNet

The winner of ILSVRC 2014 and the GoogLeNet architecture is also known as Inception Module. It was proposed by Szegedy et al[19].

This architecture consists of 22 layers in deep. The main advantage of the GoogLeNet architecture in comparison to AlexNet is the reduction of the computing resources inside the network by decreasing the number of the variables by 15 times from 60 million to 4 million. GoogLeNet incorporates 1×1 filter convolution inside the inception modules to act as a dimension reduction module, and to reduce the computation bottleneck. By doing so, the 1×1 filter convolution can increase the depth and width of the network. In order to avoid the issue of the fully connected layers, the network in GoogLeNet utilizes the global average pooling instead. Nevertheless, inception module of the network in GoogLeNet can have various types and size of convolutions, and even it can stack the outputs for same input.

In AlexNet architecture the size of convolutional layers are fixed, while in GoogLeNet the convolutional layers are 1×1 convolution, 3×3 convolution, 5×5 convolution, and 3×3 max pooling are done altogether for the previous input, and stack together again at output. At the end of network, the global average pooling is used nearly by averaging each feature map from 7×7 to 1×1.

The important idea behind the use of GoogLeNet is to widen the network (go in width) instead of increasing the number of layers (go in depth). This will increase the size of the network which can be very complex. In order to remedy to this problem, a transition from fully connected to sparsely connected architectures is used. This is presented in GoogLeNet by the use of the inception module (Figure 3). It uses a parallel combination of 1x1, 3x3, and 5x5. 1x1 convolutions to estimate reductions before performing the more computationally expensive convolutions 5x5. GoogLeNet architecture uses 9 inception modules, consists of 22 layers deep. There are also 5 pooling layers (four max pooling layers and one average pooling layer). Average pooling with 5x5 filter size and stride 3 is used before the classifier.

When combined multiple inceptions modules, the obtained deeper network, allow obtaining high accuracy.

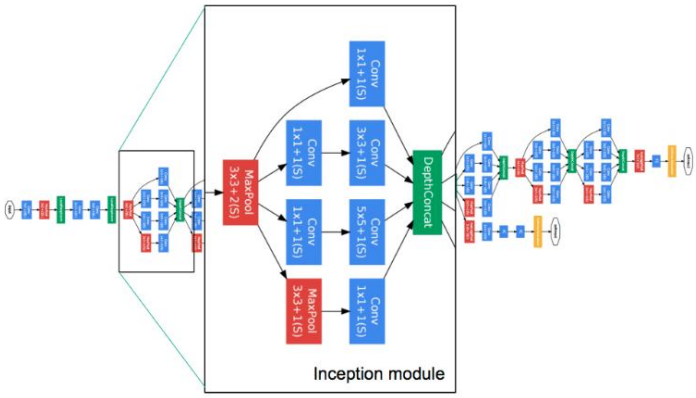


Figure 3: Overview of the GoogLeNet Architecture [19]

3.4 Training

CNNs is an efficient recognition and classification algorithm on large scale datasets. This is due to its ability of joint feature and classifier learning at the same time. However, the accuracy and efficiency of the CNNs algorithms specifically reliant on the availability of large training datasets. The concept of transfer learning is an alternative to overcome this problem.

This process is based on the idea that it is much faster and easier to improve the classification of the pre-trained network on new collection of images by using a smaller number of training images, rather than generating a new network classification model. At test time, the full architecture is used to predict the class label.

In this work, we used the transfer learning concept to train AlexNet and GoogLeNet architectures. So, we used the whole images of the publically available RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset (see figure 4). The use of a very large labeled dataset is beneficial for the supervised training stage to initialize networks weights. The obtained models can be considered as document feature extractors.

4. EVALUATION FRAMEWORK

4.1 Datasets

To evaluate the performance and the effectiveness of the features generated by the two networks AlexNet and GoogLeNet in the previous section, we carried out experiments on four different datasets: RVL-CDIP [23], NIST2 [24], MARG3 [25], and CLEF -IP4.

Table 1: Details of the benchmark datasets

Dataset	Category	# of Images	# of classes
RVL-CDIP	Document	400000	16
NIST	form	5590	12
MARG	layout	1553	9
CLEF-IP	patent image	38081	9
HEL-ADU	document	622	10

We also confirm our conclusions on our constructed datasets, which we refer to as HEL-ADU. Short descriptions of the 5 used datasets are summarized in table 1.

