Visual and Textual Deep Feature Fusion for Document Image Classification

Souhail Bakkali[1] Zuheng Ming[1] Mickaël Coustaty[1] Marçal Rusiñol[2]
[1]L3i, University of La Rochelle, France
[2]CVC, Universitat Autònoma de Barcelona, Spain
{souhail.bakkali, zuheng.ming, mickael.coustaty}@univ-lr.fr, marcal@cvc.uab.es

Abstract

The topic of text document image classification has been explored extensively over the past few years. Most recent approaches handled this task by jointly learning the visual features of document images and their corresponding textual contents. Due to the various structures of document images, the extraction of semantic information from its textual content is beneficial for document image processing tasks such as document retrieval, information extraction, and text classification. In this work, a two-stream neural architecture is proposed to perform the document image classification task. We conduct an exhaustive investigation of nowadays widely used neural networks as well as word embedding procedures used as backbones, in order to extract both visual and textual features from document images. Moreover, a joint feature learning approach that combines image features and text embeddings is introduced as a late fusion methodology. Both the theoretical analysis and the experimental results demonstrate the superiority of our proposed joint feature learning method comparatively to the single modalities. This joint learning approach outperforms the state-of-the-art results with a classification accuracy of 97.05% *on the large-scale RVL-CDIP dataset.*

1. Introduction

For many public and private organizations, understanding and analyzing data from documents manually is time consuming and expensive. Unlike the general images, document images may be presented in a variety of forms due to the different manners of organizing each document. However, extracting an accurate and structured information from the wide variety of documents is very challenging considering their visual structural properties and their textual heterogeneous content. From a computer vision perspective, earlier studies that have been using deep neural networks for document analysis tasks focused on their structural similarity constraints and their visual features [5, 29, 20, 21]. As most recent deep learning methods do not require extracting features manually, the state-of-the-art approaches based on visual information of document images treat the problem as a conventional image classification. Additionally, from a natural language processing perspective, Yang *et al.* [37] presented a neural network to extract semantic information based on word embeddings from pretrained natural language models. Nevertheless, classifying documents with only visual information may encounter the problem of low inter-class discrimination, and high intra-class structural variations of highly overlapped document images [1] shown in Fig. 1. As such, jointly learning visual cues and text semantic relationships is an inevitable step to mitigate the issue of highly correlated classes. Recent methods have used multimodal techniques to leverage both image and text modalities extracted by an OCR engine to perform fine-grained document image classification [4, 9, 36, 3].

Therfore, we study the capability of static and dynamic word embeddings to extract meaningful information from a text corpus. While static word embeddings fail to capture polysemy by generating the same embedding for the same word in different contexts, dynamic word embeddings are able to capture word semantics in different contexts to address the issue of polysemous and the context-dependent nature of words. We explored and evaluated both static and dynamic word embeddings on the large RVL-CDIP [1] [15] dataset. Furthermore, we propose a cross-modal network to learn simultaneously from the visual structural properties and the text information from document images based on two different models. The learned cross-modal features are combined as the final representation of our proposed network to boost the classification capacity of document images. To perform text classification, an optical character recognition (OCR) is employed to extract the textual content of each document image. The latent semantic analysis is following the OCR. We utilize the pretrained Glove and FastText [27, 26] as static word embeddings, followed by a gated recurrent unit (GRU) mechanism introduced by J.Chung *et al.* and K.Cho *et al.* [6]. GRU is a simplified variant of LSTM architectures introduced by S. Hochreiter

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

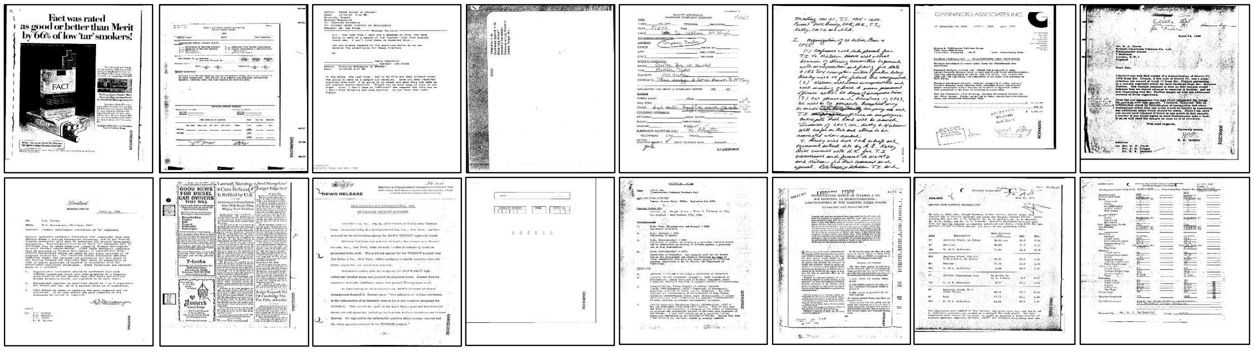

Fact was rated as good or better than Merit by 66% of low 'tar' smokers!

Figure 1. Samples of different document classes in the RVL-CDIP dataset which illustrate the low inter-class discrimination and high intra-class variations of document images. From left to right: *Advertisement, Budget, Email, File folder, Form, Handwritten, Invoice, Letter, Memo, News article, Presentation, Questionnaire, Resume, Scientific publication, Scientific report, Specification*

and J. Schmidhuber [13] to overcome the vanishing gradient problems. Moreover, based on both left and right context, the deep bidirectional pretrained BERT model [11] is utilized as a contextualized dynamic word embedding to learn the text semantic features.

