Fusion multimodale image/texte par réseaux de neurones profonds pour la classification de documents imprimés

Thibault Magallon, Frédéric Béchet, Benoit Favre

Fusion multimodale image/texte par réseaux de neurones profonds pour la classification de documents imprimés

Thibault Magallon — Frederic Bechet — Benoit Favre

Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
prenom.nom@lis-lab.fr

RÉSUMÉ. *La classification de documents imprimés est une tâche réalisée en entrée de multiples chaînes de traitement et d'analyse d'archives numériques, ce qui en fait un point critique dans de tel systèmes. Afin d'extraire des éléments caractéristiques de chaque catégorie parmi lesquels ces pièces doivent être classés, des données textuelles ou des images sont utilisés. Nous présentons dans cet article une analyse de différentes approches pour la catégorisation de documents exploitant des données textuelles ou des images en entrée, ainsi qu'un système de classification utilisant l'information du texte et de l'image de façon jointe en un modèle de réseau de neurone convolutionnel.*

ABSTRACT. *Document classification is an important task in the analysis and processing of digital collections as it is mainly used for input pipeline of such systems. To extract features allowing algorithms to categorize the elements, text and pictures are used. We present in this paper different approaches for document classification using textual datas and pictures, as well as a classification model using both of this datas in single model of convolution neural network.*

MOTS-CLÉS : *classification de documents$_1$, fusion multimodale$_2$, réseau de neurones$_3$.*

KEYWORDS: *document classification$_1$, multimodal fusion$_2$, neural network$_3$.*

1. Introduction

La classification de documents imprimés est une tâche cruciale dans de nombreuses chaînes de traitements, qu'il s'agisse d'automatisation de tâches bureautiques, de faciliter l'indexation dans des bibliothèques d'archives numériques, ou comme prétraitement avant de procéder à des analyses de leurs contenus. Nous considérons ici la classification en *catégories* d'archives, tels que documents comptables, lettres, publicités, plans et cartes, article de presse, etc. De plus en plus de services ont besoin de cette catégorisation (Berry et Castellanos, 2004), que cela se fasse dans un contexte de gestion administrative ou dans celui de la gestion d'archives numériques. Cette utilisation intensive génère d'importants volumes de données à analyser, mais aussi une demande ne pouvant plus être gérée manuellement.

Ces besoins font aujourd'hui encore de la catégorisation de documents imprimés un champ de recherche actif, puisque même si les systèmes à l'état de l'art permettent d'obtenir des performances satisfaisantes, cette tâche reste cruciale, notamment dans l'analyse fine de contenu d'archives potentiellement dégradés (mauvaise impression, rajout d'écriture manuscrite, documents tronqués, ...). En effet, lorsqu'une erreur d'aiguillage survient en amont d'un tel système, celle-ci est alors répercutée sur le reste des traitements obligeant une intervention humaine ou pire, la mauvaise gestion d'informations critiques, par exemple dans le contexte de traitement automatique d'archives numériques comptables.

L'une des caractéristiques principales du traitement de documents imprimés est de devoir combiner à la fois du traitement d'image pour isoler les zones de textes, transcrire les caractères (Optical Character Recognition - OCR), et du traitement automatique de la langue pour l'exploitation des sorties OCR.

La catégorisation de documents peut ainsi être vue comme une tâche de classification d'images, lorsque le support s'y prête, par exemple avec l'utilisation de documents papiers scannés. Des méthodes d'habitude réservées à différencier un jeu de données d'images photos (Deng *et al.*, 2009) n'ayant aucune similarité avec les archives comme celles que nous souhaitons traiter, peuvent être appliquées avec succès, ainsi que l'ont démontré (Sicre *et al.*, 2017) avec la catégorisation de pièces d'identités.

Dans le cadre du corpus MAURDOR (Moyens AUtomatisés de Reconnaissances de DOcuments écRits) (Brunessaux *et al.*, 2014), nous proposons d'étudier un modèle de classification multimodal basé sur des réseaux de neurones profonds exploitant de manière jointe les informations issues du canal de l'image et celles obtenues par des modèles basées sur le texte provenant de l'OCR.

