CROSS-MODAL DEEP NETWORKS FOR DOCUMENT IMAGE CLASSIFICATION

Souhail Bakkali[1] *Zuheng Ming*[1] *Mickaël Coustaty*[1] *Marçal Rusiñol*[2]
[1]L3i, University of La Rochelle, France
[2]Computer Vision Center, Universitat Autònoma de Barcelona, Spain

{souhail.bakkali, zuheng.ming, Mickael.coustaty}@univ-lr.fr, marcal@cvc.uab.es

ABSTRACT

As a fundamental step of document related tasks, document classification has been widely adopted to various document image processing applications. Unlike the general image classification problem in the computer vision field, text document images contain both the visual cues and the corresponding text within the image. However, how to bridge these two different modalities and leverage textual and visual features to classify text document images remains challenging. In this paper, we present a cross-modal deep network that enables to capture both the textual content and the visual information included in document images. Thanks to the efficient jointly learning of text and image features, the proposed cross-modal approach shows its superiority to the state-of-the-art single-modal methods. In this paper, we propose to use NASNet-Large and Bert to extract image and text features respectively. Experimental results demonstrate that the proposed cross-modal approach achieves new state-of-the-art results for text document image classification on the benchmark Tobacco-3482 dataset, outperforming the current state-of-the-art method by 3.91% of classification accuracy.

Index Terms— Text document image classification, cross-modal feature learning, deep CNNs.

1. INTRODUCTION

Document images can be classified into various classes based on their textual content and/or their structural properties [1]. It is an important step of various document image processing tasks such as document retrieval, information extraction, and text classification. Benefiting from the great progress of image feature learning by convolutional neural networks (CNNs), document images can be classified based on their visual information seen as a conventional image classification problem as in [2]. However, the classification of document images based only on visual information may encounter the problem of low inter-class and high intra-class structural variations of text document images [3] as shown in Fig. 1. Rather than the general image classification problem, document images contain both the visual cues and the corresponding text which present visual-text semantic relationships. The distinct semantic relationships can be used to mitigate the issue of ambiguity between inter/intra-class documents with only visual information.

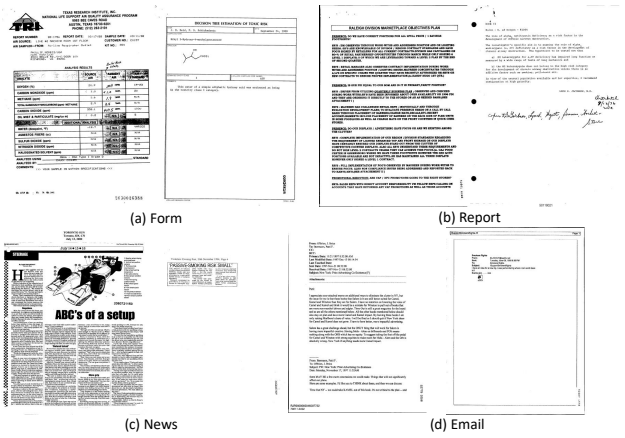

Fig. 1. Sample images from Tobacco-3482 dataset showing the low inter-class and high intra-class of structural variations of document images.

In this paper, we propose a cross-modal approach with deep two-headed networks which is capable to learn simultaneously the text content and the structural visual information from document images as shown in Fig. 1. In order to represent the document text content, an optical character recognition technique (OCR) is considered to extract the text within the document image. The latent semantic analysis is following the OCR based on both left and right context. Instead of ELMO (shallow bidirectional) [4] and OpenAI-GPT [1](one direction, left to right), the deep bidirectional pretrained BERT model [5] is introduced to learn the text semantic features. On the other hand, Neural Architecture Search Network (NASNet) [6] pretrained on ImageNet is proposed to extract relevant visual features of the document images in Tobacco-3482 dataset [7]. Rather than the [8, 9], we propose mainly two different methods, i.e. concatenating and superposing the text and image features to generate the fusing features. Finally the best performance is obtained with the superposing features.