To conduct image classification task, we investigate the impact of both heavyweight (*i.e*. with a large amount of paramaters) and lightweight (*i.e*. with a much lower amount of paramaters) deep neural network architectures on learning deep structural properties from document images. The heavyweight models with large size parameters such as NasNet$_{Large}$ [39], Inception-ResNet-v2 [32] can achieve state-of-the-art classification accuracy on the widely used ImageNet [10] dataset in the cost of the computational complexity and time consuming. Instead, the lightweight models with less parameters designed for the constrained environment, *e.g*. real-time environment, mobile application with less hardware resources, focus on the trade-off between the efficiency and the model accuracy. Amongst all classes of the RVL-CDIP dataset, some samples from specific categories present particular layout properties and document structures. Most classes are mainly composed of text information, while the classes like Advertisement, File folder contain only images with very few text information. Specifically, some samples do not contain any text data. Another class such as Handwritten, which is composed of handwritten text characters, produces noisy output text resulted by the processing of OCR engine. The idea behind this work relies on whether combining learned visual features with textual features could effectively benefit for the document image classification task, to achieve accurate results for the categories (Advertisement, File folder, Handwritten). In order to get the fusion embeddings, we adopt a late fusion scheme methodology. Our main contributions of this paper are as follows:

- We propose a cross-modal deep network that leverages textual contents and visual features to classify document images. We show that the joint learning methodology boosts the overall accuracy comparatively to the single-modal networks.
- We evaluate the performance of static and contextualized dynamic word embeddings to classify textual content of document images.
- We review the impact of training heavyweight and lightweight deep neural networks on learning relevant structural information from document images.

2. Related work

In the recent past, several studies have been made to automatically classify document images. Earlier attempts have focused on region-based analysis by detecting and analyzing certain parts of a document. Hao *et al*. [14] proposed a novel method for table detection in PDF documents based on conventional neutral networks. An alternative strategy for region-based approaches is to learn visual shared spatial configuration of a given document image [15]. In [8], a combination of holistic and region based modelling for document image classification was conducted with intra-domain transfer learning. However, many researchers utilized deep learning methods among hand-crafted feature techniques and representations, as they have shown their notable performance for the text document image classification task. In [35], transfer learning was used to improve the classification accuracy on the standard RVL-CDIP dataset, using AlexNet [19] architecture pre-trained on ImageNet dataset. A large empirical study was conducted to find what aspects of CNNs most affect the performance on document images. Results showed that CNNs trained on RVL-CDIP dataset learn region-specific layout features. Also, [2] investigated most used deep learning architectures such as AlexNet, VGG-16 [31], GoogLeNet [34] and ResNet-50 [17], using transfer learning on the RVL-CDIP and To-

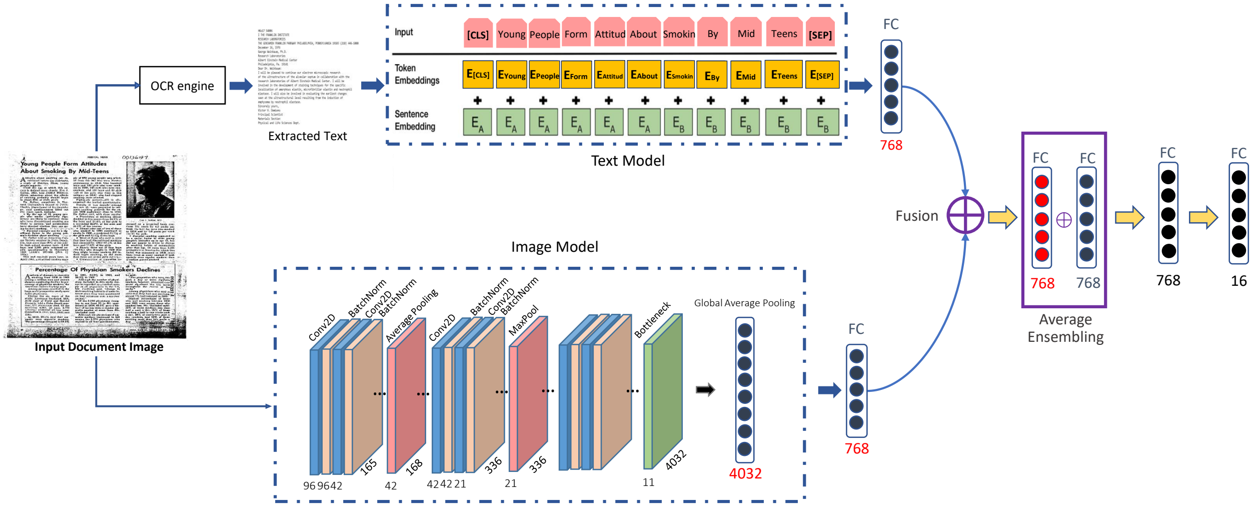


Figure 2. The proposed cross-modal deep network

bacco3482 datasets. In comparison, [22] concentrated on speed by replacing the fully connected portion of the VGG-16 architecture with extreme learning machines (ELM).