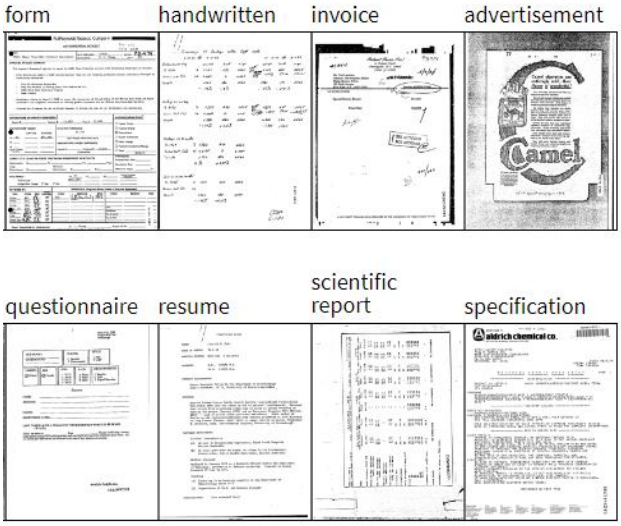


Figure 4: Illustration of the RVL-CDIP Dataset [23]

Our constructed dataset that we named HEL-ADU regroups document images from KSA hospitals and from many other health organisms. It contains 622 scanned documents from 10 different medical document categories such as medical or analysis report, medicine prescription, sick leave reports etc. (see figure 5).

Figure 5: Examples of HEL-ADU Dataset

4.2 Experimental Results

In this experiment a non-parametric k-nearest neighbor's algorithm (k-NN) has been chosen for evaluation and the classification the targeted task. The input for the k-nearest neighbor's algorithm is the k closest training samples in the feature space, while the output for the such an algorithm is a class membership. A document is classified by a plurality vote of its neighbors, with the model being assigned to the corresponding class of its k-nearest neighbor (where K=1). The results are reported in table 2.

Table 2: Description of the evaluation datasets: knn-based classification accuracy for the deep learning approaches

Datasets	MARG	CLEF-IP	NIST	HEL-ADU
CNN-AlexNet	65.7	75.0	100	93.8
CNN-GoogLeNet	75.8	75.8	100	94.2

The results obtained on the NIST database are the best and reach the 100% of good classification. However, classifying forms types from the NIST database was not a difficult task because the templates used prescribe visible high intra-class variability and reduced inter-class variability. But this confirms the effectiveness of our features and we can try them for the dataset of our application. In MARG dataset, the results reach only 75.8% with the CNN-GoogleNet architecture. This decrease in performance compared to those obtained with NIST database is due to the fact that MARG dataset presents visually similar specification aspects of the document layout which can belong to different classes of document. We have encountered the same problem of intra-class variation with 4 classes out of 10 of the HEL-ADU database, which has caused some confusion. The CLEF-IP presents images of documents with a high variable size and aspect ratio. As one of the drawbacks of the CNNs with image size is to have a fixed input size of (224×224). Such an issue in the input image for CNNs can force us to deal with limited input images in terms of both the aspect ratio and scale. To address this issue, we had fit the input images to the fixed size

via warping. This might have negative effects on CNNs performances due to undesirable geometrics distortion, which lead in obtaining only 75.8% of good recognition rate with CNN-GoogleNet architecture.

HEL-ADU dataset contains 10 categories of medical forms. The result obtained by transfer learning via the AlexNet architecture allows a success rate of 73.8%. This result was slightly improved by GoogleNet by obtaining 94.2% of good recognition rate. Six classes among the 10 are well distinguished. The four others present confusion cases.

5. CONCLUSION

In this paper, we have presented a novel classification approach to offer solutions that help with the automatic and semi-automatic processing of health forms in KSA's hospitals. The proposed work represents deep features based on Convolutional Neural Networks. It uses transfer learning from pre-trained models via fine-tuning. The best accuracy is achieved when fine-tuning with pre-trained googLeNet network. The obtained results are effective on the 10 medical form classes. We can extend our work on other forms to cover as many medical administrative documents as possible.

.

ACKNOWLEDGEMENT

This research is a part of a project entitled "SMART-PHS: A Smart Card for Automated Patient Health Service: Medical Documents Analysis, Recognition and Data-storage". This project was funded by the Deanship of Scientific Research, Albaha University, KSA (Grant No. 1439/33). The assistance of the deanship is gratefully acknowledged.