2. Travaux reliés

La classification de documents imprimés est une problématique antérieure aux questions d'automatisation que nous connaissons aujourd'hui avec la numérisation de

supports papier où tout simplement avec l'archivage de ces derniers. Historiquement, cette tâche est effectuée à partir de textes, cette information permettant au travers des termes qui la compose d'extraire le vocabulaire d'une thématique ou d'un sens commun à une catégorie. À cet effet, des approches par sacs de mots suivi d'algorithmes de classification supervisé, par Boosting (Schapire et Singer, 2000) ou réseau convolutionnel (Johnson et Zhang, 2014) sont réalisés.

La catégorisation de documents papiers scannés peut aussi utiliser pour ce même objectif les images, afin d'apprécier la similarité structurelle de ces dernières (Kumar *et al.*, 2014) pour une classe donnée. Plus récemment, il fut démontré un possible transfert d'utilisation de caractéristiques ayant été apprises par des réseaux convolutionnels sur des images autres que celle de documents pour cette même tâche (Harley *et al.*, 2015).

D'autres tâches telles que l'analyse de vidéo ont su tirer parti d'une fusion des différentes sources qui la compose (Ngiam *et al.*, 2011). Il existe cependant peu de méthodes faisant usage des informations provenant des deux modalités sources qu'offrent les documents numériques, tel que (Chen *et al.*, 2009), qui reposent sur l'utilisation de SVM et de Boosting. Plus récemment et dans un autre contexte pouvant être associé à une catégorisation, les systèmes de labellisation de photos utilisent des représentations jointes d'images et de textes pour faciliter l'indexation et la recherche des ces dernières (Chen *et al.*, 2013 ; Gong *et al.*, 2013). Nous proposons dans ces travaux de suivre ce même raisonnement et de comparer les différentes méthodes de catégorisation existantes par modalités séparées, puis par fusion au travers d'un réseau convolutionnel pour la tâche de classification de documents.

3. Le corpus MAURDOR

La campagne MAURDOR vise à évaluer les systèmes de traitement automatique de documents numérisés à travers la quantification et la qualification de ces derniers à extraire des informations pertinentes. Cette campagne met l'accent sur la recherche de solutions performantes de Reconnaissance Optique de Caractères (OCR) pour des documents de nature et de contenu variés, issus de situations réelles. Cette campagne d'évaluation a été organisée par le Laboratoire National de métrologie et d'Essais (LNE) et l'entreprise CASSIDIAN, avec un financement de la Direction Générale de l'Armement (DGA) au cours de deux éditions, l'une en 2013 et la seconde en 2014.

Le corpus mis à disposition regroupe 10 000 numérisations annotés, dont 8 129 sont fournis comme données d'entraînements (comprenant un corpus de développement et de test de 1 000 éléments chacun) et 1 871 documents utilisés lors des deux sessions d'évaluation.

Les annotations regroupent des informations concernant les zones sémantiques présentes dans chaque document, à savoir les blocs contenant des photos, des contenus graphiques, des tableaux, des zones bruitées ou tâchées, ou encore du texte. Pour chacune d'elles, des informations concernant leur contenu (pour le texte), leur fonction

dans le document ou encore les relations qu'elles observent entre elles lorsque cela est possible. Il est à noter que le texte présent dans les documents peut être rédigé en plusieurs langues, dont le Français (50%), l'Anglais (25%) et l'Arabe (25%). Chaque annotation fait ainsi référence à un langage principal mais est à même de contenir une à quatre autres langues (le contenu multilingue étant évalué à 10%). De plus, ces documents sont regroupés en cinq différentes catégories (C1 à C5) comportant des documents spécifiques tels que les formulaires imprimés, les documents commerciaux, les correspondances privées manuscrites, les correspondances professionnelles ou privées imprimées et enfin les autres documents.

