Apprentissage multimodal basé sur des modèles d'attention pour la classification de documents dans un contexte déséquilibré

Ibrahim Souleiman Mahamoud*, Joris voerman*, Mickaël Coustaty**
Aurélie Joseph*, Vincent Poulain d'Andecy*, Jean-Marc Ogier**

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

Résumé. Les documents administratifs ont la particularité d'être identifiables par leur contenu textuel (contenu sémantique) ou par leur mise en page (contenu visuel) et pourtant la classification de ces documents ne se fait généralement qu'à partir d'une de ces informations. Chacune d'entre elles constitue pourtant une part essentielle du document qui peut rendre impossible la distinction entre certaines classes. Les méthodes multimodales de l'état de l'art nécessitent une large base étiquetée pour l'ensemble des classes alors que dans la vie réelle les données sont généralement déséquilibrées. Nous proposons ici un modèle adapté à cette contrainte composé d'un RNN texte et d'un CNN visuel. Leur combinaison permet d'obtenir une description multimodale. Un modèle d'attention est également proposé pour chaque modalité afin de classifier plus efficacement une large variété de documents administratifs. Cette combinaison offre un gain de performance de 1% sur notre base de données privée et 3% sur la base de données publique RVL-CDIP.

1 Introduction

Les entreprises ont besoin de gérer chaque jour une grande quantité de documents. Ces documents représentent le cycle de vie de l'entreprise et sont très variés en termes de classes et d'origine. Ils sont généralement liés à la partie administrative et comptable de l'entreprise (factures, lettres, reçus, ...) ou directement associés au coeur de ses activés. De nombreuses entreprises font appel à des systèmes de "Digital Mailroom" (Schuster et al. (2013)) pour automatiser la gestion des documents. L'entrée de ces systèmes se modélise sous la forme d'un flux de documents définissant plusieurs contraintes : un fort déséquilibre de représentation entre les classes au sein de l'ensemble d'entrainement, un temps de traitements qui doit être très faible pour traiter de grands volumes de documents, ou encore la nécessité de limiter les erreurs qui

peuvent engendrer de lourdes conséquences (partage d'une information sensible, absence d'information pour la prise de décision, etc).
La classification de documents est traditionnellement divisée en deux approches : une basée sur l'analyse d'image et l'autre sur l'analyse de texte. Ces deux méthodes ont chacune leurs limites. Les méthodes basées sur l'analyse de texte offrent de bonnes performances lorsque le contenu textuel est très présent cependant il dépend fortement de la qualité de l'extraction (généralement faite par une Reconnaissance Optique de Caractères - OCR) notamment lorsque les documents contiennent des annotations manuscrites. Enfin, certaines classes de documents ne peuvent être décrites par une seule modalité. Par exemple, les publicités ne contiennent que peu de texte, tandis que les approches images ne pourront distinguer deux documents visuellement proches comme un devis et une facture. Il apparaît donc nécessaire de tirer profit des forces de chaque modalité pour extraire un résumé visuel et sémantique pertinent.

Récemment, des méthodes multimodales ont prouvé qu'elles pouvaient égaler, voire dépasser, l'état de l'art en matière d'analyse d'image sur la base de données publique RVL-CDIP (Bakkali et al. (2020)). Dans ce papier, nous voulons approfondir l'évaluation de ces nouvelles approches sur le terrain de la classification déséquilibrée de flux de documents en les comparant avec les méthodes plus classiques de l'état de l'art. De plus, nous proposons notre propre système multimodal intégrant un modèle d'attention conçu pour contraindre le système à n'entrainer que des caractéristiques pertinentes. Ce modèle est également utilisé pour faciliter l'interprétation des décisions du réseau profond en mettant en avant les caractéristiques principales calculées pendant l'entrainement et diminuer l'effet boîte noire inhérente à l'apprentissage profond.

2 État de l'art

Les méthodes de classification de documents peuvent être divisées en deux catégories : visuelles et textuelles. Les approches visuelles utilisent principalement une architecture de réseau convolutif profond pré-entrainé sur la base publique ImageNet (Russakovsky et al. (2015)). On compte parmi elles, les architectures InceptionResNet (Szegedy et al. (2017)), NasNet (Zoph et al. (2018)) et VGG (Simonyan et Zisserman (2014)) pour ne citer que celles-ci. Les approches textuelles quant à elles résument, dans un premier temps, le contenu du document via un système de représentation des mots. La stratégie la plus utilisée actuellement est la méthode dite d'encapsulement de mots (ou "word embedding") qui s'appuie sur des réseaux de neurones récurrents bidirectionnels dont les principaux représentants sont Word2Vec Mikolov et al. (2013), Fast-Text (Joulin et al. (2016)) ou plus récemment BERT (Devlin et al. (2018)).

