Multimodal Attention-Based Learning for Imbalanced Corporate Documents Classification

Ibrahim Souleiman Mahamoud[1,2](✉), Joris Voerman[1,2], Mickaël Coustaty[1], Aurélie Joseph[2], Vincent Poulain d'Andecy[2], and Jean-Marc Ogier[1]

[1] La Rochelle Université, L3i Avenue Michel Crépeau, 17042 La Rochelle, France
{joris.voerman,mickael.coustaty,jean-marc.ogier}@univ-lr.fr
[2] Yooz 1 Rue Fleming, 17000 La Rochelle, France
{ibrahim.souleimanmahamoud,aurelie.joseph,
vincent.poulaindandecy}@getyooz.com

Abstract. The corporate document classification process may rely on the use of textual approach considered separately of image features. On the opposite, some methods only use the visual content of documents while ignoring the semantic information. This semantic corresponds to an important part of corporate documents which make some classes of document impossible to distinguish effectively. The recent state-of-the-art deep learning methods propose to combine the textual content and the visual features within a multi-modal approach. In addition, corporate document classification processes offer a particular challenge for deep learning-based systems with an imbalanced corpus. Indeed the neural network performances strongly depend on the corpus used to train the network, and an imbalanced set generally entails bad final system performances. This paper proposes a multi-modal deep convolutional network with an attention model designed to classify a large variety of imbalanced corporate documents. Our proposed approach is compared to several state-of-the-art methods designed for document classification task using the textual content, the visual content and some multi-modal approaches. We obtained higher performances on our two testing datasets with an improvement of 2% on our private dataset and a 3% on the public RVL-CDIP dataset.

Keywords: Document classification · Imbalanced classifcation · Multimodal classification · Attention mechanism

1 Introduction

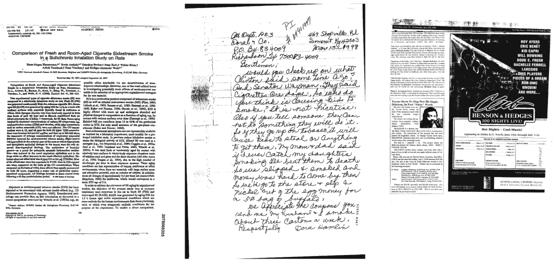
Companies need to manage each day a large number of documents. Those documents represent the life of the company and can be of a large variety of types, classes and origins. They are generally linked to the administrative part (like

J. Lladós et al. (Eds.): ICDAR 2021, LNCS 12823, pp. 223–237, 2021.
https://doi.org/10.1007/978-3-030-86334-0_15

invoices, letters, receipts) or to the core activity of the company. Such documents are of primary importance as they generally validate an action or a decision inside and/or outside the company. The management of those documents is a challenge between speed and precision. Indeed, an error could have an heavy cost by causing a wrong action or decision. In this context, precision is then preferred to recall.

Many companies use Digital Mailroom system [1] to automatize document processing and so reduce workload and time required. Those systems entries can be modeled as document streams which define many constraints. One of these constraints is a strong imbalanced representation between classes that could affect any learning methods based on a training set. Even if deep learning models recently offer impressive performance in multiple difficult situations, the document stream classification process remains a challenge. This challenge is linked to the training set used to define the model: a strong imbalance inside the training set reduce neural network performances for low represented classes and they become like noises for greater classes. Moreover, all the classes need to be known when training the system or they will be ignored and classified as a wrong known class.

Document classification processes are traditionally divided into two categories: The image processing and the textual analysis categories. Language processing methods offer good performances for classification of corporate documents where textual content is the main resource (Fig. 1a). But their performances depend on the quality of the text extraction process (generally done by Optical Characters Recognition—OCR). OCR is not perfect and an error on most important words could make the classification impossible. OCR quality has recently greatly increased but remains affected by noises and can not recognize handwritten texts (Fig. 1b). In addition, some documents classes contain few text, like advertisement (Fig. 1c). It is then hard to classify a document with this type of method. This is where the image-based approaches are more suitable. In the same way, image processing classify with efficiency documents classes with a template or with a common structure but many classes are visually very close and could only be distinguished by their semantic or few keywords.