3.1. *Le sous-corpus utilisé*

Pour mener à bien nos expériences sur la classification de documents, nous n'utilisons qu'une sous-partie du corpus original, correspondant à l'une des cinq catégories précédemment citées. Cette dernière n'est autre que la seconde regroupant les archives issues de pièces à but commerciales, plus particulièrement, ceux rédigés en français. Nous avons fait ce choix car, parmi ces groupes, certains souffrent d'un manque de représentativité dans les zones sémantiques d'intérêt, comme le décrit le Tableau 1. De plus, si le corpus MAURDOR original différencie ces cinq catégories, il n'est cependant pas annoté à une granularité plus réduite. Or, nous ne souhaitons pas orienter la classification vers des catégories de documents possédant de trop grandes disparités dans la représentation de leurs zones sémantiques, cela les rendant bien plus facilement différentiables.

Sept classes ont été retenues parmi la seconde catégorie du corpus, correspondant respectivement aux tracts, aux articles de journaux ou pages de magazines, bons de livraisons, plans ou cartes, les documents personnels de types diplômes ou visas, les pièces comptables tel que des factures ou notes de frais et pour finir les contrats. Cette sous-catégorisation du recueil initial compte 1 280 éléments pour le corpus d'entraînement, 228 pour celui de développement et 177 pour celui de test. La répartition de ces classes sur l'ensemble de la catégorie C2 est quant à elle consultable dans le Tableau 2.

3.2. *Les transcriptions OCR du corpus MAURDOR*

En complément des fichiers images de chaque document, différentes transcriptions du texte contenu dans chacun d'eux sont disponibles dans le corpus MAURDOR. Outre la transcription de référence, les sorties OCR des zones de texte provenant de plusieurs participants à la campagne d'évaluation sont disponibles. Ces sorties sont l'occasion de tester les méthodes de classification de documents à partir du texte selon plusieurs niveaux de *bruit* dans les transcriptions.

		Texte	Image	Graphique	Ligne	Tableau
TRAIN	C1	3.92	x	3.66	3.51	0.02
	C2	3.28	1.23	1.31	0.43	0.004
	C3	1.37	x	1.01	x	x
	C4	1.45	x	1.03	0.13	x
	C5	6.52	x	6.16	0.24	0.44
DEV	C1	3.72	x	3.51	3.33	x
	C2	2.04	0.22	1.44	0.40	0.04
	C3	1.41	x	1.00	x	x
	C4	1.51	x	1.04	0.11	x
	C5	1.81	x	1.54	x	0.27
TEST	C1	3.69	x	3.57	3.40	0.03
	C2	2.06	0.89	1.13	0.37	x
	C3	1.37	x	1.01	x	x
	C4	1.26	x	1.06	0.31	0.02
	C5	x	x	x	x	x

Tableau 1. *Moyenne d'apparition des différents blocs sémantiques par document pour chaque catégorie du corpus*

	Tracts	Journaux et Magazines	Bons de livraison	Plans et Cartes	Documents personnels	Documents comptables	Contrats
TRAIN	17.50%	17.66%	2.19%	11.95%	7.73%	28.85%	14.14%
DEV	26.32%	11.84%	2.19%	8.77%	9.21%	24.12%	17.54%
TEST	7.91%	35.03%	1.13%	22.03%	12.43%	11.30%	10.17%
Eval 2013	26.32%	11.84%	2.63%	9.65%	8.77%	22.81%	17.98%
Eval 2014	7.91%	35.03%	1.13%	22.60%	12.43%	10.73%	10.17%

Tableau 2. *Répartitions des classes dans les différents corpus*

Comme on peut le voir dans le Tableau 3, les taux d'erreurs en caractères (CER) et en mots (WER) sont élevés, en particulier pour les participants 4 à 6, montrant ainsi la difficulté de la tâche et des documents à traiter.