[1]https://openai.com/blog/better-language-models/

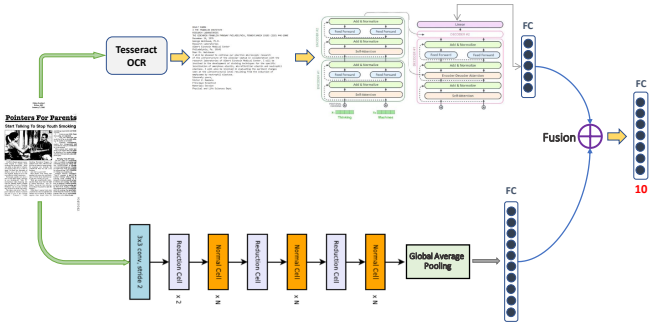

Fig. 2. The proposed cross-modal deep network performs document image classification by jointly learning text and image features.

The contributions of our paper are as follows:

- We propose a cross-modal deep network that jointly learns text-image features to classify document images which achieves new state-of-the-art results.
- We propose to use NASNet-Large and BERT to extract image-text features respectively, which achieve state-of-the-art results among the single-modal methods.
- We introduce three feature fusion methods to merge text-image features in the cross-modal framework.

2. RELATED WORK

Recently, several deep learning techniques have been proposed and have shown their notable performance for the text document image classification task. In [10], the concept of transfer learning was used to improve the classification accuracy on the standard Tobacco-3482 dataset, using AlexNet architecture pre-trained on ImageNet dataset. Another work in the area introduced region based modelling [11] by using transfer learning on the larger RVL-CDIP[2] dataset. Also, in [12] AlexNet, VGG-16, GoogLeNet, and ResNet-50 models were tested using transfer learning on the RVL-CDIP and Tobacco3482 datasets. In comparison, [13] concentrated on speed by replacing the fully connected portion of the VGG architecture with extreme learning machines (ELM). Therefore, there have been lately numerous studies on text-based document classification, following [14] work, structure-based features were used to classify text content of a document image. Moreover, [15] experimented with one-class SVM for document classification based on various text features such as TF-IDF. Furthermore, [16] presented a survey on different text classification algorithms. Recent learned word embedding approaches such as Word2vec [17], Glove [18], ELMO, FastText[3], XLNet [19], have led to significant improvements in the area by learning lexical, syntactic, semantic, and pragmatic approaches to extract relevant information from text-based documents. Latest works have achieved state-of-the-art results by combining both textual and visual features to perform text image classification. In [20], a hybrid approach was proposed to capture contextual information using an RNN, followed by a CNN to extract features. Another recent pipeline [9] has jointly performed visual and textual feature extraction using existing models MobileNet [21] and FastText. Moreover, [8] has come with a novel approach, leveraging textual and visual features to classify document images, by introducing a filter based feature-ranking algorithm, followed by a multi-channel CNN, jointly performed with a InceptionV3 [22] network.

[2] https://www.cs.cmu.edu/~aharley/rvl-cdip/

3. CROSS-MODAL LEARNING

3.1. Image Features

In the recent past, deep learning has been widely used in multiple domains including text document image classification. Specifically, transfer learning pretrained models on the large-scale ImageNet dataset to the document classification problem is very common. In this work, we trained the NASNet-Large model pretrained on ImageNet to extract the image features for document classification on the Tobacco-3482 dataset. For effective training, we applied shear transform data augmentation with a range of 0.1, which enables to augment the training data by stochastically transforming each input sample during SGD training. Also, we introduced cutout data augmentation [23] to improve regularization of CNNs, which removes the redundancy of the images and augments the dataset by partially occluded versions of existing samples. For the final layers of the NASNet-Large, the feature maps obtained are pooled via global average pooling and passed to a fully connected layer to perform classification via a softmax layer. The categorical cross-entropy loss function of softmax is given by:

$$\begin{aligned}\mathcal{L}_{s1}(\mathbf{X}_1;\Theta_1) &= \sum_{k=1}^{K} -y_k log P(y_k|\mathbf{X_1},\theta_k) \\ &= -\sum_{k=1}^{K} y_k log \frac{e^{f^{\theta_k}(\mathbf{X}_1)}}{\sum_1^K e^{f^{\theta_{k'}}(\mathbf{X}_1)}}\end{aligned} \tag{1}$$

Where $\{\mathbf{X}_1, \Theta_1\} \in \mathbb{R}^{d_1}$, and d_1 is the dimension of X_1 features of the Image branch. K is the number of classes in the dataset, $y_k \in \{0, .., 9\}$ is the one-shot label of the feature $\mathbf{X_1}$, $P(y_k|\mathbf{X_1}, \theta_k)$ is the softmax function over the activation function $f^{\theta_k}(\mathbf{X_1})$ where $\{\theta_k\}_{k=1}^K = \Theta_1, \theta_k \in \mathbb{R}^{d_1}$.

