DOI: 10.3217/978-3-85125-954-4-04

A Modular Model Combining Visual and Textual Features for Document Image Classification

Amer Duhan
TU Wien
amer1.duhan@gmail.com

Robert Sablatnig
TU Wien
sab@cvl.tuwien.ac.at

Abstract

Document image classification is the classification of digitized documents. Typically, these documents are either scanned or photographed. One page of such a document is referred to as a document image. Classifying document images is a crucial task since it is an initial step in downstream applications. Most state-of-the-art document image classification models are based on a transformer network, which are pretrained on millions of scanned document images and thus require a huge amount of training resources. Additionally, this and other state-of-the-art document image classification models have well beyond 100 million parameters. In this work, we address both challenges. First, we create a model capable of competing with the current state-of-the-art models without pretraining on millions of scanned document images. Second, we create a model several times smaller than current state-of-the-art models in terms of parameters. The results show that the developed approach achieves an accuracy of 93.70% on the RVL-CDIP dataset, and a new state-of-the-art accuracy of 96.25% on Tobacco3482.

1. Introduction

The increasing digitalization has led companies to digitize their processes and content [4], and organize their information to improve the search and access to relevant data [6]. Thus, paper documents are subject to digitization, and document images are the output [21]. The task of document image classification is to categorize a given document image into a set of defined classes [12].

Due to its high importance, document image classification has been explored extensively [1]. However, most of the current State-Of-The-Art (SOTA) methods have either parameters in the hundreds of millions, pretrain on a larger dataset, or both, such as [32] or [33].

Thus, we propose a multimodal system based on SOTA image and language models, which are relatively small in their size (less than 100 million parameters). Furthermore, the amount of training data is limited to the RVL-CDIP dataset. Due to the modular nature of the architecture, we tested two model combinations to analyze their impact on the overall test set accuracy. Our experiments show that an image-only system achieves a higher test set accuracy than a multimodal system.

The contributions are the following:

- Developing a model that can compete with current SOTA models on the RVL-CDIP dataset without requiring millions of document images. Moreover, the developed model is much more efficient than the current SOTA models.
- Achieving a new SOTA on the Tobacco3482 dataset with 96.25% accuracy.

The remainder of this paper introduces the datasets in Section 2, discusses related work in Section 3, presents the methodology in Section 4, depicts the results in Section 5, and concludes the paper in Section 6.

2. Datasets

In the following, the two datasets used in this paper are discussed. First, the dataset on which the proposed architecture is trained and evaluated, and second on which it is finetuned and evaluated.

2.1. RVL-CDIP

This work is based on the RVL-CDIP [11] dataset since it was specifically created to test image classification algorithms on document images [7]. RVL-CDIP is a subset of the IIT-CDIP Test Collection (11 million documents) [20], which itself is a subset of the LTDL dataset [26] (14 million documents), that was created from public records of lawsuits against American tobacco companies [11]. The RVL-CDIP dataset contains 400,000 grayscale images with 16 classes, split evenly in an 8:1:1 ratio of training, validation, and test set.