Quelques travaux récents proposent enfin de combiner les deux approches en une seule pour prendre en compte l'ensemble du document (Sandler et al. (2018),Kim (2014)) L'article Bakkali et al. (2020) propose trois techniques pour combiner les caractéristiques extraites à partir du texte et de l'image. Bien que ce travail démontre de très bonnes performances, il n'intègre pas la question du déséquilibre entre les classes lors de l'apprentissage, et de son impact sur les performances, alors que cela est un réel problème pour les entreprises.

Le déséquilibre des classes d'un corpus pose de nombreux problèmes pour les réseaux d'apprentissage profond. Notre modèle apporte une première solution à ce problème en donnant plus de poids aux classes faiblement présentes lors de la rétropropagation via l'utilisation de modèles d'attention. L'idée est d'obliger le modèle à se concentrer sur l'essentiel Jetley

et al. (2018) afin de mieux séparer les classes. L'un des avantages principaux de ces modèles d'attention (Jain et Wigington (2019),Górriz et al. (2019)) est leur capacité à visualiser les caractéristiques utilisées par le réseau neuronal pour prendre leur décision. Cela offre ainsi la possibilité d'interpréter les erreurs sur des modèles utilisant les caractéristiques visuelles ou textuelles. Durant le processus de décision, le système renforce leur importance et supprime les informations hors sujet ou portant à confusion.

3 Modèle proposé

Le modèle multimodal proposé dans cet article se compose d'un modèle textuel et d'un modèle visuel détaillés avec leurs mécanismes d'attention respectifs. Ces mécanismes permettent de focaliser chaque modalité sur ce qui est important tout en facilitant la compréhension du modèle. Enfin, nous présentons notre approche multimodale qui permet de nous assurer que les informations fusionnées sont pertinentes et les moins redondantes possible.

3.1 Modèle d'attention

3.1.1 Module d'attention textuelle

Pour la partie textuelle nous nous appuyons sur une chaîne classique d'extraction d'information textuelle (suppression des "stop-words", des chiffres, des symboles et enfin une lemmatisation des mots). Ce contenu prétraité est ensuite résumé sous la forme d'un vecteur à l'aide du modèle Bert Devlin et al. (2018). Pour cela, nous avons converti les 250 premiers mots de chaque document en un vecteur grâce à une architecture Bert pré-entrainé sur l'anglais et le français Martin et al. (2020) car nos documents sont soit en anglais soit en français. La taille de ce vecteur est de 768. Une fois le texte résumé sous la forme d'un vecteur, nous utilisons une architecture récurrente de type LSTM afin d'apprendre les relations intrinsèques entre les mots du document. Enfin, nous avons placé notre mécanisme d'attention juste après la sortie du réseau récurrent. Le modèle d'attention utilise la sortie de notre modèle LSTM noté $\boldsymbol{h}$ de dimension (250,768). Ce modèle d'attention est composé d'une couche dense suivie d'une couche d'activation *softmax*. Cela permet d'obtenir un vecteur d'attention $\boldsymbol{t}$ de dimension (250). Le produit scalaire entre le vecteur $\boldsymbol{t}$ et $\boldsymbol{h}$ sera l'entrée du classifieur texte. Nous utilisons le vecteur $\boldsymbol{t}$ pour analyser l'attention des mots.

3.1.2 Module d'attention visuelle

Le mécanisme d'attention visuel s'inspire des travaux réalisés par Górriz et al. (2019). L'utilisation de ce modèle permet de concentrer le système sur la partie la plus pertinente d'une image (comme les logos, les signatures ou tout autre élément graphique distinctif). Afin de tester cette idée, nous proposons d'utiliser l'architecture VGG16 de Simonyan et Zisserman (2014) pour extraire les caractéristiques visuelles. Suite aux performances rapportées par Jain et Wigington (2019), nous avons utilisé une approche classique d'apprentissage par transfert, avec un modèle VGG16 pré-entraîné sur le jeu de données d'ImageNet. Nous proposons de coupler ce réseau aux travaux décrit dans Górriz et al. (2019) comme modèle d'attention. Ce modèle propose d'ajouter une couche d'activation de type sigmoïde pour générer un masque

d'attention. Ce masque permet ainsi de focaliser le système sur des zones visuelles via une fusion dans la dernière couche de l'architecture VGG16.