[3] https://fasttext.cc/docs/en/english-vectors.html

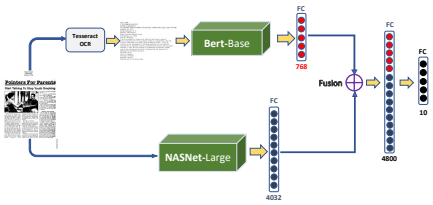

Fig. 3. (a) Naive concatenation

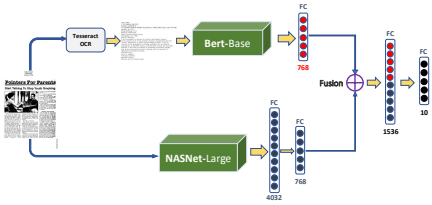

Fig. 4. (b) Equal concatenation

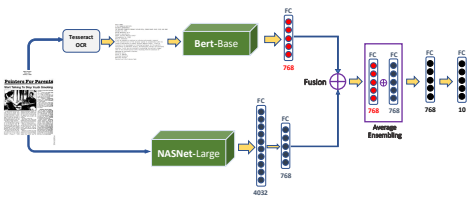

Fig. 5. (c) Superposing fusion

Fig. 6. Three different methods to merge both visual and textual streams.

3.2. Text Features

Intuitively, document images are characterized by their textual content within the second branch of the network. As most document classifiers are based on text content, we use an off-the shelf optical character recognition (OCR) engine to extract text, i.e. Tesseract OCR[4], which is noisy and not as clean as the training data. This OCR engine can conduct a fully automatic page segmentation and deskewing which performs affine transformation and rotation if needed. Recently, different contextualized word embedding procedures (i.e. ELMO, BERT, GPT, XLNet) have been proposed aiming to capture and map semantic meaning in different contexts in order to address the issue of polysemous and the context-dependent nature of words. Traditional word vectors are shallow representations (i.e. word2vec, Glove, FastText), and fail to capture higher-level information that might be more relevant.

Therefore, we focused on pre-trained bidirectional BERT-base model, based on the transformer architecture using a much faster and highly-efficient attention-based approach. The tokenization process involves splitting input text into a list of tokens that are available in the vocabulary, and gets the embedding for each word in the sequence. Each word of the sequence is mapped to an embedding vector that the model will learn during training. BERT uses WordPiece tokenization technique to deal with out-of-vocabulary (OOV) words, in which every OOV word is progressively splitted into subwords. For the final layer of the BERT model, we passed the last decoder stack output layer to a softmax layer with categorical cross-entropy loss function.

3.3. Cross-Modal Features

As described in Fig.6, we represent three methods to combine both visual and textual features from the two branches of our cross-modal approach.

Naive concatenation: we directly concatenate the obtained image embedding features $X_1 \in \mathbb{R}^{d_1}$ and text embeddings $X_2 \in \mathbb{R}^{d_2}$ with their original dimensions. The generated cross-modal features are given by:

$$X_a = [X_1, X_2], \quad X_a \in \mathbb{R}^{d_1+d_2} \tag{2}$$

[4]https://github.com/tesseract-ocr/tesseract

Equal concatenation: we add a fully connected layer following the original image embeddings to generate new image embeddings having the same dimension as the text embeddings . Thus, the cross-modal features are the concatenation of the two embedding features with same dimension:

$$X_a = [X_1, X_2], \quad X_a \in \mathbb{R}^{2d_1} \tag{3}$$

Superposing fusion: Rather than the concatenation, we show that the two embedding features are superposed directly to generate the cross-modal features. Note that the two embedding features have the same dimension as the equal concatenation.

$$X_a = [X_1 + X_2], \quad X_a \in \mathbb{R}^{d_1} \tag{4}$$

The softmax with categorical cross-entropy loss function is introduced to perform document classification based on the cross-modal embedding features.