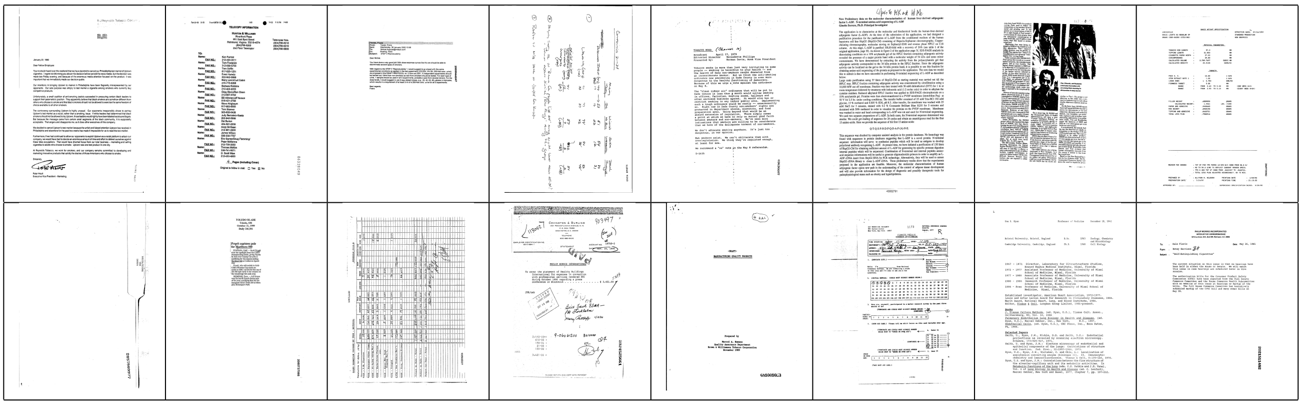

Figure 1. An example document image for each class from the RVL-CDIP dataset. From the top left image, the labels are the following: Letter, Form, Email, Handwritten, Advertisement, Scientific report, Scientific publication, Specification, File folder, News article, Budget, Invoice, Presentation, Questionnaire, Resume and Memo.

2.2. Tobacco3482

The Tobacco3482 [18] dataset, created from the same dataset as RVL-CDIP, the IIT-CDIP Test Collection, contains 3,482 grayscale document images. These images are split into 10 classes, which are not evenly distributed as in the RVL-CDIP dataset.

3. Related Work

The methods in all of the following works are tested on the RVL-CDIP test set.

Harley et al. [11], who have created the RVL-CDIP dataset, stack 5 CNNs, one of which is trained on the whole document image, and the others are trained over the header, footer, left body, and right body. These CNNs are either trained from scratch or transfer-learned from AlexNet [17]. Das et al. [6] use a similar technique. However, their CNNs are transfer-learned from VGG-16 [27]. A MLP, a class of artificial neural networks, is then found to perform as the best ensemble technique.

Afzal et al. [1] show that even though the ImageNet and RVL-CDIP datasets have different domains, a pretrained network on ImageNet, such as VGG-16, has a better accuracy score on the RVL-CDIP test set than no pretraining.

Tensmeyer and Martinez [29] train CNNs from scratch, i.e., randomly initialized. Various modifications are performed, such as changing the network depth, width, or input size. The authors show that the input size significantly impacts the performance.

Sarkhel and Nandi [25] utilize a spatial pyramid model to extract highly discriminative multi-scale feature descriptors from a visually rich document by leveraging the inherent hierarchy of its layout.

Ferrando et al. [10], Jain and Wigington [12], Audebert et al. [4], Kanchi et al. [14], and Bakkali et al. [5] combine image and text features in a two-stream approach by utilizing a CNN for image and an embedding for text. Jain and Wigington [12] use the VGG-16 to get image features and use different methods to extract text features, representing text at the sequence, word, and character level. Audebert et al. [4] utilize the MobileNetV2 [23] for image feature extraction, which has a similar performance in terms of accuracy, compared to VGG-16 while being significantly faster. As in [12], word-level text features are generated with FastText [13], a word embedding technique. Ferrando et al. [10] combine EfficientNet [28] for image features and a reduced version of BERT [8], a transformer model, for text features. Kanchi et al. [14] propose a hierarchical attention network for the textual stream, with fine-tuned BERT embeddings as input and an EfficientNet-B0 for the image stream. Bakkali et al. [5] combine NasNet$_{Large}$ [34] with BERT to achieve a SOTA accuracy of 97.05%, using an average ensembling for the image and text stream.

A transformer [30] architecture for document image classification is used in the work of Xu et al. [32]. This architecture is an extended version of BERT [8]. However, the model is pretrained on the IIT-CDIP Test Collection, which contains more than 11 million scanned document images. Another major difference, compared to all previous mentioned approaches, is that this method is suitable for classifying document images, and, for example, for form understanding, where the goal is to extract key-value pairs from document images. Xu et al. [33] extend [32]. The authors integrate visual information in the pre-training stage and use 2-D relative position representation for token pairs instead of absolute 2-D position embeddings, which Xu et al. [32] use to model the page layout. Just as its predecessor, this model is also suitable for other tasks outside of classifying document images.

Similarly, Powalski et al. [22], Wang et al. [31]. and

Srikar et al. [2] develop each a multimodal transformer based architecture, which performs a pretraining step. [22] simultaneously learns layout information, visual features, and textual semantics. In [31] the layout knowledge from monolingual structured documents is learned and then generalized to deal with multilingual ones. [2] combines textual, visual, and spatial features using a novel multi-modal self-attention layer.

4. Methodology

In this section, both streams (image and text) are elaborated, covering the preprocessing steps and the training strategy and architecture. Then, the method to combine both streams to form the final piece of the document image classification system is covered.

