Multimodal Deep Neural Networks for Banking Document Classification

Deniz Engin, Erdem Emekligil, Mehmet Yasin Akpınar, Berke Oral, Seçil Arslan

R&D and Special Projects Department, Yapi Kredi Technology
Istanbul, Turkey
Email: {deniz.engin, erdem.emekligil, mehmetyasin.akpinar, berke.oral, secil.arslan}@ykteknoloji.com.tr

***Abstract*—In this paper, we introduce multimodal deep neural networks to classify petition based Turkish banking customer order documents. These petition based documents are commonly free-formatted texts, which are created by customers, but some of them do have a specific format. According to the structure of the banking documents, some documents containing tables and specific forms are convenient for visual representation, while some documents consisting of free-formatted text are convenient for textual features. Since the texts of these documents are obtained via Optic Character Recognition technology which does not work well on handwritten, noisy, and low-resolution image documents, text classification methods can fail on them. Therefore, our proposed deep learning architectures utilize both vision and text modalities to extract information from different types of documents. We conduct our experiments on our Turkish banking documents. Our experiments indicate that combining visual and textual modalities results in better recognition of documents compared to text or vision classification models.**

Keywords–Multimodal Deep Learning; Document Classification.

I. INTRODUCTION

Every bank puts different channels *e.g.*, fax, email, scanner into service to receive its customers' orders for banking transactions. More than 6.5 million transactions are completed in a medium-large scale bank in Turkey received from these channels yearly [1]. Customers share their orders in free-formatted petitions to declare money transfer, tax payments, salary payments which leads to more than 60 various banking process types. Those orders received in image format are mostly multi-page and low in resolution. In traditional process workflow, when the customer order is received, a back-office operator views the order and investigates all pages of the document to detect the process types in the document in order to split (if needed) and direct the order to the correct back-office data entry team. Accordingly, the classification of the customer order is one of the most time-consuming and human workforce required steps of the overall workflow management. Therefore, document classification systems play a crucial role in the banking domain. An overview of our document classification flow can be seen in Figure 1.

Document classification methods can be based on image classification, text classification on obtained text from Optical Character Recognition (OCR) of images, and multimodal classification. In recent works, several deep learning based methods have been focused on document classification by using only the image of documents [2]-[5]. In [2], Convolutional Neural Networks (CNNs) are used as a feature extractor on a specific-region in a document. Also, these region-based

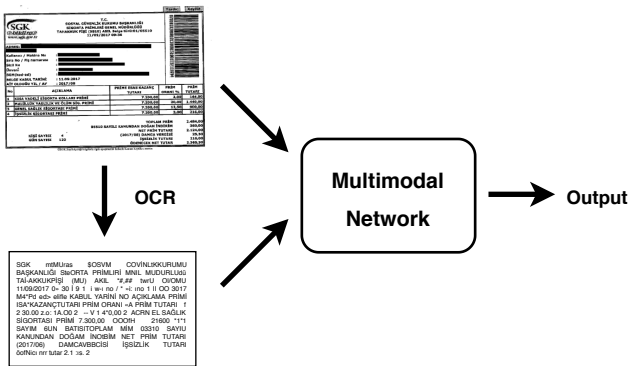

Figure 1. An overview of document classification flow.

features are concatenated before classification. Data augmentation has been applied and CNN architectures have been proposed in [3]. Several well-known CNN architectures, *e.g.*, AlexNet [6], VGG-16 [7], GoogLeNet [8], Resnet-50 [9] have been investigated for document classification by using transfer learning in [4]. Transfer learning from VGG-16 network pre-trained on ImageNet dataset [10] is utilized for region-based document classification in [5]. These proposed methods have been trained on the Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) dataset [2], which has 16 classes, such as letter, form, email, etc. While classes of this dataset are distinguishable from each other and images of inter-class are consistent, banking order documents have different structures, which can be seen in Figure 2.

