Automatic classification of company's document stream: Comparison of two solutions

Joris Voerman, Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain D'andecy, Jean-Marc Ogier

▶ **To cite this version:**

Joris Voerman, Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain D'andecy, et al.. Automatic classification of company's document stream: Comparison of two solutions. Pattern Recognition Letters, 2023, 172, pp.181-187. 10.1016/j.patrec.2023.06.012 . hal-04678432

HAL Id: hal-04678432
https://hal.science/hal-04678432v1

Submitted on 9 Jul 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Version of Record: https://www.sciencedirect.com/science/article/pii/S0167865523001915
Manuscript_c4e33e972be220923669b1c372308a6f

ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

Automatic classification of company's document stream: comparison of two solutions

Joris Voerman[a], Ibrahim Souleiman Mahamoud[a], Mickael Coustaty[a], Aurélie Joseph[b], Vincent Poulain d'Andecy[b], Jean-Marc Ogier[a]

[a] *La Rochelle Universite, L3i, Avenue Michel Crepeau, La Rochelle, 17042, France*
[b] *Yooz, 1 Rue Fleming, La Rochelle, 17000, France*

Article history:

Document Processing, Imbalanced Classification, Neural Network

ABSTRACT

Documents are essential nowadays and present everywhere. In order to manage the vast amount of documents managed by companies, a first step consists in automatically determining the type of the document (its class). Even if automatic classification has been widely studied in the state of the art, the strongly imbalanced context and industrial constraints bring new challenges which were not studied till now: how to classify as many documents as possible with the highest precision, in an imbalanced context and with some classes missing during training? To this end, this paper proposes to study two different solutions to address these issues. The first is a multimodal neural network reinforced by an attention model and an adapted loss function that is able to classify a great variety of documents. The second is a combination method that uses a cascade of systems to offer a gradual solution for each issue. These two options provide good results as well in ideal context than in imbalanced context. This comparison outlines the limitations and the future challenges.

© 2023 Elsevier Ltd. All rights reserved.

1. Introduction

Private companies and public administrations have to deal with huge amounts of documents every day [1]. These documents come from multiple sources, whether from internal processes or from external entities such as subcontractors or other administrations. Processing such a large amount of documents requires a lot of resources if an automatic system is not used. Moreover,

e-mail: joris.voerman@univ-lr.fr (Joris Voerman), ibrahim.souleiman_mahamoud@univ-lr.fr (Ibrahim Souleiman Mahamoud), mickael.coustaty@univ-lr.fr (Mickael Coustaty), aurelie.joseph@getyooz.com (Aurélie Joseph), vincent.poulaindandecy@getyooz.com (Vincent Poulain d'Andecy), jean-marc.ogier@univ-lr.fr (Jean-Marc Ogier)

© 2023 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

these documents are usually related to the core business of a company. They are therefore of the utmost importance as they can be used to validate actions or decisions with internal and/or external effects. The management of these documents becomes a challenge of speed and precision, since any error can have serious consequences by causing erroneous actions or decisions, but also losses of possibly crucial information. The implications are diverse, ranging from a simple unpaid invoice to the classification of an urgent email as a document for archiving. In this context, it is necessary for the automatic processing system to have high precision, or at least as high as possible. Consequently, Document processing is a major concern for many actors.

Our partner, Yooz, seeks to offer a solution of document processing for small and medium-sized companies through a generalist web service that allows the automatic processing of a company's internal document stream without requiring too many adaptations. This solution is intended to be complementary to the existing system and much more adapted to those who cannot be satisfied by bigger software, too costly or too complex to set up.

Consequently, the objective here is to propose a generic system allowing the automatic processing of a heterogeneous set of documents from a company while satisfying specific industrial constraints. The first is to reduce as much as possible the number of parameters, to ensure accessibility to non-expert users. The second is to minimize errors to ensure the reliability of results even with more rejections. An error is considered here to be always more costly than a rejection (while keeping the classification rate as high as possible). The last is to have a low processing time per document in order to deal with the quantity to be processed, the order of magnitude should be close to a second.

2. Industrial stream classification issues

To model this input, we based ourselves on the "document stream" proposed by [2]. A document stream is defined as a sequence of very heterogeneous documents that appear over time. The stream is composed of numerous classes more or less close to each other and very unevenly represented (hence its heterogeneity). As a result, any training corpus that we want to generate from a document stream will inherit two constraining properties: it will be imbalanced and incomplete. The imbalanced property is linked to its uneven distribution of documents (see Fig. 1). It takes the form of a small set of strong classes which constitute the core of the stream and a large number of smaller classes. The incompleteness comes from the sequential aspect of document streams: the quantity of documents for each class will increase as time goes on and the stream evolves. However, new classes may appear spontaneously and others may disappear. Consequently, a portion of the classes is missing during the training phase.