Nonetheless, document images can be characterized based on their text contents and/or their visual structural properties. Intuitively, the textual contents are mainly extracted using optical character recognition (OCR) techniques. Following [30] work, structure-based features were used to classify text content of document images. Moreover, [24] experimented with one-class SVM for document classification based on various text features like TF-IDF. The Recent appearance of learned word embedding approaches such as Word2Vec [25], Glove, ELMO [28], FastText, XLNet [38], have led to significant improvements from a natural language processing perspective. The different types of static and dynamic word embeddings aim to learn lexicon related to the words or vocabulary of a language, syntactic to create well-formed sentences in a language, semantic related to meaning in language, and pragmatic approach related to proximity between words and documents. Topic modelling used as a generative approach was performed using latent Dirichlet Allocation technique [18].

In addition, latest works focused on combining textual and visual features in a multimodal network to perform document image classification. [37] presented an end-to-end multimodal approach to extract semantic structure from documents using fully convolutional neural networks. In [23], a hybrid approach was proposed to capture contextual information using a RNN followed by a CNN to extract features. FastText word embedding, and MobileNetv2 [4] image embedding was introduced to jointly perform visual and textual feature extraction. As well, a novel approach is presented in [3], which leverages text-image features by introducing an InceptionV3 [33] network with a filter text-based feature-ranking algorithm. A modular multimodal architecture is presented in [9] for document classification followed with a XGBoost meta-classifier. Finally, [36] has proposed a novel architecture to merge textual and layout information from scanned document images in a single framework.

In this paper, we propose a cross-modal network to perform image and text feature extraction relying on off-the-shelf image-based deep networks and word embedding algorithms. We attempt to bridge the two modalities in an end-to-end network to simultaneously learn from image and text features. The built-in network is based on the performance of lightweight, heavyweight architectures used in our experiments for image stream, and static, dynamic word embeddings used to perform text classification.

3. Network architecture

This section briefly presents the deep conventional neural networks and word embedding procedures used in this work. On the one hand, we intend to investigate the impact of training lightweight and heavyweight deep networks on the classification accuracy on the RVL-CDIP dataset. On the other hand, we attempt to compare the performance of static and dynamic word embedding procedures used to generate features to process the text classification task. Fig. 2 illustrates the proposed cross-modal network.

3.1. Image stream

For the document visual embeddings, we propose to explore two well-known networks (NasNet and Inception-ResNet-v2) as backbones to extract the image features.

NasNet-A(6@4032): The NasNet architecture [39] is composed of two types of layers: Normal layer, and Reduction layer. The Normal layer is a convolutional layer that returns a feature map of the same dimension, where the Reduction layer is a convolutional layer that returns a feature map, where the feature map height and width is reduced by a factor of two. For NasNet-A(6@4032), 6 means N=6, *i.e.* number of layers repeated, 4032 means the number of filters in the penultimate layer of the network. It has 88.02 M parameters. We denote the model as NasNet$_{Large}$.

NasNet-A(4@1056): A second architecture based on the same network was studied with N=4 layers repeated and 1056 filters in the penultimate layer of the network. This light network only has 4.23 M parameters. We denote it as NasNet$_{Mobile}$.

Inception-ResNet-v2: Inception-ResNet-v2 [32] is a convolutional neural network that achieved state-of-the-art results on the ILSVRC image classification benchmark. Inception-ResNet-v2 is a variation of the earlier Inception V3 model by introducing the bypass connection as in ResNet[16]. The model has 54.36 M parameters.

3.2. Text steam

For the textual part of documents, we use three recent word-embedding mixing static and dynamic approaches for the text classification.

Glove: GloVe [27] is an unsupervised learning algorithm that generates word embeddings by aggregating global word-word co-occurrence matrix from a corpus. The resulting embeddings show interesting linear substructures of the words in vector space. We used a pretrained GloVe on Wikipedia 2014 and Gigaword 5 (6B tokens, 400K vocab, uncased, 50d vectors).

FastText: FastText [26] is a library for efficient learning of word representations and sentence classification. FastText breaks words into several character n-grams, which allows computing word representations for words that did not appear in the training data, known as out-of-vocabulary words. FastText algorithm we used was pretrained on 2 million word vectors trained on Common Crawl (600B tokens), and uses 1,999,996 word vectors.

Bert: Bert [11] is a contextualized bidirectional word embedding based on the transformer architecture. Bert representations are jointly conditioned on both left and right context in all layers, using a faster highly-efficient attention-based approach. The Bert$_{Base}$ model we will be using in this work has 12 attention layers, 768 hidden layers, 12 heads, 109.19 M parameters, and uses a vocabulary of 30,522 words.

The next section illustrates in detail the components of each stream of our proposed cross-modal approach.

4. Cross-modal feature learning

In this section, we illustrate the proposed cross-modal stream for document image classification. In the first stream, we feed input document images to the backbone model. In the second stream, we extract the textual information from document images with an OCR engine, then we feed the text strings generated as input to the word embedding algorithm. Finally, we consider a late fusion process to merge the two modalities to enhance the performance comparatively to single-modals.