Different structures can be categorized as follows:

- free-formatted texts,
- large tables,
- customer arranged forms which are unique for each customer,
- forms that are pre-defined by certain organizations.

Due to structural variation of documents in our dataset, only the vision method is not sufficient for our classification task. Similarly, documents belonging to the same class can be a form or a free-formatted text. Sample documents for this problem can be shown in Figure 3. To overcome these difficulties, we decide to utilize textual information obtained via OCR besides visual information. After the text is obtained from documents, this problem becomes a text classification task. The main idea behind the recent methods is to capture the document representations from characters, words or sentences, by using CNN or Long Short-Term Memory (LSTM), to

 ISBN: 978-1-61208-731-3

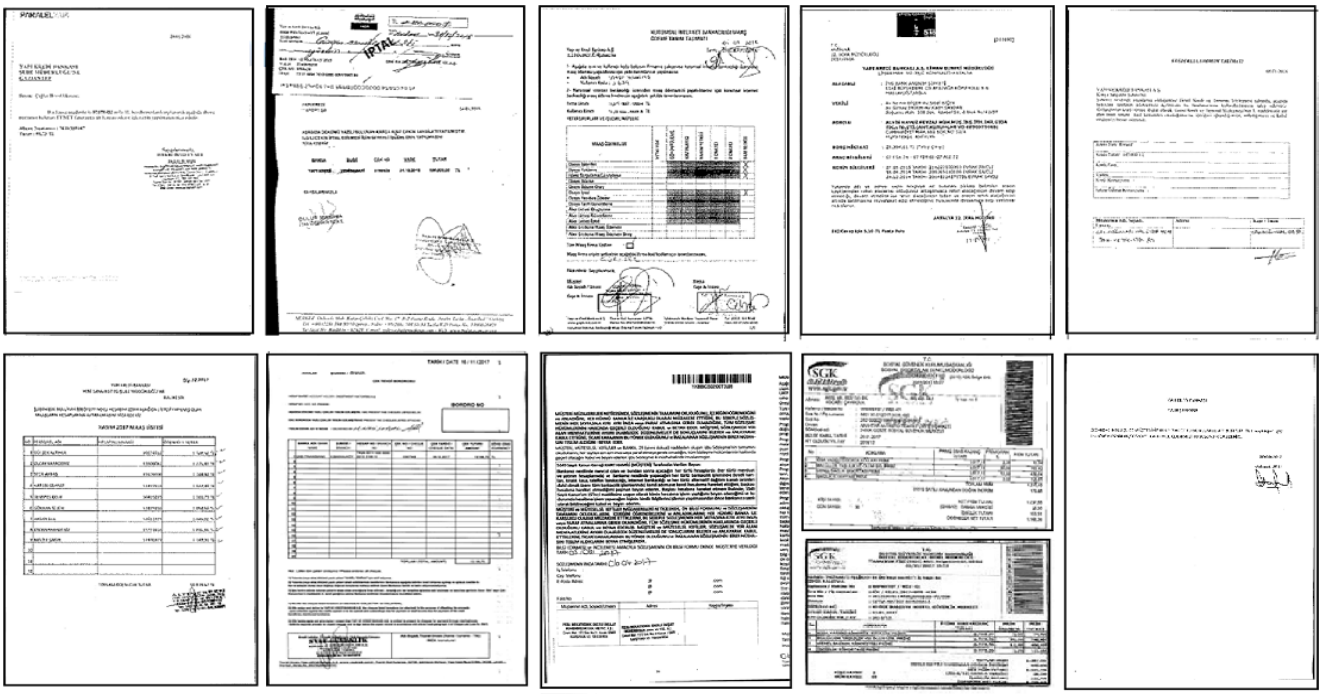

Figure 2. Sample images from the Turkish Banking Documents.

classify documents using these representations. Using CNNs to extract information from characters or word embeddings in order to classify texts have been proposed in [11][12]. In addition, [13] and [14] suggested to use embeddings with sequential models like Gated Recurrent Units (GRUs) or LSTMs.