4.1. Image Stream

Compared to textual features, image features are preferred for the problem of document image classification [16]. The current SOTA CNN architecture, EfficientNet [28], is used for the image stream. The image stream and text stream are two independent parts of the whole model, which are combined in a later stage. The preprocessing steps, the training strategy, and the architecture are explained in the following.

4.1.1 Preprocessing steps

In our method, the image stream consists of five EfficientNets, each focusing on a input part. The preprocessing steps partly follow the work of [11]. First, all images are resized to 936 × 720. Then, 5 regions are defined for an image; holistic, header, footer, left body, and right body. The holistic region is the whole image itself. The header is defined as the first 307 pixel rows. Similarly, the footer is defined as the last 307 pixel rows. The left body is defined as the 480 central pixel rows and the first 360 pixel columns; similarly, the right body is defined as the 480 central pixel rows and the last 360 pixel columns. A slight intersection exists between the left and right body areas with the header and footer. Finally, each image is resized to 384 × 384.

The focus on specific regions of a document follows from the fact that certain categories show a low interclass variability, as seen in Figure 1 when comparing memo and letter. While memos often have a complete address section, letters typically have a ”To:” and ”From:”. Having a CNN to classify documents using only this region will much more likely learn those differences than a holistic CNN [11]. Similar to the header region, different CNNs are applied to each region described in the previous paragraph.

Since the document images are in grayscale, they are transformed into images with three channels, i.e., copied two times and stacked depth-wise along the third axis.

4.1.2 Training strategy and architecture

The training strategy and architecture on the full dataset are inspired by [6]. The main benefit of the following training strategy is reducing computational complexity. A three-level transfer learning achieves this.

The first level of transfer learning (L1) is initializing the weights of the holistic model from the corresponding EfficientNet-B1 model, trained on the ImageNet dataset. To train the holistic model, only the classifier added on top of the EfficientNet-B1 model is trained first, and all other weights of the model are frozen, such that they are not updated during backpropagation. This model’s weights are then used to initialize the same model (L2), but with all layers unfrozen, including the batch normalization layers. Now, all weights can be updated to further increase the prediction accuracy.

Next, its weights are taken to initialize the remaining four models (L3), i.e., the models for the header, footer, left body, and right body region. Like the holistic model, these four models are trained with early stopping on the validation loss and patience of 10. ReLU [9] is used as the activation function.

4.2. Text stream

The recent development in this field suggests that textual features are necessary to achieve SOTA results. A distilled version of BERT [8], called DistilBERT [24], is used as the backbone in our work since it is 40% smaller in size compared to BERT while retaining 97% of its language understanding capabilities. In the following sections, the preprocessing steps, as well as the training strategy, are explained.

4.2.1 Preprocessing steps

The Tesseract OCR system (version 4.1.1) extracts the text from the document images. Once this is done, the next step is preprocessing the extracted text before feeding it into a neural network. This is even more important when the text is extracted from document images, instead of, for instance, scraping the text from the web. Everything that is not a letter or a digit is removed. It is ignored if the text is less than two characters long, but single-digit numbers are kept. Moreover, the text is lowercased. There are some pages where no text can be extracted by Tesseract. In this case, or where the whole extracted text of a document image is removed due to the preprocessing steps, the extracted text is set to ””, i.e., a string of length zero.

4.2.2 Training strategy and architecture

Following the results from the image stream, the training strategy in the text stream takes a similar approach. Only