(a) (b) (c)

Fig. 1. Three examples of documents that can only be classified based on their textual content (a), graphical content (b) or a combination of them (c)

With the spread of multimodal approaches in the deep-learning era [2], some methods have been proposed to benefit from both sides. Multimodal or Cross-modal systems propose a solution by making a decision based on a combination of graphical and textual information. The way to combine those two modalities then becomes the new main problem in order to keep the best of the two previous approaches. Recently, cross-modal methods demonstrate that they could be considered as the best ones compared to the state-of-the-art image processing system on RVL-CDIP dataset [3] and offering impressive performances.

In our paper, we want to deepen the evaluation of these new approaches in the field of imbalanced classification of document flows by comparing them with more classical methods of the state-of-the-art. In addition, we propose our multimodal system integrating an attention model per modality designed to force the system to learn the most relevant features even with the least represented classes. The second advantage is its ability to visualize features used by the deep neural network in the decision process, thus reducing the black box effect of our architectures.

2 Related Work

Few works have been proposed on multimodal analysis and classification of documents, and even fewer on the impact analysis of imbalanced number of document per class during the learning process. We will first provide an overview of articles related to one modality (based on visual or textual features), followed by the multi-modal/cross-modal approaches. The last part will focus on attention-based architectures.

Regarding the visual-based approaches, many image classification methods have been efficiently applied on the document classification task since the RVL-CDIP dataset [4] have been proposed. They are mainly based on deep convolution architecture pre-trained on ImageNet dataset [5]. Multiple architectures were proposed for this task like the InceptionResNet [6], NasNet [7] and VGG [8] for naming the most used ones. In first hand, a majority of those methods isn't especially adapted for document. In another hand, some methods propose to take into account structural aspect of documents and divided them in multiple sections like in [4].

Regarding text content based neural network, they need firstly a word representation system. Most used systems are currently word embedding like FastText [9], BERT [10] has became probably the most interesting word embedding technique and seems to be more efficient that previous methods. Those methods are based on various elder statistical strategy like bag-of-words and co-occurrences matrix. They reinforce them with deep learning method trained on huge corpus of text, with more than one million words, like Wikipedia's articles. For the network himself, current research are concentrated on recurrent strategy to keep sentence sequential information with Recurrent Neural Network (RNN) [11] and bidirectional-RNN [12].

Some recent works proposed to combine textual and visual features to take the whole document's content into account. [13] uses a MobileNetV2 [14] architecture combined with a CNN model inspired of [15] by concatenating the network outputs. [3] introduced another architecture with a NasNet [7] network for the image part and a bidirectional BERT architecture for the text. The paper tested several combinations between the networks. Even if the performances improved recently, the neural network results interpretation remains difficult. This is moreover becoming more important with the development of multimodal classification process and the need to a better understanding of limitations (which modality is penalising the other one).

Attention is one of the most influential ideas in the Deep Learning community. Even though this mechanism is now used in various problems like image captioning and others, it was initially designed in the context of Neural Machine Translation using Seq2Seq Models [16]. One important advantage of those attention models is their ability to visualize the features used by the network to make a decision, and then offer a possible interpretation of errors. This system could also be used to reinforce network performance as in [17] where the integration of an attention model into a recurrent LSTM offers better performances for image classification. Another work [18] proposed an attention model architecture combined with the VGG16 network with several layers at different resolutions to generate more global features. The attention model could force the networks to learn relevant patterns. [19] propose to use attention model to identify salient image regions at different stages of the CNN. During decision, the system reinforces their importance and therefore suppress irrelevant or confusing information.

Our work is particularly interested in the classification of an imbalanced corpus. This subject is close to two image classification challenges known as zero-shot and one-shot learning. The first introduces a context where a part of classes is unknown during training. The second is close but classes are represented by at least one or few documents. Zero-shot solutions [20] mainly use a transfer learning strategy based on semantic description of image. Our context does not offer access to such descriptions and we do not have any prior knowledge about unknown classes. So those solutions can not be used in our situation. One-shot learning methods offer more interesting option with Bayesian-based techniques like in [21] or adapted deep learning methods like in [22] and [23] but those methods are not directly compatible with our multi-modal approach. In addition, some methods are designed for imbalanced classification out of those challenge like in [24]. This propose to apply reinforcement learning with higher rewards and penalties on low represented classes.