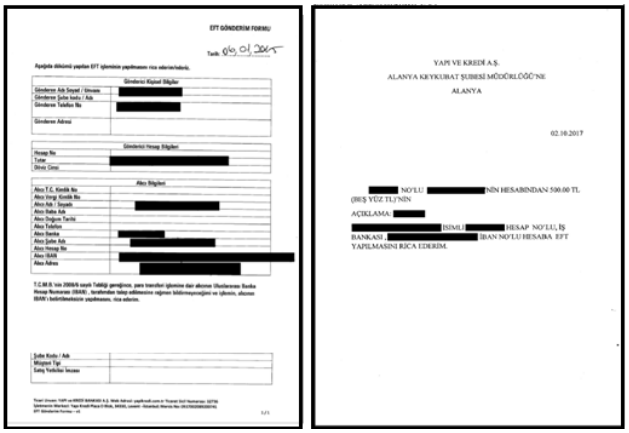

Figure 3. Sample images for the formatted document (left) and free text document (right) belonging to the money transfer class.

Although multimodal deep learning methods have been proposed for classification [15][16], visual question answering [17]-[19], image captioning [20][21], photo editing [22][23], and many other tasks, to the best of our knowledge, these methods have not been used for multimodal document classification. One of the recent works on this specific field proposed methods that use hand-crafted features and SVM for classification in early and late fusion strategies [24]. We propose multimodal classification networks by utilizing the deep neural networks on both modalities, which is different from previous works on document classification.

In this work, our main purpose is to be able to classify Turkish banking documents by utilizing the images of documents and the text which belongs to document. Textual and visual modalities are combined in two methods as early and late fusion in our proposed multimodal deep neural networks. In late fusion, we use our pre-trained LSTM model for text modality and pre-trained CNN model for vision modality to obtain probabilities. Then, we train a small decision level network, which predicts a class by using these probabilities. On the other hand, by feeding FastText embedding to LSTM and images to CNN, the proposed early fusion method jointly learns the image and text representations.

The rest of the paper is organized as follows: our methods are described in Section II, and the experimental results are discussed in Section III. Finally, the conclusion is given in Section IV.

II. METHODS

A. Word Vectors

To train word vectors, we organized an unsupervised dataset with 4.9M documents. Then, we used the Abbyy FineReader OCR tool [25] to extract texts from the data. The obtained text data consists of a total of 787M words and 55.5M vocabulary size. Since OCR generates faulty texts because of noisy images and misspellings, our vocabulary size is unusually large. To overcome problems caused by OCR noises and misspellings, we chose FastText embedding [26], since it works in agglutinative languages more effectively and is able to capture spelling errors better. In Table I, we have show that the most similar words to word ”nezdinizdeki” (a frequently used Turkish word that means *in care of*) are the

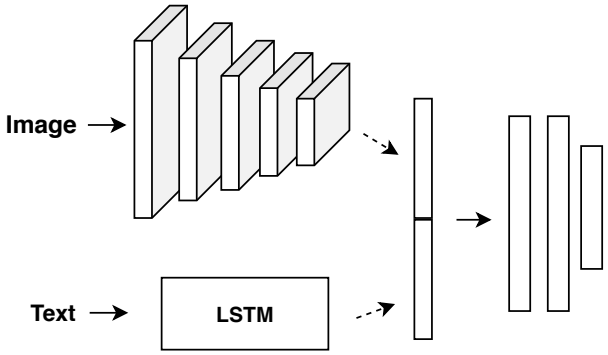

Figure 4. An overview of early fusion network.