4. Modèles de classification utilisant uniquement le texte

Afin de mesurer les performances de classifieurs sur le texte uniquement, nous avons comparé plusieurs approches utilisant 3 paradigmes différents :

		CER	WER
Eval2013	participant 1	24.42%	40.63%
	participant 3	27.78%	37.76%
	participant 4	87.46%	153.40%
	participant 5	37.46%	74.45%
	participant 6	93.45%	94.56%
Eval2014	participant 1	8.29%	14.66%
	participant 4	76.95%	92.05%
	participant 5	29.41%	50.52%
	participant 6	91.03%	92.91%

Tableau 3. *Taux d'erreurs en caractères (Character Error Rate - CER) et en mots (Word Error Rate - WER) parmi les sorties OCR des différents participants à la campagne MAURDOR*

– les sacs de mots : méthodes Naïve Bayes et Support Vector Machine sur des sacs de termes

– les sacs de n-grammes de mots : méthode AdaBoost sur des séquences de mots

– réseaux de convolution : utilisation de réseaux de neurones convolutionnels (CNN)

4.1. *Approches sacs de mots*

Comme nous ne pouvons donner en entrée à ces méthodes du texte dans sa forme brute, il nous faut constituer un jeu de caractéristiques (ou *traits*) représentatif de chaque extrait textuel pour chaque document. Pour ce faire, une pratique courante est d'utiliser des traits provenant de *sacs de mots*. Ces dernières ne sont autre qu'une collection de mots mise sous forme de vecteur de taille fixe à une dimension. Le corpus d'entraînement est alors utilisé pour collecter les différents symboles du dictionnaire, pouvant amener à une représentation vectorielle. Il est à noter que chaque mot est lemmatisé de façon à retirer les suffixes des terminaisons de verbes ou de genre et de nombre. Cela est fait à l'aide de l'algorithme *Snowball* de la librairie NLTK (Bird et Loper, 2004). De plus, les mots syntaxiques, tel que les pronoms ou les articles, sont retirés avant de constituer le lexique. L'un des paramètres importants de ce type de modèle est le nombre de mots conservés pour fixer la taille du vecteur. Nous avons simplement considéré ici les n-mots les plus fréquents avec $n = 100$ ou $n = 2000$.

Deux type de méthodes ont été testées sur ces sacs de mots : un classifieur Naïve Bayes et un Support Vector Machine (SVM).

4.2. *Approche par boosting de n-gramme*

La représentation par sac de mots a un inconvénient majeur : elle ne tient pas compte ni du contexte d'apparition de chaque mot composant le document à analyser. En considérant des petits arbres de décision représentant des n-grammes de ces derniers de longueur variable, le *boosting* de classifieurs faibles représente une alternative intéressante à la représentation par sac en implémentant une certaine dépendance au contexte dans la prise en compte des mots.

Nous avons utilisé l'implémentation *ICSIBoost* (Schapire et Singer, 2000) de l'algorithme de boosting AdaBoost, avec des n-grammes de taille 2 et 1 000 itérations lors de l'entraînement.

4.3. *Approche par réseaux de neurones convolutionnels*

Afin de prendre en compte un contexte plus large que des simples séquences de n-grammes, nous avons implémenté un réseau de neurones à convolution (CNN) utilisant les 32 premiers mots inclus dans le vocabulaire apparaissant dans le texte de chaque document en entrée. L'utilisation d'un tel type de structure peut être vu comme une équivalence aux n-grammes, puisque cette dernière se focalise sur des enchaînements locaux via l'utilisation de filtres, qui peuvent être vus comme le déplacement d'une fenêtre le long de la séquence textuelle. Cette architecture est comparable au modèle utilisé par (Johnson et Zhang, 2014), puisqu'il est composé d'une couche de représentation vectorielle des mots pré-entraînée et non raffinée, selon le modèle GloVe (Pennington *et al.*, 2014) ayant une dimension de 100 pour un vocabulaire comptant les 10 000 mots les plus fréquents, suivit d'une couche de convolution de 200 filtres de taille 3 avec activation ReLU suivi d'un maxpooling d'une fenêtre de 3, puis consécutivement d'une couche dense de 1 024 unités et de la couche de sortie pour nos 7 classes sur laquelle est appliqué une activation softmax. De plus, un dropout avec une probabilité de 0,5 est utilisé avant chacune de ces dernières.