Le modèle d'attention incorporé dans le réseau de classification d'image utilise la sortie de la 5ème couche du modèle VGG16 ***D*** (composée de 128 filtres) et de dimension (112, 112, 128). Plusieurs couches de convolution avec un noyau de taille $1 * 1$ sont ensuite empilées avec des filtres de dimensions respectives 64 et 16. La couche d'attention correspond alors à la sortie de ce banc de filtres (***A***) et est de dimension (112, 112, 1). La matrice ***D*** est multipliée par ***A*** pour obtenir la matrice ***D'***. Enfin, une réduction de la dimension est réalisée par l'application d'une couche de type "average pooling" sur ***D'*** pour générer un vecteur ***f*** de dimension (128). La dernière étape consiste alors à normaliser ce vecteur d'attention par une couche de pooling sur ***A***. Cette normalisation est la fonction g(x1,x2)=x1/x2 et l'entrée du classifier visuel sera la sortie de cette fonction de normalisation.

3.2 Approche multimodale - texte et image

Bien que les modèles d'attention permettent d'améliorer les performances intrinsèques de chaque modalité, nous proposons en plus dans cet article, d'étudier la combinaison de celles-ci. L'objectif étant de démontrer qu'elles permettent de se compléter et d'améliorer les performances globales en termes de classification de documents. En ce sens, l'architecture multimodale proposée (voir Fig. 1) combine les deux modèles présentés précédemment. De manière pratique, Nous commencons par apprendre l'importance de chaque modalité (c'est à dire, les sorties des classifiers visuels **I1** et textuels **I2**) au travers d'un pondération W. Nous concaténons ensuite la sortie de la multiplication du poids W par les vecteurs ***I1*** et ***I2*** pour l'utiliser dans un classifieur final. Ce dernier vecteur (de dimension 2 * nombres de classes) sera l'entrée de notre classifieur final. Les couches de classification s'appuient sur la même architecture pour le modèle visuel, textuel et multimodal, avec deux couches densément connectées séparées par une couche d'oubli (dropout). Chacun de ces modèles est entraîné séparément dans un premier temps. Puis, ces modèles sont également "fine-tunés" afin d'améliorer leurs performances respectives conjointement.

Le deuxième défi adressé dans cet article concerne le déséquilibre fort entre les classes dans le flux de documents. Cela influe sur les performances globales du système, et touche en particulier les classes faiblement représentées. Afin de prendre en compte le déséquilibre entre les classes dans la phase d'apprentissage du réseau, nous proposons de pondérer ces probabilités au travers d'une d'un vecteur de poids N, comme présenté dans l'équation 1. Le poids attribué à chaque classe est le pourcentage inverse des exemples présents dans l'ensemble d'apprentissage. Ainsi, plus la présence d'une classe est faible, plus son poids durant l'entraînement sera important.

$$Loss_{CE}(i) = -Nt_i log(P(i)) \tag{1}$$

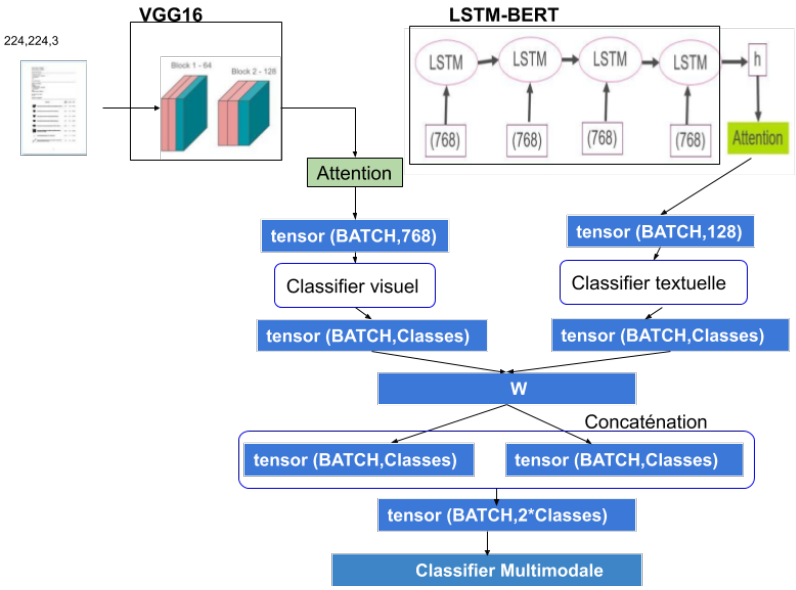

FIG. 1 – *Modèle proposé utilisant la multimodalité et le modèle d'attention*

4 Expérimentation

4.1 Données

Afin d'évaluer notre modèle multimodal basé sur l'attention, nous avons utilisé deux corpus différents. Le premier est un corpus privé de l'entreprise Yooz et déjà déséquilibré provenant de nos clients. Le corpus YOOZ se compose de 15 000 documents pour l'apprentissage, 2 000 pour la validation et 5 000 pour les tests. Il est composé de 47 classes (Facture, chèque, etc.). Certaines classes sont très similaires du point de vue de la structure et du contenu, tandis que d'autres se distinguent facilement.