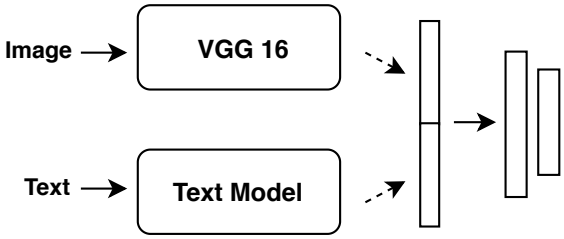

Figure 5. An overview of late fusion network.

words with OCR noises and misspellings. We trained 50, 100 and 300 dimensional FastText embedding vectors and chose to use 100 dimensional vectors in our experiments since it achieved the best accuracy.

B. Text Model

We used a rather simple neural network model with approximately 123k trainable parameters to classify documents using their texts. Our model consists of an embedding layer, a dropout layer with 0.4 rate, an LSTM layer with 128 hidden nodes and, finally, a dense layer that outputs the class predictions. Sequential capabilities of LSTMs fit well in our problem since each document in the data has different number of tokens and each of them is fed to LSTM in a single step. This rather simple model is able to perform good predictions thanks to our pre-trained word embeddings.

TABLE I. WORDS THAT ARE MOST SIMILAR TO "nezdinizdeki"

Word	Cosine Distance
nezdinizdekı	0.033
nezdinîzdeki	0.033
nezdînizdeki	0.035
nezdınizdeki	0.037
nezdinizdekî	0.038
nezrînizdeki	0.041
nezdınızdeki	0.043
nezinizdeki	0.046

C. Vision Model

We employed VGG-16 [7] network, which is a well-known architecture as the winner of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 [27]. This network includes 5 convolutional blocks, which have 13 convolutional layers, and these layers are followed by three fully connected layers. VGG-16 model, pre-trained on ImageNet dataset which includes approximately 1 million images from 1000 classes, is utilized for transfer learning. Firstly, this pre-trained model is fine-tuned on the RVL-CDIP dataset [2] which consists of 320,000 train, 40,000 validation, and 40,000 test images from 16 classes: letter, memo, email, filefolder, form, handwritten, invoice, advertisement, budget, news article, presentation, scientific publication, questionnaire, resume, scientific report, and specification. Thus, the pre-trained model is utilized to understand document images. Finally, this model is fine-tuned on our dataset to classify 45 classes.

D. Multimodal Classification Models

Multimodal learning allows using different modalities, *e.g.*, text, vision, speech, sensor data, which are related to each other during learning [28]. Utilizing different modalities, called fusion, can take place in a different phase of learning. Correspondingly, fusion methods can be categorized into late fusion, early fusion, and hybrid fusion. Early fusion methods are based on feature learning for different modalities, while late fusion methods are defined as the decision-based.

Early Fusion. The proposed early fusion multimodal network learns embeddings jointly for vision and text modality. Our proposed network is demonstrated in Figure 4. Firstly, the output of the last convolutional layer of the VGG-16 model pre-trained on ImageNet was used to obtain $7 \times 7 \times 512$ dimensional visual features. After this convolutional layer, global average pooling layer is added to reduce the dimension of the visual features to 512. Weights of the vision model are fixed and fine-tuned during training. Also, the text model, which is explained in Section II-B, is trained from scratch by using FastText word embedding as an input. The output of the text model is 128 dimensional. Textual and visual features are concatenated and fed into three fully-connected layers to classify documents. These fully connected layers are shared during training.

Late Fusion. The late fusion method combines different model probabilities to capture two model results as a kind of decision mechanism. According to our preliminary analysis on results of text and vision models, while text model predicts the wrong label for some documents, the vision model can predict the correct label even if the vision model accuracy is lower than the text model accuracy on classification. Our late fusion network takes as an input the concatenated probabilities obtained from text and vision models. This network consists of two fully-connected layers for the training of the classification model. An overview of our late fusion network is illustrated in Figure 5. Since each model has 45 probability score, the input is 90 dimensional vector for this network.