4.1. Image features

Deep neural networks have exhibited their exceptional performance in both general image recognition and document image classification tasks. The concept of transfer learning from the object recognition domain was used to improve the recognition accuracy on smaller datasets. To investigate this approach more efficiently, we train the three deep CNNs discussed above pre-trained on the ImageNet weights. The image stream extracts visual features that are passed to a global average pooling layer to reduce the spatial dimensions of a three-dimensional tensor. It performs also a more extreme type of dimensionality reduction. For the final layers of the three deep CNNs, the global average pooling layer is passed to the last fully connected layer to perform classification with a softmax layer. The categorical cross-entropy loss function of softmax is given by:

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

Where $\{\mathbf{X}_1, \Theta_1\} \in \mathbb{R}^{d_1}$, and d_1 is the dimension of X_1 features of the image stream. K is the number of classes in the dataset where K=16, y_k is the one-shot label of the feature $\mathbf{X_1}$, $P(\hat{y}_k|\mathbf{X_1}, \theta_k)$ is the estimated probability of y_k calculated by the softmax function over the activation function $f^{\theta_k}(\mathbf{X_1})$, where $\{\theta_k\}_{k=1}^{K} = \Theta_1$, $\theta_k \in \mathbb{R}^{d_1}$. The bottleneck layer of the image branch is extracted as the feature $\mathbf{X}_1$ of the input image.

4.2. Text features

As textual content is required to perform text classification, we process all document images with an off-the shelf optical character recognition (OCR) engine, *i.e.* Tesseract OCR[2]. It is based on LSTM layers and includes a neural network subsystem configured in English as a text line recognizer. Besides, the output text extracted is noisy and not clean due to the different ways of presenting documents from plain, handwritten, and curved text, exotic fonts, multi-column layouts, the wide variety of tables, forms, and figures. Many word embeddings process a good tokenization of the words by getting the embedding (*i.e.*

[2] `https://github.com/tesseract-ocr/tesseract`

a vector of real numbers) for each word in the sequence, where each word is mapped to a emb_dim dimensional vector that the model will learn during training. In average, for GloVe word embedding, we found 3,581,896 unique tokens and a total number of 400,000 word vectors on RVL-CDIP corpus. As well, we found 3,601,377 unique tokens, 24,109 of null word embeddings, and a dictionary size of 3,601,377 for FastText word embedding on the same standard dataset. Contrary to traditional shallow representations (*i.e.* word2vec, GloVe, FastText), as they fail to capture higher-level information, many different dynamic word embedding procedures (*i.e.* ELMO, BERT, XLNet) have been proposed to capture semantic meaning to deal with the context-dependent nature of words. For Bert$_{Base}$ model, we processed the tokenization by splitting the input text into a 128 sequence list of tokens. To deal with out of vocabulary (OOV), Bert$_{Base}$ uses a WordPiece tokenization technique in which every OOV word is splitted into subwords. the input embeddings are computed then by summing the corresponding word embeddings, and segment embeddings. Then, the input embeddings are passed to the attention-based bidirectional transformer. After pre-processing the textual content extractred by the OCR engine form document images, we pass the input embeddings of both Glove and FastText to a GRU network of 32 nodes and 3 hidden layers. The final layers of the three models are passed to a softmax layer with categorical cross-entropy loss function.

4.3. Cross-modal features

In this part, we intend to study the effectiveness of the cross-modal features that are jointly learned from the image stream and text stream for the classification of document images. We adopt the late fusion process with two different methodologies, *i.e.* equal concatenation and average ensemble fusion. We assume that the dimension of the features extracted from the image stream or the text stream is denoted as d.

(a) Equal concatenation: We add a fully connected layer to the image stream, having the same dimensional output vector as the text stream. The final cross-modal feature is the concatenation of the two equal embedding features given by:

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

Where $X_1 \in \mathbb{R}^{d_1}$ is the obtained image embedding feature, and $X_2 \in \mathbb{R}^{d_2}$ is the text embedding feature, $d_1 = d_2$ and $|$ is the concatenation operation.

(b) Average ensembling fusion: We employ a pixel-wise addition between the image and text embedding features, *i.e.* superposing directly the two embeddings to generate the cross-modal features. Note that the obtained cross-modal features have the same dimension as the image or text embedding features.

$$X_{Av} = [X_1 + X_2], \quad X_{Av} \in \mathbb{R}^{d_1} \tag{3}$$

Training protocol: The learning of the cross-modal features include two main parts: the learning of the parameters of the image stream Θ_1 and the parameters of the text stream Θ_2. Then, the parameters of the network $\Theta = \{\Theta_1, \Theta_2\}$ are optimized by the global cross-entropy loss function $\mathcal{L}(\Theta)$ given by:

$$\mathcal{L}(\Theta) = \sum_{k=1}^{K} -y_k log P(\hat{y}_k | \mathbf{X}, \Theta) \tag{4}$$

where $\mathbf{X}$ is the cross-modal features $\mathbf{X_a}$ or $\mathbf{X_{Av}}$.

5. Experiments and analysis

5.1. Dataset

In order to evaluate the performance of the cross-modal learning approach, we introduce the publicly available RVL-CDIP dataset used in our experimentation. The Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) dataset consists of 400,00 grayscale labeled document images in 16 classes (advertisement, budget, email, file folder, form, handwritten, invoice, letter, memo, news article, presentation, questionnaire, resume, scientific publication, scientific report, specification), with 25,000 images per class. There are 320,000 training images, 40,000 validation images, and 40,000 test images.