Afin de permettre une comparaison équitable de notre approche avec les modèles de l'état de l'art, nous utilisons également le corpus RVL-CDIP de Harley et al. (2015). Ce vaste corpus public est composé d'une grande variété de classes de documents équilibrées entre elles. Afin d'évaluer notre contribution, nous proposons un protocole pour le déséquilibrer. RVL-CDIP est un jeu de données équilibré composé de 20 000 images par classe en apprentissage et 2500 images par classe dans les jeux de validation et de test. Afin de valider notre approche, nous avons réduit le nombre de documents pour chaque classe. Nous avons pris 100%, 50%, 10% ou 5% des documents dans chaque classe. Nous avons donc obtenu un total de 33 000 documents pour l'apprentissage et 16 000 pour la validation et les tests.

Pour notre jeu de données privé, nous avons extrait le contenu textuel des documents avec un OCR (ABBYY FineReadear 12), tandis que la version OCR des documents RVL-CDIP est fournie par leurs auteurs(Harley et al. (2015)).

4.2 Résultats

Le tableau 1 résume les résultats des principales méthodes de l'état de l'art et celles que nous proposons. Nous tenons à rappeler que les autres approches de l'état de l'art ont été entraînées sur l'ensemble des données alors que nos méthodes n'ont été entraînées que sur un ensemble déséquilibré de données. Les auteurs de la première méthode Das et al. (2018) ont utilisé un réseau de neurones VGG16 et ont obtenu une valeur de performance de 91% en precision,leurs images ont été divisées en quatre zones (en-tête, pied de page et côtés gauche et droit de la page) afin d'avoir des processus de classification spécifiques par zones. Nous pouvons observer que les systèmes que nous avons proposés ont obtenu des performances inférieures sur l'image uniquement. Ceci valide notre hypothèse qu'un ensemble d'entrainement réduit a un impact sur les performances finales (leur réseau VGG16 a été entrainé sur l'ensemble du jeu de données RVL-CDIP contrairement au nôtre).Pour la partie textuelle seule, Bakkali et al. (2020) a utilisé une approche similaire à la nôtre mais nous pouvons observer que nous avons de meilleures performances. En effet bien qu'ayant la même architecture, l'utilisation de notre modèle d'attention nous permet d'augmenter la précision d'un point.

Les meilleurs résultats pour la classification de documents sont obtenus grâce à l'utilisation d'une architecture multimodale, L'approche multimodale proposée dans Audebert et al. (2019) utilisait un réseau VGG16 et un réseau LSTM récurrent avec ELMO pour le contenu textuel. Nous comparons notre travail avec ce modèle et nous pouvons observer que nous avons une performance légèrement supérieure alors que nous avons utilisé un modèle attaché qui a été entrainé sur un ensemble de données plus petit et déséquilibré. Cela met clairement en évidence la robustesse du modèle que nous proposons pour les données déséquilibrées. La multimodalité a beaucoup mieux fonctionné sur le corpus RVL-CDIP car les classes se distinguent surtout par leur structure, de sorte que le modèle s'appuie sur l'image lorsqu'elle n'atteint pas la classe avec le texte.

Modèles	Precision	Rappel
Modèle proposé utilisant la partie visuelle (VGG16 + Attention)	83.4%	**83.2** %
Partie Visuelle (VGG16_pretained Das et al. (2018))	**91.1** %	-
Partie textuelle (BERT Bakkali et al. (2020))	86%	86%
Modèle proposé utilisant la partie textuelle (LSTM + attention)	**87.1**%	**86. 9**%
Approche multimodale Audebert et al. (2019)	90.6 %	-
Modèle multimodal proposé avec attention (LSTM + VGG16 + Attention)	91.1 %	90.7%
Multimodalité + Attention + Pondération (LSTM + VGG16 + Attention + Weighted loss)	**92.3**%	**92.0**%

TAB. 1 – *Évaluation des performances sur RVL-CDIP*

Maintenant que nous avons validé les performances de notre approche proposée sur un ensemble de données publiques, nous présentons dans le tableau 2 les performances obtenues sur notre ensemble de données privée. On peut observer qu'une fois de plus, nous avons obtenu les meilleurs résultats avec l'utilisation d'une approche mutlimodale. De plus, les documents de ce jeu de données se distinguent principalement par le texte. L'utilisation de l'image dans un seul canal ne permet pas dans la plupart des cas de classer correctement le document.