Implementation Details. We implemented our models in Keras [29]. Training of all models was done over GTX 1080Ti GPU with the batch size 32. We performed between 40-50 epochs for all models. We used ADAM [30] optimizer with the learning rate in the range of 0.001 and 0.0001 and early

 ISBN: 978-1-61208-731-3

stopping on the validation dataset by controlling validation loss for specified consecutive epochs.

III. Experimental Results

A. Dataset

In the banking domain, customer orders have over 200 distinct process types, however, most of these processes are scarcely used. Therefore, we selected the most commonly used 45 distinct process types and limited our problem scope to these. We create a dataset with approximately 27k banking order documents labeled with 45 different classes. Sample images from several classes are shown in Figure 2. Each class has a variable number of instances that changes between 100 and 1000. We split the data by 70%, 15%, and 15% in order to create train, validation, and test sets, respectively. Since we require a different train set for training vision/text model and late fusion model, we split our main train by 75% train1 and 25% train2; similarly, we split the main validation by 75% validation1 and 25% validation2.

Challenges on the dataset.

The main challenge is that the documents were wrongly labeled by the back-office operator at the bank due to operational mistakes. The second significant challenge is that several classes have similar documents visually and textually. In addition, inter-class variation is quite common in the dataset. Moreover, some of the documents are filled with handwriting and such documents do not have textual information since the OCR tool does not support handwritten documents. In addition to the valid customer orders, irrelevant documents also reside in the data.

These irrelevant documents are:

- ID cards,
- driving licenses,
- property ownership documents (deeds),
- credit cards photocopies,
- registry newspapers,
- blank documents.

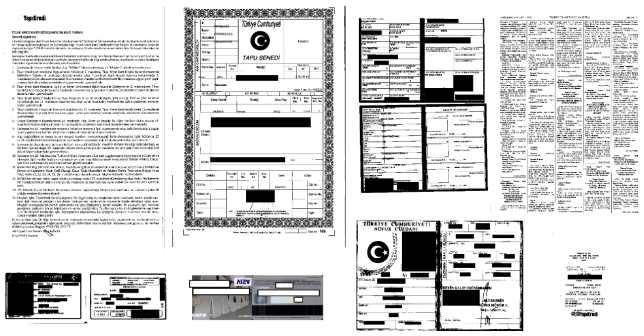

Figure 6. Sample images from the irrelevant documents.

Sample images of the irrelevant documents are illustrated in Figure 6. These irrelevant documents also have labels, similarly to the main documents. This situation causes confusion during training and testing. These dataset specific problems make our task harder than a document classification. Therefore, we considered our dataset a noisy dataset.

B. Results and Discussion

We evaluated our proposed methods on our Turkish banking dataset. We primarily investigated the effects of training the vision model on the early fusion multimodal network. To be able to do this, we employed two training strategies in the early fusion model. Firstly, ImageNet pre-trained model was utilized without updating weights, namely the fixed vision. Secondly, the last convolution block was fine-tuned in the model called fine-tuned vision. The text model was trained from scratch in both of them. As seen in Table II, fine-tuning vision model enhances the classification accuracy since the vision model adapts to the document images.

TABLE II. RESULTS ON EARLY FUSION

Learning joint embeddings	Accuracy (%)
fixed vision trained text from scratch	83.82
fine-tuned vision trained text from scratch	**85.29**

We have mainly four models to conduct experiments in order to classify banking documents: vision, text, early fusion and late fusion models, hence we have four main results. To compare all proposed models fairly, all experimental results are reported on the same test set. The results of all models are provided in Table III.