4. EXPERIMENTS AND ANALYSIS

4.1. Dataset

In this section, we introduce the public Tobacco-3482 dataset to conduct our experiments to demonstrate the effectiveness of our cross-modal feature learning method. It contains 3482 grayscale scanned document images of ten categories: ADVE, Email, Form, Letter, Memo, News, Notes, Report, Resume and Scientific. Some representative images from the dataset are shown in Fig. 1.

4.2. Implementation Details

We describe in this subsection the implementation details used for our proposed cross-modal approach. We have trained our networks on a NVIDIA Quadro GP100 GPU. All input images are downsized to 331x331. The trained networks are optimized with SGD, momentum of 0.9, learning rate of 0.001, and a step decay schedule. For the visual stream, L2 regularization was applied with a batch size of 16 for 100 epochs. Dropout is applied to the final softmax layer with probability of 0.5. To minimize the high intra-class similarity variations in document images, we trained our network with shear transforms, as [10] suggested, with $\theta \in [-10°, 10°]$.

Table 1. Overall accuracy on the Tobacco-3482 dataset

Model	Accuracy(%)	ADVE	Email	Form	Letter	Memo	News	Notes	Report	Resume	Scientific
Single-Modal (Image-NASNet)	96.25	1	1	0.96	0.94	0.98	1	0.90	1	0.78	0.90
Single-Modal (Text-Bert)	97.18	0.97	0.99	0.98	0.93	0.97	0.98	0.89	1	0.96	0.95
Multimodal Model [9]	87.8	0.93	0.98	0.88	0.86	0.90	0.90	0.85	0.71	0.96	0.68
Two Stream Model [8]	95.8	0.94	0.98	0.95	0.98	0.97	0.97	0.88	0.92	1	0.93
Cross-Modal (Naive Concat.)	99.14	1	0.99	0.96	1	1	1	1	0.98	1	0.98
Cross-Modal (Equal Concat.)	98.42	0.98	0.99	0.95	1	0.98	0.97	1	1	0.96	0.98
Cross-Modal (Superposing fusion)	99.71	1	1	0.97	1	1	1	1	1	1	1

Table 2. The Image-stream evaluation of best models from different methods on Tobacco-3482 dataset

Method	Accuracy(%)
AlexNet [12]	90.04
GooGleNet [12]	88.4
VGG-16 [12]	91.01
ResNet-50 [12]	91.13
MobileNetV2 [9]	84.50
InceptionV3 [8]	93.2
NASNet-Large	96.25

Table 3. Accuracy comparison of Text-stream state-of-the-art models on Tobacco-3482 dataset

Method	Accuracy(%)
FastText-CNN [9]	73.8
Feature Ranking (ACC2) [8]	87.1
Glove-CNN1D-LSTM	51
Glove-GRU	61
Bert	97.18

As for the text stream, it is trained with a batch size of 40, and a sequence length of 128 for 50 epochs.

4.3. Evaluation and Ablation Analysis

To evaluate the effectiveness of our proposed cross-modal approach for document image classification, we firstly investigate the performance of the single modalities based on visual and textual features. Then, we compare our cross-modal method to the single modalities, and finally, to the state-of-the-art methods based on two stream deep network.

In this work, we propose to use NASNet-Large to classify the document images with only visual features. As shown in Table 2, the NASNet-Large gains the best result of 96.25% which outperforms the state-of-the-art single-modal method based on the InceptionV3 network by a 3.05% margin. Note that the NASNet-Large is pretrained on the ImageNet.

Besides, for the single-modal text pipeline, we tested combined architectures such as CNN-LSTM and GRU on top of Glove word embeddings as shown in Table 3. Results demonstrate that BERT model achieves a new state-of-the-art result of 97.18%, outperforming all existing methods with a very high margin of 10.08%. Therefore, attention-based approaches are highly-efficient operations thanks to their fast run-time characteristics.