The originality of our approach lies therefore in the proposal of a model based on attention models that allow us to have a better performance and a better interpretation of the results. In the next section, we will describe our own method for imbalanced multi-modal documents classification.

3 Problem Definition

The main objective of our proposal is to provide solutions that improve the predictive accuracy in forecasting tasks involving imbalanced dataset. The task of prediction of document classes in an industrial environment involves input flows $s_1, s_2, s_3, ..., s_n \in S$, where s_n is the number of samples from classes n in the training set. The number of samples per class is by definition imbalanced due to the industrial context. Some document classes are very recurrent (i.e. there will be thousands of invoices) while others are very rare with a very few number of samples (ten or even 1 document) as displayed in Fig. 2.

The objective of this predictive task is to predict the class of samples S. The overall assumption is that an unknown function correlates the samples and ground truth classes of S, i.e. $Cn = f(sn)$. The goal of the learning process is to provide an approximation of this unknown function whatever the quantity of samples available by class. To better approximate this function f, we must know the are significant intra-class variation and inter-class similarity caused by different structure documents for each client brings a great deal of difficulties to classify (see the Fig. 3).

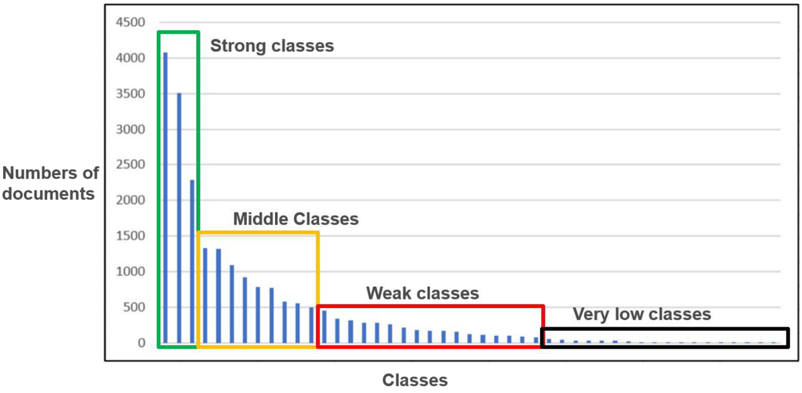


Fig. 2. Distribution of samples by classes for the private data YOOZ

The hypotheses tested in our experimental evaluation are:

1. A higher inter-class variance in the representation space will reduce the confusion between classes and also better predict weakly represented cases
2. The use of an attention mechanic for textual and visual parts will allow to better focus on the important part of the document content and thus reduce intra-class variation errors

4 Proposed Approach

To see the relevance of our hypotheses, we inspired by state-of-the-art models to propose a classification architecture with three attention models. The combination of these three attention models has not been studied in the state-of-the-art and moreover on imbalanced data. In a first step we will present the architecture without the attention mechanism and then we discuss each of the attention mechanisms.

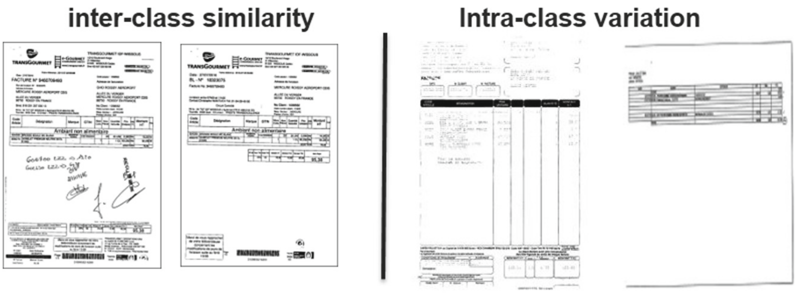


Fig. 3. The two documents on the left come from two different classes even though their structure and contents are similar. The documents on the right come from the same class and we can see that they have a different structure.

4.1 Model Without Attention

The initial system is composed of two classifiers, one for the visual part and the other one for the text extracted from the image. We use a bidirectional recurrent LSTM network with bert embedding for the text classification and a VGG16 network pre-trained on ImageNet for the image part. The choice of this architecture is inspired by the methods from the state of art for document and image classification. Here, the idea is to use the best systems from the literature as a baseline.