TABLE III. RESULTS ON ALL MODELS

Method	Accuracy (%)
Vision Model	70.53
Text Model	84.56
Early Fusion	**85.29**
Late Fusion	**85.42**

We also analyzed the performance of our multimodal networks by comparing with text and vision models. The highest improvement on the accuracy is observed when comparing the proposed multimodal networks with the vision model. However, we obtained a slightly better improvement when we compared the multimodal networks with the text model. This indicates that text embeddings are more beneficial than vision embedding for this task on these kinds of documents. Most of our documents in the dataset are free-formatted, thus this difference in accuracy between the vision and the text model is expected. On the other hand, multimodal networks achieve better results than the text model for class-based accuracies, especially when a class has visually rich documents, as expected. In addition, these improvements can be seen as fairly good results because of the constraints discussed in Section III-A. We observed that our fusion methods especially benefit from the text model since the text model and the multimodal model have approximately equal results. 1% improvement of the accuracy might seem unsatisfactory, but in a real-world scenario, where the prediction of each document is critical, such improvement diminishes the requirement of manpower.

 ISBN: 978-1-61208-731-3

IV. Conclusion

In this work, we proposed deep multimodal networks for document classification by using two fusion methods. The petition based customer order documents have different types of format, therefore, we focused to understand all types of documents by learning visual and textual features. We performed our experiments on our Turkish banking order dataset. The experimental results indicate that both early and late fusion multimodal models outperform text and vision models.

Acknowledgment

This work was supported by The Scientific and Technological Research Council of Turkey with the project no 3180571. We would like to thank our colleagues for their valuable discussions and support.

References

[1] EY Consulting LLC (UAE) and Microsoft, "Artificial Intelligence Maturity in Middle East Africa," Tech. Rep., 2019.

[2] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015, pp. 991–995.

[3] C. Tensmeyer and T. Martinez, "Analysis of convolutional neural networks for document image classification," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1. IEEE, 2017, pp. 388–393.

[4] M. Z. Afzal, A. Kölsch, S. Ahmed, and M. Liwicki, "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), vol. 1. IEEE, 2017, pp. 883–888.

[5] A. Das, S. Roy, U. Bhattacharya, and S. K. Parui, "Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks," in 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018, pp. 3180–3185.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[8] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[10] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," International Journal of Computer Vision (IJCV), vol. 115, no. 3, 2015, pp. 211–252.

[11] Y. Kim, "Convolutional neural networks for sentence classification," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1746–1751.

[12] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in Neural Information Processing Systems, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 649–657.

[13] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2015, pp. 1422–1432.

[14] Z. Yang et al., "Hierarchical attention networks for document classification," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2016, pp. 1480–1489.

[15] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in Advances in Neural Information Processing Systems, 2012, pp. 2222–2230.

[16] D. Wang, K. Mao, and G.-W. Ng, "Convolutional neural networks and multimodal fusion for text aided image classification," in 2017 20th International Conference on Information Fusion (Fusion). IEEE, 2017, pp. 1–7.

[17] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in International Conference on Machine Learning, 2016, pp. 2397–2406.

[18] S. Antol et al., "Vqa: Visual question answering," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.

[19] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 804–813.

[20] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in International Conference on Machine Learning, 2015, pp. 2048–2057.

[21] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.

[22] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," ICLR 2017, 2016.

[23] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in European Conference on Computer Vision. Springer, 2016, pp. 597–613.

[24] M. Rusiñol, V. Frinken, D. Karatzas, A. D. Bagdanov, and J. Lladós, "Multimodal page classification in administrative document image streams," International Journal on Document Analysis and Recognition (IJDAR), vol. 17, no. 4, 2014, pp. 331–341.

[25] Abbyy finereader. [Online]. Available: https://www.abbyy.com/en-us/finereader/

[26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, 2017, pp. 135–146.

[27] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision (IJCV), vol. 115, no. 3, 2015, pp. 211–252.

[28] J. Ngiam et al., "Multimodal deep learning," in Proceedings of the 28th International Conference on Machine Learning (ICML), 2011, pp. 689–696.

[29] F. Chollet et al., "Keras," https://keras.io, 2015.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," ICLR, 2015.

 ISBN: 978-1-61208-731-3