4.4. *Évaluation*

Les résultats sur les transcriptions de référence des corpus de développement et de test sont regroupés dans le Tableau 4. On y observe que les caractéristiques décrivant les documents sous une forme de séquence de n-grammes (IcsiBoost) sur l'ensemble du vocabulaire donnent les meilleures performances. Les modèles de classifications par sac de mots, tel que les SVMs (Chang et Lin, 2011) ne parviennent pas aux même résultats, même si ces dernières semblent plus performantes lorsqu'un nombre réduit de termes est utilisé. Les résultats par réseau convolutionnel sont bien meilleurs que les approches par sacs de mots mais sont inférieurs d'environ 3% à ceux obtenus par IcsiBoost. Cela s'explique sans doute par la faible taille du corpus d'apprentissage.

	DEV	TEST
Naïve Bayésienne 100 mots	79%	78%
SVM 100 mots	88%	84%
SVM 2000 mots	47%	42%
IcsiBoost	**96.50%**	**97.75%**
CNN Texte	96.49%	94.92%

Tableau 4. *Résultats de différentes méthodes de classification de documents (7 classes) utilisant les données textuelles de référence.*

En complément des résultats obtenus sur les transcriptions de référence, nous avons voulu mesurer l'impact des erreurs d'OCR sur la classification de documents. Si l'on observe les résultats utilisant Icsiboost (que nous considéreront comme notre *baseline* pour l'approche texte) sur le Tableau 6, on note une dégradation proportionnelle des performances à l'augmentation des erreurs présentes dans le texte.

Comme on peut le voir, l'impact des erreurs d'OCR est très important sur les performances de la classification. C'est dans cette optique que nous avons voulu développer des approches pouvant être robustes à la dégradation des sorties OCR en considérant l'ensemble de l'image d'un document pour réaliser cette tâche, et non plus uniquement le contenu de ses boites textuelles.

5. Classification multimodale image/texte de documents imprimés

5.1. *Classification utilisant uniquement l'image*

Pour la catégorisation de documents disposant comme seule source d'information les images, nous avons utilisé une implémentation du réseau AlexNet (Krizhevsky *et al.*, 2012), que nous considérerons comme notre *baseline* pour la modalité d'image. Ce dernier utilise des réseaux de Convolution profonds pour extraire les caractéristiques des images. Si celui-ci est connu pour ses résultats sur la classification d'images, il a aussi été utilisé avec succès pour la classification de documents (Afzal *et al.*, 2015).

Nous avons adapté la taille des images de chaque document en la réduisant à des carrés de *227x227* pixels, afin d'entraîner seulement les trois dernières couches du modèle de classification, dont les poids chargés pour les autres couches correspondaient à un entraînement sur le corpus d'ImageNet. Cette stratégie de *finetuning* permettant de conserver les extractions de traits d'images faites sur une grande quantité de données que nous ne possédons pas sur ce corpus de taille réduite.

Lors de l'entraînement, nous avons utilisé le corpus de développement comme validation afin de prévenir tout sur-apprentissage. Sur les données d'entraînement, ce modèle est fiable à 99,22%, sur celles de développement les bonnes classifications montent à 92,19% et 96,88% pour les images du corpus de test.

5.2. *Fusion des classifications Image-Texte*

Pour procéder à une fusion des modèles provenant des images et du texte des différents documents, nous devons utiliser une représentation commune basée sur les réseaux de neurones. En effet, ceci nous permet de d'obtenir directement une représentation interne des caractéristiques d'images ayant été extraites par le réseau AlexNet avec celles du texte. Même si le réseau CNN utilisant les données textuelles n'est pas le modèle ayant obtenu les meilleurs résultats, nous espérons que sa fusion avec celui basée sur les images puisse améliorer globalement les performances.