Each of these models has different input feature (denoted X and Z in the Eq. 1). The features of each sample i are decomposed in $x_i \in R^{d_1}$ for the textual branch and $z_i \in R^{d_2}$ for the visual branch. d_1 and d_2 are the features dimension respectively for textual and visual parts. N is the size of the corpus (i.e. $1 \leq i \leq N$).

The one-hot vector corresponding the label as $y_i \in R^L$ and L corresponds the total number of classes. The one-hot vector values are defined as 1 for the corresponding class in the groundtruth, and 0 for the others.

For the textual content, we uses the first 150 words and extracted the Camembert features [25], where each word is represented by 768 values. We then have d_1 of size (150,768). The input images are color images of size d_2 =(224,224,3) as an input for the VGG16 network.

$$\begin{aligned} X &= [x_1, x_2, ..., x_i]^T \in \mathbf{R}^{N\times d_1} \\ Z &= [z_1, z_2, ..., z_i]^T \in \mathbf{R}^{N\times d_2} \\ Y &= [y_1, y_2, .., yi]^T \in \mathbf{R}^{N\times L} \end{aligned} \tag{1}$$

4.2 Attention Model

The use of Attention Mechanism aims at focusing the network to the most relevant features related to our task (*i.e.* document classification). This corresponds to learning a matrix (or a mask) which weights features for each class. We propose to adapt this to each branch of our architecture.

Self-attention Mechanism for Label Fitting. The self-attention mechanism used in this article is inspired by [26]. They propose a self-attention mechanism on the input image to consider the inherent correlation (attention) between the input features themselves, and then use a graph neural network for the classification task. We use this self-attention on both the image and the text to focus our network on common features from the input. This will allow us to exploit the interclass and intersample correlation at the initial stage. Contrary to our baseline, we use the label Y as input and not only to help error propagation. We calculate the correlation between the input data and the label to predict. To do this, we transform X, Z to X', Z' which will be the input of our two classifiers. To achieve that, we follow several steps, the first step is to calculate the sample and label correlation matrices as:

$$\begin{aligned} C^x &= softmax(XX^T) \\ C^z &= softmax(ZZ^T) \\ C^y &= softmax(YY^T) \end{aligned} \tag{2}$$

Here softmax($\cdot$) denotes a softmax operator. The inputs of this function have the same dimension where $XX^T, ZZ^T, YY^T \in R^{N\times N}$

The self-attention module exploits C^x, C^z, C^y. Thus, the next step is to fuse Cx and Cy using trainable 1 × 1 kernels as:

$$\begin{aligned} C^{xy} &= fusion([Cx, Cy]) \in R^{N\times N} \\ C^{zy} &= fusion([Cz, Cy]) \in R^{N\times N} \end{aligned} \tag{3}$$

where e.g. [Cx, Cy] denotes the attention map concatenation, This fusion function $fusion$ is equivalent e.g. $C_x y = w_1 Cx + w_2 Cy$, where the weighted parameters w_1, w_2 are learned adaptively. We will use C^{xy}, C^{zy} to update both the visual feature X and the textual feature Z.

$$X' = XC^{xy}, Z' = ZC^{zy} \tag{4}$$

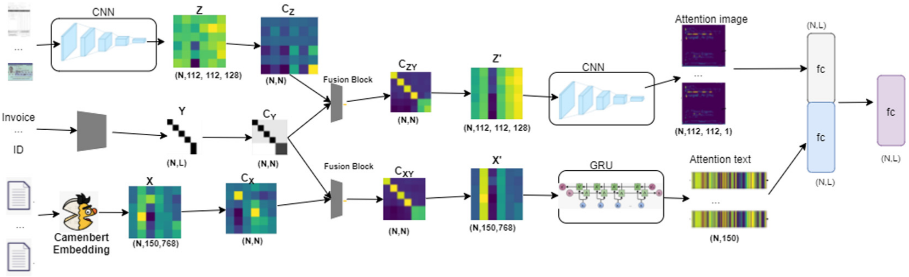


Fig. 4. The proposed multimodal model uses the three attention mechanisms described above. We used a small colored matrix to illustrate how the attention mechanism is used (Figure description see part 4)

Attention for the Textual Content. The Text Attention Model allows us to see the relevancies of the 150 words, to determine which ones have which has more impact in the classification.

$$A_t = f(softmax(h_f)) \odot h \tag{5}$$

Here $\boldsymbol{h}$ is the hidden LSTM and h_f is after having flatten $\boldsymbol{h}$, $\odot$ denotes a point-wise product operator, and softmax$(\cdot)$ denotes a row-wise softmax operator for the sample. The f function is a dropout which is intended to prevent overfitting on the data by activating neurons with probability p or kept with probability $1 - p$ (p fixed to 0.3 in our experiments). We analyze the output of the softmax to determine the words the models of which gave more weight.