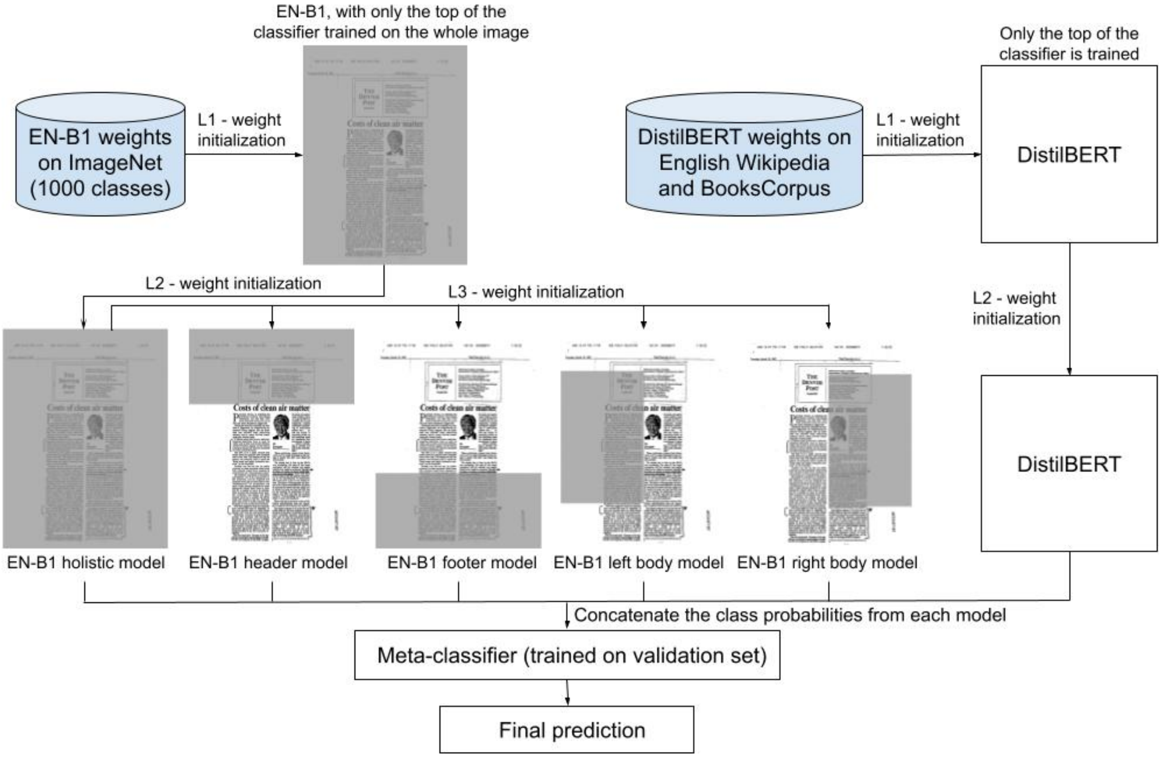
EN-B1, with only the top of the classifier trained on the whole image
EN-B1 weights on ImageNet (1000 classes)
L1 - weight initialization
DistilBERT weights on English Wikipedia and BooksCorpus
L1 - weight initialization
Only the top of the classifier is trained
DistilBERT
L2 - weight initialization
L3 - weight initialization
L2 - weight initialization
DistilBERT
EN-B1 holistic model
EN-B1 header model
EN-B1 footer model
EN-B1 left body model
EN-B1 right body model
Concatenate the class probabilities from each model
Meta-classifier (trained on validation set)
Final prediction

Figure 2. Proposed architecture for document image classification based on SOTA architectures, with an image stream, text stream, and utilizing different levels of transfer learning. EN = EfficientNet.

the classification head added on top of DistilBERT is trained first, with the features extracted from the base model.

DistilBERT and the original BERT model have two unique tokens: [CLS], a classification token, and [SEP], a separator token. The [CLS] token is used for classification tasks and is added in front of every sequence. Specifically, the last hidden state representation of the [CLS] token is used. This hidden state representation is then used as an input to the classification head.

Like in the image stream, the classification head is first trained, then the whole model. The final model is trained with early stopping and patience of 10, with ReLU as the activation function.

4.3. Stacked generalization

The last part of the system is to train a meta-classifier, which outputs the final predictions. It is adopted in document image classification models, such as in [6], [4], [10], [12], [3], and works by combining the (intermediary) output of one or more classifiers and feeding that as an input to a meta-classifier. To reduce overfitting, the meta-classifier is trained on the validation set. The goal of stacked generalization is to provide a lower generalization error than the base models. The meta-classifier is the last module of the document image classification system, and the full architecture is shown in Figure 2.

The input for the meta-classifier are the class probabilities (i.e. the softmax output). In this work, the meta-classifier, a 3-layer neural network, combines visual and textual features by concatenating them and producing the final output of the document image classification system.

Adam is chosen as the optimizer. Moreover, an image-only system versus a multimodal system is tested.

4.4. Tobacco3482

The document image classification model is also finetuned and evaluated on the Tobacco3482 dataset. To make results comparable with other works, such as [11], [15], [18], [19], or [10], the dataset is split as follows. From 3,482 images, 100 images per class are randomly selected. This constitutes the training set; and the remaining 2,482 images are the test set. This process is repeated 10 times, such that there are 10 different training and test sets, from which the median test set accuracy is reported. From the 1,000 training images, 200 are used for the validation set.