Enfin, nous pouvons conclure que l'approche proposée améliore toujours les résultats même lorsque l'ensemble d'apprentissage est réduit et déséquilibré.

Modèles	Precision	Rappel
Modèle proposé utilisant la partie visuelle (VGG16 + Attention)	87.9%	87.4%
Modèle proposé utilisant la partie textuelle (LSTM + attention)	96.2%	95.7%
Modèle multimodal proposé avec attention (LSTM + VGG16 + Attention)	96.8%	96. 1%
Multimodalité + Attention + Pondération (LSTM + VGG16 + Attention + Weighted loss)	**97.3**%	**97.18**%

TAB. 2 – *Évaluation des performances sur les données YOOZ*

5 Conclusion

Dans cet article, nous avons proposé des méthodes utilisant la multimodalité et les modèles d'attention sur le visuel et le textuel pour la classification de document. L'utilisation de la multimodalité est nécessaire afin de tirer parti des modèles textuel et de l'image pour renforcer les performances. Nous utilisons ici des modèles d'attention pour améliorer les performances mais également pour mieux interpréter les sorties des modèles. Avec l'utilisation de notre système multimodal pondéré avec attention, nous obtenons une augmentation de précision d'environ 2% comparé à de l'état de l'art. Malgré les bonnes performances que nous avons obtenues grâce à la multimodalité, tant sur notre jeu de données que sur RVL-CDIP, il reste encore beaucoup à faire. Nous avons envisagé l'utilisation d'un système d'attention multimodal et des fonctions de coût plus adaptés. D'autres pistes pour mieux extraire les mots les plus pertinents dans un document sont en cours.

Références

Audebert, N., C. Herold, K. Slimani, et C. Vidal (2019). Multimodal deep networks for text and image-based document classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 427–443. Springer.

Bakkali, S., Z. Ming, M. Coustaty, et M. Rusinol (2020). 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.

Das, A., S. Roy, U. Bhattacharya, et S. K. Parui (2018). Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Górriz, M., J. Antony, K. McGuinness, X. Giró-i Nieto, et N. E. O'Connor (2019). Assessing knee oa severity with cnn attention-based end-to-end architectures. *arXiv preprint arXiv :1908.08856*.

Harley, A. W., A. Ufkes, et K. G. Derpanis (2015). 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.

Jain, R. et C. Wigington (2019). Multimodal document image classification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 71–77. IEEE.

Jetley, S., N. A. Lord, N. Lee, et P. H. Torr (2018). Learn to pay attention. *arXiv preprint arXiv :1804.02391*.

Joulin, A., E. Grave, P. Bojanowski, M. Douze, H. Jégou, et T. Mikolov (2016). Fasttext.zip : Compressing text classification models. *arXiv preprint arXiv :1612.03651*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv :1408.5882*.

Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, et B. Sagot (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision 115*(3), 211–252.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, et L.-C. Chen (2018). Mobilenetv2 : Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520.

Schuster, D., K. Muthmann, D. Esser, A. Schill, M. Berger, C. Weidling, K. Aliyev, et A. Hofmeier (2013). Intellix–end-user trained information extraction for document archiving. In *2013 12th International Conference on Document Analysis and Recognition*, pp. 101–105. IEEE.

Simonyan, K. et A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*.

Szegedy, C., S. Ioffe, V. Vanhoucke, et A. A. Alemi (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Zoph, B., V. Vasudevan, J. Shlens, et Q. V. Le (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710.

Summary

The corporate documents classification process may rely on the use of image analysis approach considered separately of textual features.

The recent state of art deep learning methods propose to combine those two within a multimodal approach. In addition, corporate documents classification processes offer a particular challenge for deep learning based systems with an imbalanced corpus. This paper presents an evaluation of several state of the art methods designed for document classification task using the textual content, the visual content and some multi-modal approaches. We complete this evaluation with our own method, a multi-modal network with an attention model .This combination offers a performance gain of 1% for our private database and 3% for the public RVL-CDIP database.