REFERENCES

1. P. Héroux, S. Diana, A. Ribert, and E. Trupin, "**Classification method study for automatic form class identification**", *Proc. of the 14th Int. Conf. on Pattern Recognition (PR).*, pp. 926–929, 1998
2. C. Shin, D. Doermann, and A. Rosenfeld, "**Classification of document pages using structure-based features**," Int. *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 3, no. 4, pp. 232–247, 2001.
   https://doi.org/10.1007/PL00013566
3. A. Gordo, F. Perronnin, and E. Valveny, "**Large-scale document image retrieval and classification with runlength histograms and binary embeddings**," in *Pattern Recognition (PR)*, vol. 46, no. 7, pp. 1898–1905, 2013.
   https://doi.org/10.1016/j.patcog.2012.12.004
4. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "**Gradient-based learning applied to document recognition**," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
   https://doi.org/10.1109/5.726791
5. L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, "**Convolutional neural networks for document image classification**," in *International Conference on Pattern Recognition (ICPR)*, pp. 3168–3172, 2014.
6. M. Z. Afzal, S. Capobianco, M. I. Malik., S. Marinai, T. M. Breuel, A. Dengel, and M. Liwicki, "**Convolutional neural networks for document image classification**," *Proc. of the 13th Int. Conf. on Document Analysis and Recognition (ICDAR)*, pp. 1111–1115, 2015.
7. S. J. Pan and Q. Yang, "**A survey on transfer learning**," *IEEE Trans. on Data & Knowledge Engineering (DKE)*, vol. 22, no. 10, pp. 1345–1359, 2010.
   https://doi.org/10.1109/TKDE.2009.191
8. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "**Imagenet: A large-scale hierarchical image database**," *Proc. of IEEE in Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.
   https://doi.org/10.1109/CVPR.2009.5206848
9. C. Tensmeyer and T. Martinez, "Analysis of convolutional neural networks for document image classification," *arXiv preprint arXiv:1708.03273*, 2017.
10. M. Z. Afzal, A. K ¨olsch, S. Ahmed, and M. Liwicki, "**Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification**," *arXiv preprint arXiv:1704.03557*, 2017.
11. Matthew D. Zeiler and Rob Fergus. "**Visualizing and Understanding Convolutional Networks**". In: *Computer Vision (ECCV)* European Conference, Zurich, Switzerland, september, pp. 818–833, 2014.
12. Maxime Oquab et al. "**Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks**". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Columbus, OH, USA, June 23-28, pp. 1717–1724, 2014.
    https://doi.org/10.1109/CVPR.2014.222
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "**ImageNet Classification with Deep Convolutional Neural Networks**," in *Neural Information Processing Systems (NIPS)*, pp. 1–9, 2012.
14. A. S. Qureshi and A. Khan, "**Adaptive Transfer Learning in Deep Neural Networks: Wind Power Prediction using Knowledge Transfer from Region to Region and Between Different Task Domains**," *arXiv Prepr. arXiv1810.12611*, 2018.
15. Qiang Yang, S. J. Pan, and Q. Yang, "**A Survey on Transfer Learning**," in *IEEE Translactions on Data and Knowledge Engineering (DKE)*, vol. 1, no. 10, pp. 1–15, 2008.
16. A. Krizhevsky, I. Sutskever, and G. Hinton. "**ImageNet classification with deep convolutional neural networks**", In *Neural Information Processing Systems (NIPS)*, pp 1097-1105, 2012.
17. M.D. Zeiler, R. Fergus, "**Visualizing and understanding convolutional networks**", *European Computer Vision conference (CVC), Springer*, pp. 818-833, 2014.
18. K. Simonyan, A. Zisserman, "Very deep convolutional networks for largescale image recognition", *arXiv preprint, 1409.1556*, 2015.
19. C. Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan,

Vincent Vanhoucke, and Andrew Rabinovich, "**Going deeper with convolutions**", in *Computer Vision and Pattern Recognition (CVPR)*, pages 1-9, 2015. https://doi.org/10.1109/CVPR.2015.7298594

20. K. He, X. Zhang, S. Ren, J. Sun, "**Identity mappings in deep residual networks**", *European Computer Vision conference (CVC), Springer*, pp. 630-645, 2016.
21. G. Huang, Z. Liu, L. VanDer Maaten, K.Q. Weinberger, "**Densely connected convolutional networks**", *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700-4708, 2017. https://doi.org/10.1109/CVPR.2017.243
22. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "**Imagenet large scale visual recognition challenge**", in *International Journal of Computer Vision (IJCV)*, 115(3):211-252, 2015.
23. A. W. Harley, A. Ufkes, K. G. Derpanis, "**Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval**", in *International Conference on Document Analysis and Recognition (ICDAR)*, 2015. https://doi.org/10.1109/ICDAR.2015.7333910
24. **The NIST Structured Forms Database** (NIST Special Database 2) (2010) https://www.nist.gov/srd/nist-special-database-2. Last visited Jan 2019.
25. **The Medical Article Records Groundtruth Dataset** (2003) https://ceb.nlm.nih.gov/inactive-communications-engineering-branch-projects/medical-article-records-groundtruth-marg/. Last visited Jan 2019.