Pour la fusion des représentations d'images et de texte issue de ces deux modèles de réseaux à convolution, nous avons concaténé les deux sorties des dernières couches en une seule, afin de pouvoir l'acheminer vers une nouvelle couche dense avant d'appliquer une dernière couche d'activation *softmax*. Une représentation schématique du modèle est donné dans la Figure 1. Nous avons par la suite entraîné seulement la nouvelle partie du réseau de la même manière que les précédents, en prenant le corpus de développement comme validation.

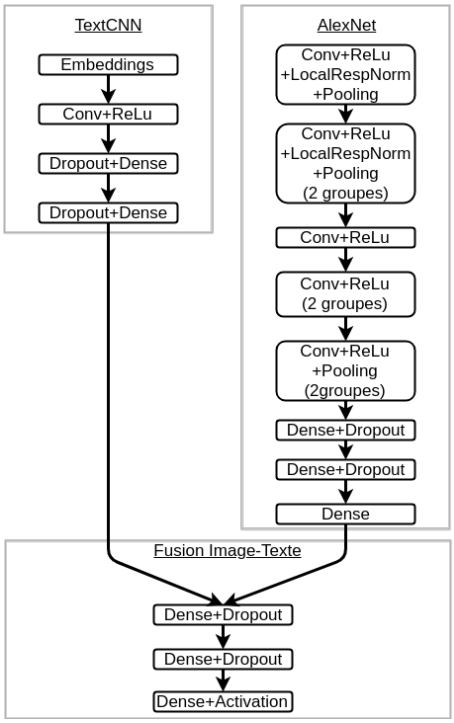

Figure 1. *Schématisation du modèle de classification image-texte.*

.

Le Tableau 5 présente les résultats des réseaux convolutionnels sur l'image et le texte, ainsi que la fusion des modèles. Les résultats de la meilleure méthode sur le

	Corpus de Test
Baseline Texte	97.75%
Baseline Image	96,88%
CNN Texte	94.92%
CNN Fusion	**98.31%**

Tableau 5. *Résultats des méthode Image, Texte et Fusion sur la tâche de classification de documents (7 classes)*

texte (IcsiBoost) et l'image sont aussi rappelés. Comme on peut le voir le modèle de fusion donne les meilleurs résultats, même si le modèle textuel n'était pas le meilleur des systèmes lorsqu'il était utilisé seul, confortant en cela l'intérêt de la fusion multimodale pour le traitement de documents texte/image.

De plus, on peut observer sur le Tableau 6 que les résultats montrent une amélioration dans la majorité des cas lorsque la méthode de fusion des modèles est employé, même lorsque des textes bruités par de nombreuses erreurs sont utilisés. Par exemple, pour le participant P4 sur *Eval 2014*, les performances avec uniquement le texte passent de 49.15% à 96.05% avec le réseau multimodal. Pour la plupart des participants, combiner texte et image apporte aussi une amélioration.

On note cependant que dans deux situations la méthode de fusion ne donne pas les meilleurs résultats, même si elle donne des performances très proches. Pour les données issues de la référence de l'évaluation de 2013, c'est le *CNN Texte* qui parvient à obtenir le meilleur score ; pour l'évaluation 2014, dans le cas de l'OCR du participant 4 qui présente un taux d'erreurs mots très élevé, c'est l'image seule qui donne les meilleures performances. Nous pouvons conclure de ces deux situations que la fusion multimodale est particulièrement intéressante lorsque les performances du classifieur visuel seul sont plus élevées que celles du classifieur textuel : si le texte seul obtient d'excellents résultats, l'image n'apporte rien. Dans le cas ou l'écart entre les performances des classifieurs est très élevé, comme dans le cas où les taux d'erreurs d'OCR sont très importants, la fusion apporte peu, voire dégrade légèrement les résultats comme pour l'éval 2014, participant 4.

6. Conclusion

Nous avons présenté dans cet article des approches de classification procédant à une extraction de caractéristiques à l'aide de données textuelles, mais aussi d'images pour catégoriser des documents numérisés provenant du corpus MAURDOR. De plus, nous avons montré au travers de la présentation d'un modèle de réseaux de neurones convolutionnel les gains de performances pouvant être engendrés par la fusion d'informations image-texte jumelées au sein d'un même système de classification.