Attention for the Visual Content. This mechanism is inspired by the work done by [18]. Their original goal was to predict the age of people from fixed flexion knee X-ray images. We use this image attention model as it brings two advantages: it helps understanding which part of the image is the most relevant (like the titles of document for instance); it allows understanding which part the attention model focused on, and then to understand its mistakes (for instance, we can determine the area of zones of interest it concentrates whereas it should not. The attention model embedded in the image classification network uses the output of the 5^{th} layer of the VGG16 model D. This layer is composed of 128 filters and has a dimension of (112, 112, 128). The last layer is our attention layer A and is of dimension (112, 112, 1). The matrix D is multiplied by D to obtain the matrix D'. Finally, the dimension is reduced by applying an average pooling layer on D' to generate a vector of size (128). The last step consists in normalizing this attention vector by a pooling layer on A. This normalization is the function $g(x1, x2) = x1/x2$ and the input of the visual classifier will be the output of this normalization function.

4.3 Combination of Modalities

As stated at the beginning of this paper, combining the two previous modalities should allow to take advantage of each other to increase the overall performance of the model. The proposed multimodal architecture is illustrated in Fig. 4.

First we apply self-attention mechanism on the inputs for a better separation of features in their representation space. Then we use the updated inputs by the self-attention mechanism to classify textual and visual parts, and each part is reinforced with its own specific attention mechanism. Our final classifier is based on the prediction output of the two classifiers. The following equation explain how we combined the two models:

$$[Y_x, Y_z] \in R^{N^2 \times C} \tag{6}$$

Where [Yx, Yz] is the concatenation of the outputs of the text Y_x and image Y_z classifier. We noticed that the model had some difficulties to classify the weakly represented classes. We then decided to use a weight on the Cross Entropy loss function presented in Eq. 7. The Cross Entropy [27] tends to counterbalance the weight of the under-represented classes towards the over-represented ones (the minority classes are often ignored as they represent only a small part of the total loss). The weight assigned to each class corresponds to the inverse percentage of the examples present in the training set. The smaller the presence of a class, the greater its weight on the loss will be.

$$Loss_{CE}(i) = -Wt_i log(P(i)) \tag{7}$$

5 Experiments

5.1 Dataset

To evaluate our multimodal attention-based model, we used two different datasets. The first one named “YOOZ dataset” is a private and already imbalanced set issuing from our customers. In order to assess the relevance of our method, and to have a comparison with the best approaches from the literature, we also use the public RVL-CDIP [4] dataset. This large public dataset composed of a large variety of document classes equally balanced between them. In order to evaluate our contribution, we propose a protocol to unbalance the RVL-CDIP dataset.

The YOOZ dataset is composed of 15 thousands training documents, 2 thousand for the validation set and 5 thousand for testing. This dataset is composed of 47 classes (Invoice, Quotation, general sales condition, check, etc.). Some classes are very similar from the layout and the content point of view while other classes are easily distinguishable. The Fig. 2 proposes an illustration of the classes distribution of our imbalanced dataset. One can observe that four groups appear (from the most frequent on the left to the least frequent classes on the right).

The second dataset, The RVL-CDIP, is balanced by definition with 20 000 images per class in the training set and 2500 images per class in the validation and test sets. In order to assess the relevance of our proposed approach (a multimodal attention-based classification model for imbalanced dataset), we reduced the number of document for each class. This reduction mimic the configuration of real-life conditions (some very frequent classes opposed to classes with very few samples).