5.2. Preprocessing of the experiments

As the DCNNs used in this paper require input images of fixed size, we first downscale all document images presented in both datasets to the expected input size of the networks. The original document images size is about 1000x750 pixels. For NasNet$_{Large}$, the images are resized to 331x331 pixels. For Inception-ResNet-v2, the images are resized to 299x299 pixels, and resized to 224x224 for NasNet$_{Mobile}$. As a data augmentation typical step, we intended to minimize the high intra-class similarity variations in document images. To do so, we applied shear transform with a range of 0.1 as in [35]. This technique is a common practice to stochastically transform each input during SGD training, to artificially enlarge the training data in order to improve the performance. Also, we randomly shifted images horizontally and vertically with a range of 0.1. For effective training, we introduced cutout data augmentation [12] that has shown its efficiency towards improving regularization of DCNNs. It consists of randomly masking a square region in an image at every training step, thus removing the redundancy of the images and augmenting the dataset by partially occluded versions of existing samples.

Table 1. The classification accuracy of the text streams for each class in RVL-CDIP dataset

Model	Adv.	Budg.	Email	File	Form	Handw.	Inv.	Letter	Memo	News	Pres.	Quest.	Res.	Public.	Report	Spec.
GloVe	0.53	0.68	0.85	0.90	0.62	0.53	0.81	0.57	0.62	0.78	0.56	0.72	0.94	0.77	0.62	0.85
FastText	0.57	0.72	0.89	0.94	0.68	0.64	0.88	0.69	0.70	0.78	0.62	0.81	0.95	0.85	0.73	0.88
Bert$_{Base}$	0.68	0.83	0.95	0.85	0.80	0.69	0.88	0.84	0.90	0.84	0.82	0.87	0.97	0.89	0.80	0.92

Table 2. The classification accuracy of the image streams for each class in RVL-CDIP dataset

Model	Adv.	Budg.	Email	File	Form	Handw.	Inv.	Letter	Memo	News	Pres.	Quest.	Res.	Public.	Report	Spec.
Inception-ResNet-v2	0.89	0.78	0.97	0.96	0.72	0.93	0.88	0.82	0.93	0.83	0.72	0.75	0.96	0.87	0.86	0.85
NasNet$_{Mobile}$	0.91	0.79	0.97	0.95	0.75	0.95	0.70	0.79	0.83	0.90	0.81	0.68	0.94	0.80	0.63	0.85
NasNet$_{Large\ (4032d)}$	0.92	0.90	0.98	0.94	0.84	0.94	0.91	0.89	0.94	0.91	0.85	0.89	0.96	0.93	0.82	0.93
NasNet$_{Large\ (768d)}$	0.94	0.90	0.98	0.96	0.83	0.95	0.93	0.90	0.93	0.92	0.85	0.89	0.96	0.93	0.82	0.93

As a final preprocessing step for image streams, we convert the grayscaled document images to RGB images.

Intuitively, the text corpus fed to the input layer of the text branch was extracted with an off-the-shelf optical character recognition OCR, *i.e.* Tesseract OCR. We utilized this OCR engine to conduct a fully automatic page segmentation, as the document images from the datasets are well-oriented and relatively clean. Since the text extracted from document images contains a lot of noise such as stop words, mis-spellings, symbols and characters, cleaning text documents is a crucial step to remove unnecessary features.

5.3. Implementation details

In this subsection, we describe the implementation details used to train the proposed single-modal and cross-modal approaches. We have trained all networks on a NVIDIA Quadro GP100 GPU, using stochastic gradient descent optimizer (SGD), with a momentum of 0.9, a learning rate of 0.001, and a step decay schedule defined as :

$$lr = initial\_lr * drop^{\left(\frac{iter}{iter\_drop}\right)} \qquad (5)$$

where $drop$ and $iter\_drop$ took a value of 0.5.

The visual streams were trained with a batch size of 16 for 50 epochs. Early stopping was considered within 5 epochs to stop training once the model's performance stops improving on the hold out validation dataset. Further, L2 regularization was applied to add a penalty for weight size to the loss function. Dropout was applied too to the final softmax layer with a probability of 0.5. For the text stream, it was trained with a batch size of 40, and a sequence length of 128 for 50 epochs. The cross-modal feature learning approach was fine-tuned using document pretraining weights obtained by the single modalities. We freezed all layers and trained our cross-modal network with both the equal concatenation and the average ensembling fusion methods, followed by the softmax layer to perform document image classification.

5.4. Overall evaluation

On the large-scale RVL-CDIP dataset, all of the adopted networks in this work achieve comparable performance with the state-of-the-art results. We report the overall accuracy results in Table. 4. The heavyweight NasNet$_{Large\ (768d)}$ model performs the best for our single image modalities at an accuracy of 91.45%, outperforming the other tested models NasNet$_{Large\ (4032d)}$, Inception-ResNet-v2, and NasNet$_{Mobile}$ at an accuracy of 91.12%, 85.04%, and 81.54% respectively. For the text modalities, the Bert$_{Base}$ model achieves comparable performance with the state-of-the-art results on the same standard dataset, with an accuracy of 84.96%. Bert$_{Base}$ manages to improve the performance thanks to its attention-based mechanism, while Glove and FastText still achieve good results on the text classification task at an accuracy of 71.54%, and 77.31% respectively. As each single modality is trained independently from one another, merging both streams boost the performance significantly for the two fusion modalities to 96.94%, 97.05% classification accuracy for equal concatenation and average ensembling respectively. Thus, exceeding the current state-of-the-art results by a 2.63% margin.