Eval 2013						
Baseline Image	91.02%					
	Ref.	P1	P3	P4	P5	P6
WER	-	*40.63%*	*37.76%*	*153.40%*	*74.45%*	*94.56%*
Baseline Texte	95.61%	71.49%	81.14%	23.68%	76.32%	31.14%
CNN Texte	**96.05%**	81.14%	80.70%	14.04%	69.30%	15.35%
CNN Fusion	95.61%	**95.61%**	**95.18%**	**92.11%**	**95.18%**	**92.98%**
Eval 2014						
Baseline Image	96.35%					
	Ref.	P1	P4	P5	P6	
WER	-	*14.66%*	*92.05%*	*50.52%*	*92.91%*	
Baseline Texte	94.35%	92.09%	63.28%	79.66%	20.34%	
CNN Texte	94.35%	90.96%	49.15%	80.79%	20.90%	
CNN Fusion	**97.74%**	**97.18%**	96.05%	**96.61%**	**97.18%**	

Tableau 6. *Résultats comparatifs pour les sorties textes des participants sur la tâche de classification de documents (7 classes)*

7. Bibliographie

Afzal M. Z., Capobianco S., Malik M. I., Marinai S., Breuel T. M., Dengel A., Liwicki M., « DeepDocClassifier : Document classification with deep convolutional neural network », *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, IEEE, p. 1111-1115, 2015.

Berry M. W., Castellanos M., « Survey of text mining », *Computing Reviews*, vol. 45, n° 9, p. 548, 2004.

Bird S., Loper E., « NLTK : the natural language toolkit », *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, Association for Computational Linguistics, p. 31, 2004.

Brunessaux S., Giroux P., Grilhères B., Manta M., Bodin M., Choukri K., Galibert O., Kahn J., « The Maurdor Project : Improving Automatic Processing of Digital Documents », *2014 11th IAPR International Workshop on Document Analysis Systems*p. 349-354, 2014.

Chang C.-C., Lin C.-J., « LIBSVM : a library for support vector machines », *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, n° 3, p. 27, 2011.

Chen M., Zheng A., Weinberger K., « Fast image tagging », *International conference on machine learning*, p. 1274-1282, 2013.

Chen S. D., Monga V., Moulin P., « Meta-classifiers for multimodal document classification », *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*, IEEE, p. 1-6, 2009.

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., « Imagenet : A large-scale hierarchical image database », *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, p. 248-255, 2009.

Gong Y., Jia Y., Leung T., Toshev A., Ioffe S., « Deep convolutional ranking for multilabel image annotation », *arXiv preprint arXiv :1312.4894*, 2013.

Harley A. W., Ufkes A., Derpanis K. G., « Evaluation of deep convolutional nets for document image classification and retrieval », *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, IEEE, p. 991-995, 2015.

Johnson R., Zhang T., « Effective use of word order for text categorization with convolutional neural networks », *arXiv preprint arXiv :1412.1058*, 2014.

Krizhevsky A., Sutskever I., Hinton G. E., « ImageNet Classification with Deep Convolutional Neural Networks », *in* F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., p. 1097-1105, 2012.

Kumar J., Ye P., Doermann D., « Structural similarity for document image classification and retrieval », *Pattern Recognition Letters*, vol. 43, p. 119-126, 2014.

Ngiam J., Khosla A., Kim M., Nam J., Lee H., Ng A. Y., « Multimodal deep learning », *Proceedings of the 28th international conference on machine learning (ICML-11)*, p. 689-696, 2011.

Pennington J., Socher R., Manning C., « Glove : Global vectors for word representation », *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532-1543, 2014.

Schapire R. E., Singer Y., « BoosTexter : A boosting-based system for text categorization », *Machine learning*, vol. 39, n° 2-3, p. 135-168, 2000.

Sicre R., Awal A. M., Furon T., « Identity documents classification as an image classification problem », *International Conference on Image Analysis and Processing*, Springer, p. 602-613, 2017.