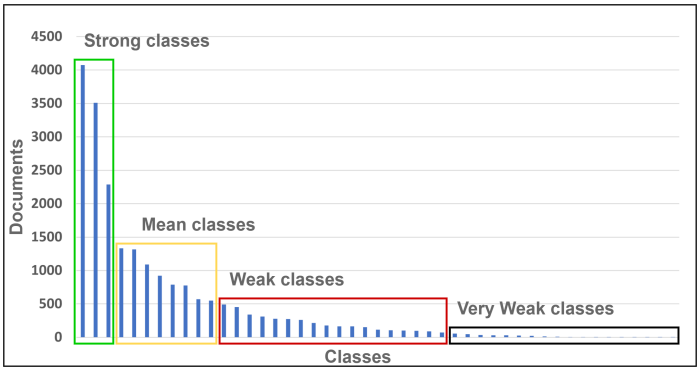

Fig. 1. Distribution of documents over class in a document stream sample.

These two properties combined with industrial constraints highly challenge the state of the art. The two best solutions for classification: expert system [3, 4] (mainly rule based system) and neural network [5] (and other machine learning methods using samples of results). For expert systems, the adaptation to the unknown is a big problem, any new class appearance introduces the addition of new rules and verification of older rules reliability. On this point, neural networks are better with only the need of some new samples, but could suffer from catastrophic forgetting [6]. An adaptation of neural networks to the imbalanced and incomplete context could be a good solution for our case.

We define in a previous work [7] the problems introduced by document stream context with the use of state of art the neural networks and deep learning models. This study shows that the adaptation of networks to imbalance spread of samples over classes is difficult and they cannot manage incompleteness, even reject unknown classes samples.

For this paper, we have done further research on the adaptation of neural networks to low sample conditions (see Fig. 2). It shows that our best neural networks are better than few-shot learning methods over 1000 samples per class and became clearly less efficient under 50. It shows equally that image convolutional networks require more samples than textual recurrent networks to keep their high performances. In this context, it seems difficult to rely on only one of these networks. None of them offer both the high precision required for strong classes and the ability to manage weaker classes.

With these further research and our previous work [7], we consider the three best state of the art network candidates for imbalanced documents classification: a biRNN [8] classifier using Bert [9] embedding, the VGG16 [10] network one of the best classifier on document image [11], ProtoNet [12] a few-shot learning methods that we have managed to apply to the document. With these three networks as benchmarks, we could see if our adaptation solutions for document stream classification are able to improve performances in imbalanced and realistic context and which one is the best. In the next sections we will present shortly these two adaptation solutions.

3. Proposed solutions

3.1. Multimodality, attention and adaptation

To create a neural network architecture adapted to document stream classification, we firstly try to adapt it with a combination of multiple state of the art solutions. It includes the use of a multimodal architecture [13] that is able to deal with visual and textual content, attention systems and an imbalanced adaptation.

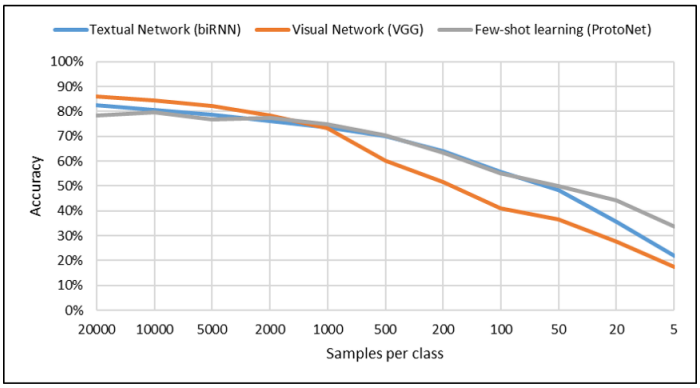

Fig. 2. Results with a decreasing number of training samples.

The proposed architecture is presented in the Fig. 3. It combines the best network evaluated for each modality: biRNN-Bert [8, 9] and VGG16 [10]. Each one is reinforced with an attention system to improve precision and accuracy, with one inspired of [14] for textual network and another inspired of [15] for image network. The combination is done in a meta-classifier of dense layers that combine as input the output of each network (provided by the last dense layer before classification). In addition, this solution uses an adaptive loss function to deal with the imbalance spread of classes in the training set.

The proposed loss function can be formally described with the following equation:

$$Loss_{CE}(i) = -Wt_j log(P(i)) \quad (1)$$

The main change from classical cross-entropy (corresponding to $-log(P(i))$ in Eq. 1) is the introduction of the Wt_j parameter. It is a weight assigned to each class which is the inverse percentage of class samples overall corpus samples, in other words the inverse frequency of class j. This architecture offers better results than the state of the art approaches for the imbalanced case, but it fails against incompleteness [16] and has low performances on very weak classes, with few-shot learning conditions.