The training approach first uses the pretrained models on the RVL-CDIP dataset and then finetunes on the Tobacco3482 dataset, where only the added classification head

Results					
Author	Accuracy	# Parameters	Modality	Extra training data	Tobacco3482 Accuracy
Afzal et al. (2017) [1]	90.97	138.36	I	No	91.13
Kang et al. (2014) [15]	-	4.21	I	No	65.35
Kumar et al. (2014) [19]	-	-	I	No	43.27
Das et al. (2018) [6]	92.21	691.87	I	No	-
Audebert et al. (2020) [4]	90.60	3.64	I + T	No	87.80
Ferrando et al. (2020) [10]	92.31	85.47	I + T	No	94.90
Harley et al. (2015) [11]	89.80	58.35	I	No	79.90
Jain and Wigington (2019) [12]	93.60	138.36	I + T	No	-
Sarkhel and Nandi (2019) [25]	92.77	-	I	No	82.78
Tensmeyer and Martinez (2017) [29]	91.03	-	I	No	-
Xu et al. (2020) [32]	94.42	160.00	I + T	Yes	-
Xu et al. (2021) [33]	95.64	426.00	I + T	Yes	-
Srikar et al. (2021) [2]	96.17	183.00	I + T	Yes	-
Wang et al. (2022) [31]	95.68	-	I + T	Yes	-
Powalski et al. (2021) [22]	95.52	780.00	I + T	Yes	-
Bakkali et al. (2020) [5]	97.05	197.21	I + T	No	-
Kanchi et al. (2022) [14]	95.48	-	I + T	Yes	95.70
Proposed approach	93.70(I) / 93.50(I+T)	40.72	I	No	95.65(I) / 96.25(I+T)

Table 1. Test set results on RVL-CDIP and Tobacco3482. Accuracy in %. The number of parameters (in millions) is either explicitly stated in the work, an estimation, or omitted. I = Image, T = Text.

is trained.

Additionally, a meta-classifier is trained to combine the softmax outputs on the training set of the image and text models. Similarly, an image-only and multimodal system is trained. The models are trained with Adam, ReLU, and early stopping with patience of 3.

5. Results

The proposed approach includes two results per dataset, each with an image-only and multimodal system. The results are depicted in Table 1.

An accuracy of 93.70% on RVL-CDIP and 96.25% on Tobacco3482 is achieved. Note that on the RVL-CDIP dataset, the image-only system achieves a higher accuracy, while on the Tobacco3482 dataset, it is the multimodal system. That is, adding textual information decreases the accuracy on the RVL-CDIP dataset, which goes against the results of other papers that have used textual information (see Table 1). The difference in the accuracy between the image-only and multimodal approach is larger on the Tobacco3482 dataset.

Most SOTA papers have used additional training data with a multimodal approach. Table 1 shows, that all papers, who have reached an accuracy of over 94%, have used an extra training data, either the full IIT-CDIP Test Collection (11 million documents) or a fraction of it, except the current SOTA [5], with 97.05% accuracy. Moreover, all papers with an accuracy of over 94% are fully based on a Transformer architecture, except [5] and [14].

The number of parameters of the proposed approach (around 41 million) is multiple times smaller than in the current SOTA methods. Even though the result on the RVL-CDIP dataset could not match them, a new SOTA has been achieved on the Tobacco3482 dataset using the multimodal approach, beating the previous SOTA result of Kanchi et al. [14] by 0.55 percentage points. Additionally, the image-only approach missed the previous SOTA result by 0.05 percentage points.

The model is trained on a NVIDIA T4 GPU with 16GB VRAM. One epoch takes about 220 minutes for the image models on the RVL-CDIP dataset. Each image model is trained for about 14 epochs, i.e., for 70 epochs combined. The text model is trained for 8 epochs, with about 136 minutes per epoch. These numbers refer to those models, where the weights of all layers are unfrozen.

6. Conclusion

The goal of the proposed approach is to develop a model, which can compete with current SOTA methods and be relatively efficient, i.e., have a relatively small number of parameters. Even though the current SOTA results on the RVL-CDIP dataset could not be quite matched, the developed model is around 5 times smaller in terms of the number of parameters. On the Tobacco3482 dataset, however, a new SOTA result is achieved. Interestingly, contrary to the papers using a multimodal approach mentioned in Table 1, the textual information decreases the accuracy on the RVL-CDIP dataset.

References

[1] Muhammad Zeshan Afzal, Andreas Kolsch, Sheraz Ahmed, and Marcus Liwicki. Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification. *ICDAR*, 1:883–888, 2017.