To reproduce this unbalanced dataset, in the first we took the first 90 first thousand dataset(image and text) of RVL CDIP according to the order present in the train.txt in the folder labels. The exact document partition ordered by classes (0 to 15) is the following, for each class we recover a proportion to have an unbalanced distribution [5673*1, 5662*1, 5630*1, 5735*1, 5613*0.5, 5584*0.5, 5559*0.5, 5622*0.5, 5658*0.1, 5649*0.1, 5593*0.1, 5642*0.1, 5553*0.05, 5639*0.05, 5619*0.05, 5569*0.05]. So we get a total of 33 thousand documents for training. we did the same procedure to unbalance the data for the test and validation data, for each we had 16 thousand documents.

The textual content of both datasets were extracted from documents with an OCR. For our private YOOZ dataset, we used ABBYY FineReader OCR, while the OCRed version of the RVL-CDIP documents is provided by their authors (more details are available in [4]).

5.2 Implementation Details

For all our experiments, we used the ADAM optimizer [28] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a batch size of 64, an initial learning rate of 10^{-3} scaled from 0.1 every 3 epochs without improvement in validation loss and an early stopping after 5 epochs without improvement.

5.3 Performance Evaluation and Discussions

The first part of our evaluation protocol consists in comparing the proposed approach with the best approaches from the state-of-the-art. Table 1 presents the precision and recall values of all these approaches on the RVL-CDIP original dataset, its imbalanced proposed version and our internal dataset.

When looking at the textual part of the documents, we can see that even if the work done by [3] used a similar approach than ours, we can observe that we have better performances. This can be explained by the fact that the authors of [3] used a LSTM network with BERT embedding. We can here observe that the proposed attention model allows us to increase the precision by 3 points on the whole dataset.

When focusing on the visual part of the architecture, we compared our proposed attention-based pattern mechanism compared to the state-of-the-art model [29]. This model uses both holistic image and image split by region(right, left, top and bottom). We can here again observe that or complex model have better performance (+1%) with less parameters.

Finally, we present the performances of our multimodal architecture on imbalanced dataset to highlight the complementary effects of the textual and visual modalities. We can observe that the increase is much important on the imbalanced RVL-CDIP dataset than on our private one. This can be explained by the fact that classes are mostly distinguishable by their structure in the RVL-CDIP dataset, so the model relies on the image when it does not reach class with the text.

Table 1. The results of the models, the models with Att are the models with attentions either on the text or the image while S_{Att} is Att added the self attention. BiRNN is the bi-directional RNN model and VGG16 is the image classification model.

Modality	Data	RVL-CDIP		RVL-CDIP unbalanced		YOOZ	
Model		Precision	Recall	Precision	Recall	Precision	Recall
Text	*Bert (Souhail_CVPR2020)*	86%	86%	—	—	—	—
	biRNN	88.58%	87.95%	79.7%	68.5%	95.7%	94.3%
	$biRNN_{Att}$	89.5%	88.1%	80.3%	69.8	96.8%	95.2%%
	$biRNN_{S_{Att}}$	—	—	80.7%	70.2	97.5%	97.3%%
Image	*VGG16_pretained (das2018document)*	91.1%	—	—	—	—	—
	VGG16	91.8%	90.4%	87.3%	86.8%	88.1%	83.5%
	$VGG16_{Att}$	**92.2%**	**91.3%**	88.5%	87.4%	88.7%	83.2%
	$VGG16_{S_{Att}}$	—	—	88.8%	87.6%	89.2%	83.4%
Multimodality	*Multimodality (audebert19multimodal)*	90.6%	—	—	—	—	—
	Multimodality	—	—	90.8%	88.7%	96.6%	96.2%
	$Multimodality_{Att}$	—	—	92.03%	90.06%	97.3%	93.18%
	$Multimodality_{S_{Att}}$	—	—	**92.90%**	**91.40%**	**98.2%**	**97.5%**

The $Multimodality_{S_{Att}}$ propose allows us to have real gains on both RVL and YOOZ data. As previously mentioned in the industrial context prefers a model with little bad prediction, for this we took a threshold of 0.99 to filter only the cases where the model has more confidence on its prediction.

The multimodal model without model attention and with attention we almost better classify about 64 documents in dataset YOOZ as illustrated in Fig. 5(a) and (b) with threshold 0.9. When we further analyze the results class by class, we find out that the true contribution of this model lies in the reduction of intra-class errors. This contribution on almost all the classes highlight a better separation in the feature representation space.