In addition, Table 1. compares the performance of the three proposed methods to perform cross-modal feature learning classification. Naive concatenation raises the performance of all classes except for the class Report where it falls by 2%, while Email and Form classes' performance did not raise. For the second concatenation method, we compress image features and concatenate them with the text features, having both the same dimensional vector. As well, our cross-modal network manages to raise the performance for all classes except for the classes News and Form, where it drops by 1%. This is mainly due to the highly overlapped categories (Form, Report, Email) shown in Fig. 1. Finally, our cross-modal feature learning approach with superposing fusion features, outperforms all current state-of-the-art methods with a significant margin of 3.91% comparing to the two-stream methods, and of 2.53% comparing to the single-modal proposed methods. Thus, the superposing approach raises the performance of all classes regarding their structural property differences.

Out of the three proposed methods that merge both text and image features, the superposing fusion method jointly learns more relevant information from textual content and visual features, achieving the best performance with 99.71% classification accuracy.

5. CONCLUSION

This paper presents a hybrid cross-modal feature learning approach that leverages both text and image features to classify document images. NASNet and BERT are proposed to extract image and text features respectively. With the proposed fusion methods, our cross-modal approach outperforms all current state-of-the-art methods either with single-modal or multi-modal approaches. In future works, we intend to investigate the effectiveness of our cross-modal approach on the larger RVL-CDIP dataset.

6. REFERENCES

[1] Nawei Chen and Dorothea Blostein, "A survey of document image classification: problem statement, classifier architecture and performance evaluation," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 10, pp. 1–16, 2006.

[2] M. Z. Afzal, S. Capobianco, M. I. Malik, S. Marinai, T. M. Breuel, A. Dengel, and M. Liwicki, "Deepdocclassifier: Document classification with deep convolutional neural network," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1111–1115.

[3] Lucia Noce, Ignazio Gallo, Alessandro Zamberletti, and Alessandro Calefati, "Embedded textual content for document image classification with convolutional neural networks," in *Proceedings of the 2016 ACM Symposium on Document Engineering*, New York, NY, USA, 2016, DocEng '16, p. 165–173, Association for Computing Machinery.

[4] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," *CoRR*, vol. abs/1802.05365, 2018.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[6] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le, "Learning transferable architectures for scalable image recognition," *CoRR*, vol. abs/1707.07012, 2017.

[7] Jayant Kumar, Peng Ye, and David S. Doermann, "Structural similarity for document image classification and retrieval," *Pattern Recognit. Lett.*, vol. 43, pp. 119–126, 2014.

[8] M. N. Asim, M. U. G. Khan, M. I. Malik, K. Razzaque, A. Dengel, and S. Ahmed, "Two stream deep network for document image classification," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019, pp. 1410–1416.

[9] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal, "Multimodal deep networks for text and image-based document classification," *CoRR*, vol. abs/1907.06370, 2019.

[10] 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)*, 2017, vol. 01, pp. 388–393.

[11] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," *CoRR*, vol. abs/1502.07058, 2015.

[12] 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," *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Nov 2017.

[13] A. Kölsch, M. Z. Afzal, M. Ebbecke, and M. Liwicki, "Real-time document image classification using deep cnn and extreme learning machines," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, vol. 01, pp. 1318–1323.

[14] Christian K. Shin, David S. Doermann, and Azriel Rosenfeld, "Classification of document pages using structure-based features," *International Journal on Document Analysis and Recognition*, vol. 3, pp. 232–247, 2001.

[15] Larry M. Manevitz and Malik Yousef, "One-class svms for document classification," *J. Mach. Learn. Res.*, vol. 2, pp. 139–154, Mar. 2002.

[16] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, pp. 150, Apr 2019.

[17] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[18] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *CoRR*, vol. abs/1906.08237, 2019.

[20] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao, "Recurrent convolutional neural networks for text classification," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015, AAAI'15, p. 2267–2273, AAAI Press.

[21] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015.

[23] Terrance Devries and Graham W. Taylor, "Improved regularization of convolutional neural networks with cutout," *CoRR*, vol. abs/1708.04552, 2017.