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. DocFormer: End-to-End Transformer for Document Understanding. *ICCV*, pages 973–983, 2021.

[3] Muhammad Nabeel Asim, Muhammad Usman Ghani Khan, Muhammad Imran Malik, Khizar Razzaque, Andreas Dengel, and Sheraz Ahmed. Two stream deep network for document image classification. *ICDAR*, pages 1410–1416, 2019.

[4] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. Multimodal deep networks for text and image-based document classification. *CCIS*, 1167:427–443, 2020.

[5] Souhail Bakkali, Zuheng Ming, Mickael Coustaty, and Marcal Rusinol. Visual and textual deep feature fusion for document image classification. *CVPRW*, pages 2394–2403, 2020.

[6] Arindam Das, Saikat Roy, Ujjwal Bhattacharya, and Swapan Kumar Parui. Document Image Classification with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks. *ICPR*, pages 3180–3185, 2018.

[7] Tyler Dauphinee, Nikunj Patel, and Mohammad Rashidi. Modular multimodal architecture for document classification. *arXiv:1912.04376*, 2019.

[8] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 1:4171–4186, 2019.

[9] Jianli Feng and Shengnan Lu. Performance Analysis of Various Activation Functions in Artificial Neural Networks. *Journal of Physics: Conference Series*, 1237(2), 2019.

[10] Javier Ferrando, Juan Luis Domínguez, Jordi Torres, Raúl García, David García, Daniel Garrido, Jordi Cortada, and Mateo Valero. Improving accuracy and speeding up document image classification through parallel systems. In *Computational Science – ICCS 2020*, pages 387–400. 2020.

[11] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. *ICDAR*, pages 991–995, 2015.

[12] Rajiv Jain and Curtis Wigington. Multimodal document image classification. *ICDAR*, 3:71–77, 2019.

[13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *EACL 2017*, 2:427–431, 2017.

[14] Shrinidhi Kanchi, Alain Pagani, Hamam Mokayed, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. EmmDocClassifier: Efficient Multimodal Document Image Classifier for Scarce Data. *Applied Sciences (Switzerland)*, 12(3), 2 2022.

[15] Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for document image classification. *ICPR*, pages 3168–3172, 2014.

[16] Andreas Kolsch, Muhammad Zeshan Afzal, Markus Ebbecke, and Marcus Liwicki. Real-Time Document Image Classification Using Deep CNN and Extreme Learning Machines. *ICDAR*, 1:1318–1323, 2018.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NeurIPS*, 25, 2012.

[18] Jayant Kumar and David Doermann. Unsupervised classification of structurally similar document images. *ICDAR*, pages 1225–1229, 2013.

[19] Jayant Kumar, Peng Ye, and David Doermann. Structural similarity for document image classification and retrieval. *PRL*, 43(1):119–126, 2014.

[20] David D. Lewis, Gady Agam, Shlomo Engelson Argamon, Ophir Frieder, David A. Grossman, and Jefferson Heard. Building a test collection for complex document information processing. *ACM SIGIR*, pages 665–666, 2006.

[21] Lawrence O'Gorman and Rangachar Kasturi. Executive Briefing: Document Image Analysis. 1997.

[22] Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. *ICDAR*, pages 732–747, 2021.

[23] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CVPR*, pages 4510–4520, 2018.

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv: 1910.01108*, 2019.

[25] Ritesh Sarkhel and Arnab Nandi. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. *IJCAI*, pages 3360–3366, 2019.

[26] Heidi Schmidt, Karen Butter, and Cynthia Rider. Building Digital Tobacco Industry Document Libraries at the University of California, San Francisco Library/Center for Knowledge Management. *D Lib Mag.*, 8(9), 2002.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR 2015*, pages 1–14, 2015.

[28] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *ICML 2019*, pages 6105–6114, 2019.

[29] Chris Tensmeyer and Tony Martinez. Analysis of Convolutional Neural Networks for Document Image Classification. *ICDAR*, 1:388–393, 2017.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Jones Llion, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, pages 6000–6010, 2017.

[31] Jiapeng Wang, Lianwen Jin, and Kai Ding. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. *ACL*, 1:7747–7757, 2 2022.

[32] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *ACM SIGKDD*, pages 1192–1200, 2020.

[33] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. *ACL 2021*, pages 2579–2591, 2021.

[34] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning Transferable Architectures for Scalable Image Recognition. *CVPR*, 2018.