Following this idea, we propose in our evaluation to illustrate the impact of our attention mechanism on the inter-class correlation. The proposed multimodal module exploits three attention mechanisms and this combination of attention mechanisms effectively guide the feature representation for each document and was designed to better separate classes in the feature space (thus reducing the confusion in the decision process). We then first propose to visualize the way documents are distributed in the feature space using a t-SNE visualization tool. One can observe in Fig. 5 (second line) that the classes are better seperated after applying our self-attention mechanism, demonstrating the effect of our architecture. In order to illustrate the impact of a better distributed and separated classes in the feature space, we also display the total number of error per class before and after applying our attention mechanisms (first line of Fig. 5). We can notice that this better separation entails a much lower number of errors for each class and then validate our initial assumptions.

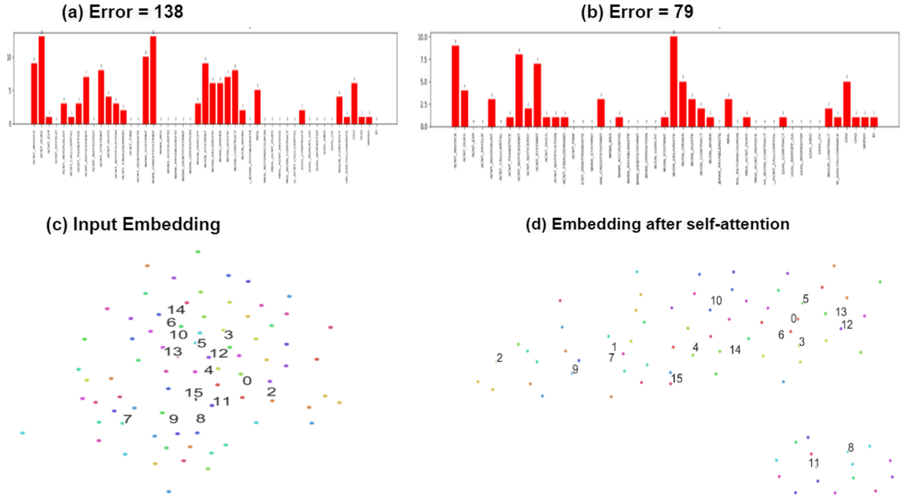


Fig. 5. In this figure we compare the errors between (a) the multimodal model with no attention and (b) the model with attentions (YOOZ dataset). These bars represent the error rate after a 99% threshold has been applied. We note that the multimodal model proposes to increase these cases of errors, which is highly appreciated in the industrial context. T-SNE visualization: (c): The embeddings without correlation, (d): The embedding created after correlation with labels (RVL-CDIP dataset). The numbers represent the class in the figure it's placed in the center of each class.

Finally, the past part of our evaluation process tends to illustrate the impact of this better separation of classes class by class. Table 2 presents the number of documents misclassified despite a high threshold of 0.99. The multimodality reduces the misclassified with threshold compared to the separate modalities for almost all classes because it manages to exploit the best of each modality. For some classes the contribution of multimodality is very low, which can be explained because one of the modalities is much more confident than the other so multimodality will use the prediction of only one modality (e.g. image) and so they will have almost the same misclassified.

Discussion. The challenge we hope to meet is to combine several modalities so that each one brings its own contribution in order to have better performance for both frequent and non-frequent classes.

Our future research will be much more focused on the training technique of our methods in order to better take into account the imbalanced data. We

Table 2. The number document of misclassified with a threshold 99 test data. hw = handwritten, adv = advertisement, s.r = scientific report, s.p=scientific publication spec = specification, fileF = file folder, newAr = news article pre = presentation, ques = questionnaire

Method	Letter	Form	email	hw	adv	s.r	s.p	spec	fileF	newAr	Budget	Invoice	pre	ques	Resume	Memo
Data	1217	1261	1289	1167	620	641	640	625	123	129	123	130	61	64	58	61
Text	380	441	220	352	467	380	322	117	28	48	36	21	30	34	26	21
Image	320	412	219	322	167	219	208	78	17	24	29	21	22	28	15	12
Multimodality	**310**	**384**	**201**	**317**	**164**	**210**	**204**	**71**	**15**	**21**	**26**	**16**	**21**	**26**	**13**	12

will test several types of specific loss for imbalanced data and will also test reinforcement learning to better train our model to force the network to be performant on the weakly represented classes.