5.5. Ablation study

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 the textual content and the corresponding visual features. As seen in Table. 1, the classification results of each class of the three word embedding procedures are very low concerning three main categories that are: Advertisement, File Folder, and Handwritten. For Glove, the classification results of the three classes are 53%, 90%, and 53% respectively. Whereas for FastText, it improved slightly the accuracy results for each class to 57%, 94%, and 64% respectively. More specifically, the GloVe method predicted 36.32%, 32.66% of Advertisement and Handwritten class documents as File Folder documents. Also, FastText managed to improve the performance and reduced the classification error by 4% where 31.13% of Advertisement, and

Table 3. The classification accuracy of the cross-modal stream for each class in RVL-CDIP dataset, with the proposed fusing modalities

Model	Adv.	Budg.	Email	File	Form	Handw.	Inv.	Letter	Memo	News	Pres.	Quest.	Res.	Public.	Report	Spec.
Equal Concatenation	0.97	0.96	0.98	0.98	0.93	0.97	0.97	0.95	0.97	0.96	0.94	0.97	0.99	0.97	0.94	0.98
Average Ensembling	0.97	0.97	0.98	0.98	0.94	0.97	0.97	0.95	0.97	0.96	0.94	0.97	0.99	0.97	0.95	0.98

Table 4. The overall accuracy of the proposed methods with different backbones and different fusion modalities on the RVL-CDIP dataset

Method	Model	Accuracy(%)	Top-5 Acc	Precision	Recall	F1-Score	#Parameters
Baseline Methods	Harley *et al.* [15]	89.80	-	-	-	-	-
	Multimodal (Nicolas *et al.*) [4]	90.06	-	-	-	-	-
	Csurka *et al.* [7]	90.70	-	-	-	-	-
	Tensmeyer *et al.* [35]	90.94	-	-	-	-	-
	Azfal *et al.* [2]	90.97	-	-	-	-	-
	Single model (Das *et al.*). [8]	91.11	-	-	-	-	-
	Region-based model (Das *et al.*). [8]	92.21	-	-	-	-	-
	Multimodal (Dauphinee *et al.*) [9]	93.03	-	-	-	-	-
	Multimodal (Dauphinee *et al.*) [9]	93.07	-	-	-	-	-
	LayoutLM (Xu *et al.*) [36]	94.42	-	-	-	-	160 M
Text Stream	Glove-GRU	71.54	0.9386	0.75	0.72	0.72	179 M
	FastText-GRU	77.31	0.9515	0.80	0.78	0.78	30.47 M
	Bert$_{Base}$	84.96	0.9674	0.86	0.86	0.85	109.19 M
Image Stream	NasNet$_{Mobile}$	81.54	0.9729	0.84	0.83	0.83	4.23 M
	Inception-ResNet-v2	85.04	0.9780	0.88	0.86	0.87	54.36 M
	NasNet$_{Large\ (4032d)}$	91.12	0.9861	0.92	0.91	0.92	84.98 M
	NasNet$_{Large\ (768d)}$	91.45	0.9860	0.92	0.92	0.92	88.02 M
cross-modal Stream	NasNet$_{Large}$+Bert$_{Base}$ (Equal concat.)	**96.94**	**0.9983**	**0.97**	**0.97**	**0.97**	197.22 M
	NasNet$_{Large}$+Bert$_{Base}$ (Average Ensemb.)	**97.05**	**0.9985**	**0.97**	**0.97**	**0.97**	197.21 M

28.28% of Handwritten class documents are predicted as File Folder documents.

Furthermore, the bidirectional Bert$_{Base}$ enhanced the performance to 68% for Advertisement, 85% for File folder, and 69% for Handwritten categories. The Bert$_{Base}$ network boosted the performance of the three classes and cut the error-classification by half where 15.98% of Advertisement, and 15.84% of Handwritten categories are predicted as File folder document images. The classification errors are mainly due to either OCR error recognition, or empty document images which result to empty text files. Advertisement documents contain mostly images with few invisible text sequences, where the corresponding text generated by OCR is too much noisy and non-recognized. File Folder class presents in most cases empty document images with no text in it to be processed by the OCR engine. Finally, OCR technique fails to recognize handwritten characters in document images as a result of the different handwriting manners.

Still, all image networks trained on the RVL-CDIP dataset achieve comparable performance with the state-of-the-art methods. Table. 2 illustrates the performance of our best single image modality NasNet$_{Large\ (768d)}$. It shows an improvement in the classification results of all classes, especially for the classes Advertisement, File Folder, and Handwritten to 94.08%, 96.04%, 95.07% in comparison of text stream results. Nevertheless, the lightweight NasNet$_{Mobile}$ network fails to improve the performance for most of the classes comparatively to Bert$_{Base}$, our best text-based model. Whereas, the Inception-ResNet-v2 network slightly outperforms our text streams with 85.04% accuracy in comparison with Bert$_{Base}$ model (84.96%), surpassing significantly both Glove and FastText word embeddings.

Besides, the aim of this work is to leverage the ability of the cross-modal network to enhance the performance comparing to the single-modals. To do so, we proposed to merge textual and visual features with two different fusion modalities. For the average ensembling fusion method, it requires two feature vectors with the same size. Since text output vector is of size 768, and image output vector is of size 4032, we added a fully connected layer on top of NasNet$_{Large\ (4032)}$. We re-trained it to study its effect on the classification results. Table. 4 shows that indeed, adding a fully connected layer slightly increases the performance of the image modality from 91.12% for NasNet$_{Large\ (4032)}$, to 91.45% for NasNet$_{Large\ (768)}$. This comparison illustrates that image features are more important than text features with both feature embeddings of size 4032 and 768.

Accordingly, Tables. 3, 4, show the accuracy of each class and the overall accuracy of the cross-modal network that merges the best single-modals NasNet$_{Large}$, and Bert$_{Base}$. Jointly learning both modalities with the late fusion scheme proposed in section. 4.3, achieves accurate results in comparison with the current state-of-the-art meth-

Table 5. The Recall and Precision metrics of image backbones of most relevant classes in RVL-CDIP

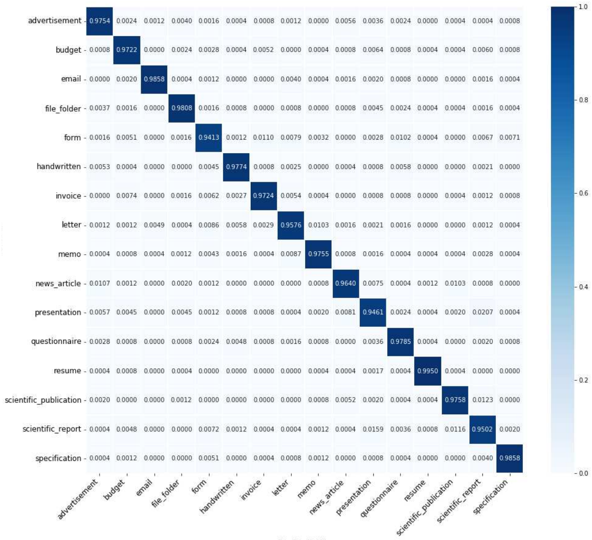
Model	Metrics	Advert.	Email	File folder	Form	Handwr.	Invoice	Presentation	Quest.	Resume	Sci.Report
NasNet$_{Mobile}$	Recall	0.91	0.97	0.96	0.75	0.95	0.71	0.82	0.69	0.95	0.63
	Precision	0.83	0.90	0.89	0.69	0.83	0.95	0.68	0.87	0.82	0.77
Inception-ResNet-v2	Recall	0.90	0.97	0.97	0.73	0.93	0.88	0.73	0.76	0.96	0.86
	Precision	0.90	0.99	0.86	0.78	0.93	0.83	0.84	0.90	0.91	0.56
NasNet$_{Large\ (4032d)}$	Recall	0.94	0.99	0.96	0.84	0.95	0.93	0.86	0.90	0.97	0.83
	Precision	0.92	0.98	0.95	0.86	0.95	0.93	0.84	0.87	0.98	0.83
NasNet$_{Large\ (768d)}$	Recall	0.93	0.99	0.95	0.84	0.92	0.92	0.85	0.89	0.97	0.82
	Precision	0.93	0.98	0.96	0.84	0.95	0.94	0.82	0.88	0.97	0.83

ods. The joint learning approach shows its capability to learn more relevant information from document images. Thus, it improves the accuracy of each class independently in comparison to single-modals. The cross-modal network manages to correct the error of the text classification generated by text-based approaches for the three main classes: Advertisement, File Folder, and Handwritten. Also, Fig. 3 shows the confusion matrix of our state-of-the-art cross-modal network with the average ensembling late fusion method. The network performs the best for the Resume category with a 99.46%, 99.50% classification accuracy for equal concatenation and average ensembling respectively. Whereas it performs the worst for the class Form with a 93.38%, 94.13% accuracy for the two fusion modalities.

To this end, we see that either Glove, FastText, or Bert are not able to outperform the image-based approaches for this task. This proves that relying only on textual content is not sufficient. Hence, it needs image features to achieve accurate results. It is clear from all reported results that combining the visual structural features with the extracted text improves the quality and accuracy of predictions for the document classification task.

5.6. Discussion

As illustrated in Table. 5, the lightweight NasNet$_{Mobile}$ framework fails to capture higher level features from Form, Invoice, Questionnaire, and Scientific_report classes. The model seems to be less sensitive with a recall rate of 75%, 71%, 69%, and 63% for the four classes respectively. Also, we measured the precision of the NasNet$_{Mobile}$ network for each class. It is less precise with a precision rate of 68%, 69% for the classes Presentation and Form. On the other hand, the Inception-ResNet-v2 framework's recall rate for the classes Form, Presentation, and Questionnaire is low in comparison with other categories. The recall for each class is of 73%, 73%, and 76% respectively, while the precision is of 78% for the class Form, with a deterioration to 56% for the class Scientific_report. Lastly, for our best heavyweight model NasNet$_{Large}$, it shows an important ability to classify document images with a lower recall and precision of 83% for the Scientific_report category. The higher recall is of 99% for the class Email, while the higher precision is of 98% for both Email and Resume classes.