6 Conclusion

In this article, we have proposed methods using multimodality and attention patterns on images and text. The use of multimodality is necessary in order to have better performance and take advantage of both the features extracted from the text and the image. The combination of three attention mechanisms allows focusing on the most relevant visual or semantic features to classify documents, and ease the understanding of results and mistakes. Our best proposed multimodal weighted system is able to increase by 2% the global precision compared to state-of-the-art architecture. We can also drastically reduce the number of errors, mainly by reducing the confusion between classes.

Even if we obtained good performances with our proposed multimodality approach on both the YOOZ and the RVL-CDIP dataset, many perspectives appear. The use of weighted categorical cross_entropy has certainly slightly improved the results on some classes with a low presence in the training dataset, but nevertheless we hope to get higher performance with the use of reinforcement learning to strengthen the answer when the models correctly predicts the less frequent classes. This could be done using the work proposed in [30] which describes the use of reinforcement learning in the case where the data is out of balance. We then expect improving our performances.

Acknowledgment. This research has been funded by the LabCom IDEAS under the grand number ANR-18-LCV3-0008, by the French ANRT agency (CIFRE program) and by the YOOZ company.

References

1. Schuster, D., et al.: Intellix-end-user trained information extraction for document archiving. In: 2013 12th International Conference on Document Analysis and Recognition. IEEE, pp. 101–105 (2013)
2. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep Boltzmann machines. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 2222–2230. Curran Associates Inc. (2012). http://papers.nips.cc/paper/4683-multimodal-learning-with-deep-boltzmann-machines.pdf
3. Bakkali, S., Ming, Z., Coustaty, M., Rusinol, M.: Visual and textual deep feature fusion for document image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 562–563 (2020)
4. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 991–995. IEEE (2015)
5. Russakovsky, 0, et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
6. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
7. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697–8710 (2018)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
9. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fast-text.zip: compressing text classification models, arXiv preprint arXiv:1612.03651 (2016)
10. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018)
11. Graves, A., Mohamed, A.-R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649. IEEE (2013)
12. Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., Xu, B.: Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling, arXiv preprint arXiv:1611.06639 (2016)
13. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. In: Cellier, P., Driessens, K. (eds.) ECML PKDD 2019. CCIS, vol. 1167, pp. 427–443. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43823-4_35
14. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
15. Kim, Y.: Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882 (2014)
16. Luong, M.-T., Manning, C.D.: Stanford neural machine translation systems for spoken language domains. In: Proceedings of the International Workshop on Spoken Language Translation, pp. 76–79 (2015)

17. Jain, R., Wigington, C.: Multimodal document image classification. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 71–77. IEEE (2019)
18. Górriz, M., Antony, J., McGuinness, K., Giró-i Nieto, X., O'Connor, N.E.: Assessing knee OA severity with CNN attention-based end-to-end architectures, arXiv preprint arXiv:1908.08856 (2019)
19. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.: Learn to pay attention, arXiv preprint arXiv:1804.02391 (2018)
20. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4582–4591 (2017)
21. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
22. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.: One-shot learning with memory-augmented neural networks, arXiv preprint arXiv:1605.06065 (2016)
23. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, vol. 2. Lille (2015)
24. Lin, E., Chen, Q., Qi, X.: Deep reinforcement learning for imbalanced classification. Appl. Intell. **50**(8), 2488–2502 (2020). https://doi.org/10.1007/s10489-020-01637-z
25. Martin, L.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
26. Cheng, H., Zhou, J.T., Tay, W.P., Wen, B.: Attentive graph neural networks for few-shot learning (2020)
27. Nasr, G.E., Badr, E.A., Joun, C.: Cross entropy error function in neural networks: forecasting gasoline demand. In: Applied Intelligence, pp. 1–15 (2002)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
29. Das, A., Roy, S., Bhattacharya, U., Parui, S.K.: Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks (2018)
30. Lin, E., Chen, Q., Qi, X.: Deep reinforcement learning for imbalanced classification (2019)