	advertisement	budget	email	file_folder	form	handwritten	invoice	letter	memo	news_article	presentation	questionnaire	resume	scientific_publication	scientific_report	specification
advertisement	0.9754	0.0024	0.0012	0.0040	0.0016	0.0004	0.0008	0.0012	0.0000	0.0056	0.0036	0.0024	0.0000	0.0004	0.0004	0.0008
budget	0.0008	0.9722	0.0000	0.0024	0.0028	0.0004	0.0052	0.0000	0.0004	0.0008	0.0064	0.0008	0.0004	0.0004	0.0060	0.0008
email	0.0000	0.0020	0.9858	0.0004	0.0012	0.0000	0.0000	0.0040	0.0004	0.0016	0.0020	0.0008	0.0000	0.0000	0.0016	0.0004
file_folder	0.0037	0.0016	0.0000	0.9808	0.0016	0.0008	0.0000	0.0008	0.0000	0.0008	0.0045	0.0024	0.0004	0.0004	0.0016	0.0004
form	0.0016	0.0051	0.0000	0.0016	0.9413	0.0012	0.0110	0.0079	0.0032	0.0000	0.0028	0.0102	0.0004	0.0000	0.0067	0.0071
handwritten	0.0053	0.0004	0.0000	0.0000	0.0045	0.9774	0.0008	0.0025	0.0000	0.0004	0.0008	0.0058	0.0000	0.0000	0.0021	0.0000
invoice	0.0000	0.0074	0.0000	0.0016	0.0062	0.0027	0.9724	0.0054	0.0004	0.0000	0.0008	0.0008	0.0000	0.0004	0.0012	0.0008
letter	0.0012	0.0012	0.0049	0.0004	0.0086	0.0058	0.0029	0.9576	0.0103	0.0016	0.0021	0.0016	0.0000	0.0000	0.0012	0.0004
memo	0.0004	0.0008	0.0004	0.0012	0.0043	0.0016	0.0004	0.0087	0.9755	0.0008	0.0016	0.0004	0.0004	0.0004	0.0028	0.0004
news_article	0.0107	0.0012	0.0000	0.0020	0.0012	0.0000	0.0000	0.0000	0.0008	0.9640	0.0075	0.0004	0.0012	0.0103	0.0008	0.0000
presentation	0.0057	0.0045	0.0000	0.0045	0.0012	0.0008	0.0008	0.0004	0.0020	0.0081	0.9461	0.0024	0.0004	0.0020	0.0207	0.0004
questionnaire	0.0028	0.0008	0.0000	0.0008	0.0024	0.0048	0.0008	0.0016	0.0008	0.0000	0.0036	0.9785	0.0004	0.0000	0.0020	0.0008
resume	0.0004	0.0008	0.0000	0.0004	0.0000	0.0000	0.0000	0.0000	0.0004	0.0004	0.0017	0.0004	0.9950	0.0004	0.0000	0.0000
scientific_publication	0.0020	0.0000	0.0000	0.0012	0.0000	0.0000	0.0000	0.0000	0.0008	0.0052	0.0020	0.0004	0.0004	0.9758	0.0123	0.0000
scientific_report	0.0004	0.0048	0.0000	0.0000	0.0072	0.0012	0.0004	0.0004	0.0012	0.0004	0.0159	0.0036	0.0008	0.0116	0.9502	0.0020
specification	0.0004	0.0012	0.0000	0.0000	0.0051	0.0000	0.0004	0.0008	0.0012	0.0000	0.0008	0.0004	0.0000	0.0000	0.0040	0.9858

Figure 3. Confusion Matrix of our best cross-modal network with the average ensembling fusion method

6. Conclusion

In this paper, we propose a hybrid cross-modal methodology that learns simultaneously from the input token embeddings extracted from text corpus, and image structural information from document images to perform end-to-end document image classification. We showed that, merging the two modalities with different fusion schemes boost the performance comparatively to single-modal networks. Further, our proposed cross-modal network outperforms the current state-of-the-art result by the figure of 2.63% classification accuracy. The dynamic Bert$_{Base}$ word embedding has proved its efficiency to learn relevant semantic information from text corpus comparatively to static word embeddings, as well as the ability of heavyweight networks to learn higher level features comparing to lightweight architectures. For the future research, we will investigate early fusion schemes with different modalities, exploring new strategies that may further improve the performance for the task of document image classification.

References

[1] 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)*, pages 1111–1115, 2015.

[2] 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.

[3] 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)*, pages 1410–1416, 2019.

[4] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. Multimodal deep networks for text and image-based document classification, 2019.

[5] Yungcheol Byun and Yillbyung Lee. Form classification using dp matching. In *SAC '00*, 2000.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[7] Gabriela Csurka, Diane Larlus, Albert Gordo, and Jon Almazan. What is the right way to represent document images?, 2016.

[8] 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, 2018.

[9] Tyler Dauphinee, Nikunj Patel, and Mohammad Rashidi. Modular multimodal architecture for document classification, 2019.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[12] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

[13] Klaus Greff, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, and Jurgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, Oct 2017.

[14] L. Hao, L. Gao, X. Yi, and Z. Tang. A table detection method for pdf documents based on convolutional neural networks. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 287–292, 2016.

[15] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval, 2015.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[18] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, 2017.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[20] J. Kumar, P. Ye, and D. Doermann. Learning document structure for retrieval and classification. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1558–1561, 2012.

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

[22] 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)*, volume 01, pages 1318–1323, 2017.

[23] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *AAAI*, 2015.

[24] Larry M. Manevitz and Malik Yousef. One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154, January 2001.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[26] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

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

[28] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[29] Christian K. Shin and David S. Doermann. Document image retrieval based on layout structural similarity. In *IPCV*, 2006.

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

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[32] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[34] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[35] Chris Tensmeyer and Tony Martinez. Analysis of convolutional neural networks for document image classification, 2017.

[36] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding, 2019.

[37] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural network, 2017.

[38] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019.

[39] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition, 2017.