3.2. Cascade of systems

In order to cope with few-shot learning and incompleteness issues, we have proposed a solution using a cascade of systems (published in [17]). It

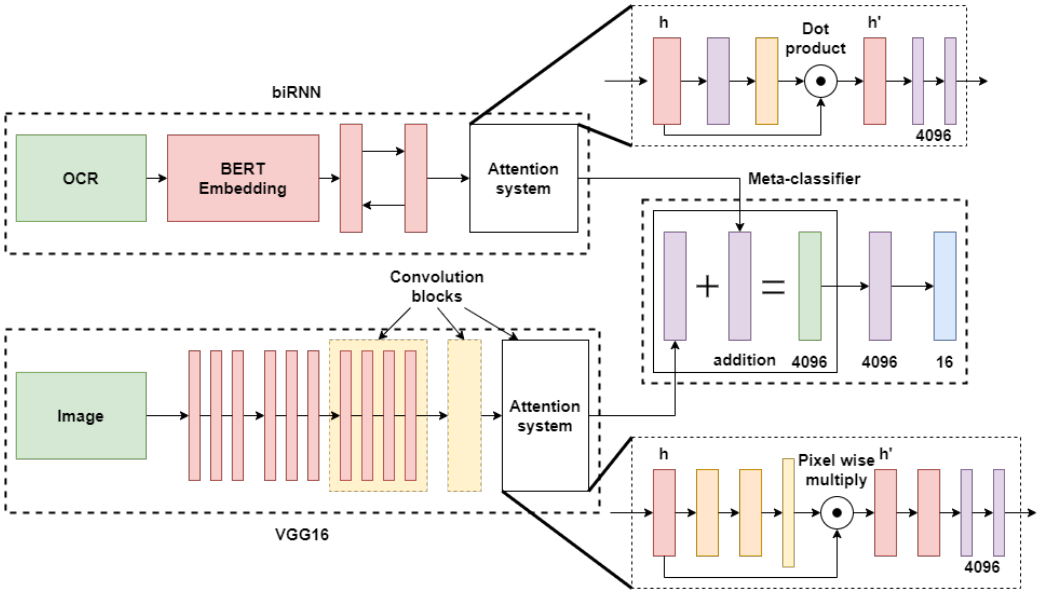

Fig. 3. Multimodal architecture with attention.

allows classifying the stream progressively with each stage specialized for a specific issue. It could be compared to a succession of increasingly thin sieves. It relies on a training corpus segmentation, with a selection function, at each stage to improve specialization (see Fig. 4).

The selection function uses a combination of four configurable criterions to assess the quality of the training on each sample/class. All samples considered as "too difficult" are gathered in a new dataset used to train the next stage, the others are removed because we consider them as manageable with the current stage.

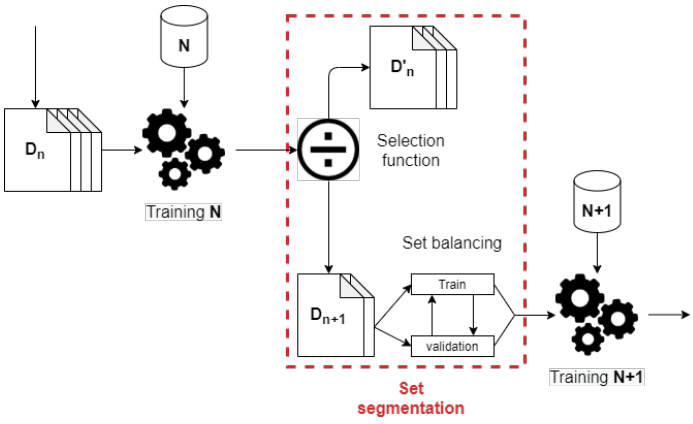

Fig. 4. Cascade of system training process.

The decision process is done stage after stage. When a stage has classified a document with a sufficiently high confidence rate, it is considered as classified and removed from the cascade. This process limits the computing time, because the majority of documents will be classified by the first stage.

In this cascade method, the order of stages is primordial. The main idea is to use classical deep networks as first stages (to deal with strong or mean classes, refer to Fig. 1) and few-shot learning or incremental methods as last stages (to manage weak and very weak classes, hard to train with deep learning methods). These two solutions are complementary, the multimodal architecture could serve as cascade's first stage.

4. Experiments

4.1. Protocol

The protocol used for experiments is based on the work done in [7] and [17]. The datasets used come from two different streams. First, an excerpt of documents issuing from the industrial stream of the Yooz company (YOOZDB), and the RVL-CDIP dataset [18]. YOOZDB is a private dataset containing confidential and personal data (which prevents us from making it public) and composed of 23 532 documents unequally distributed into 47 classes. Classes have between

1 and 3397 documents, and the distribution of documents in classes is available in Fig. 1. The dataset is divided into a training set of 15 491 documents (65.71%), a validation set of 2203 documents (9.34%) and a test set of 5883 (24,95%) documents. YOOZDB classes are mainly multiple variations of invoices, purchase orders, bank statements, tax notice, mails, cheques, identity card, contract and more. One of main problems with this dataset is its test set that was as imbalanced and incomplete as a document stream. Consequently, the overall result gathered on this dataset appears like it is balanced. Strong classes have a much greater impact than weak classes on the overall outcome, simply because there are more of them.

As there is no public documents stream dataset, we created two from the RVL-CDIP in order to simulate YOOZDB challenges. These are respectively the Imbalanced and the Realistic datasets and have been generated by a simulation protocol, summarized in Fig. 5.

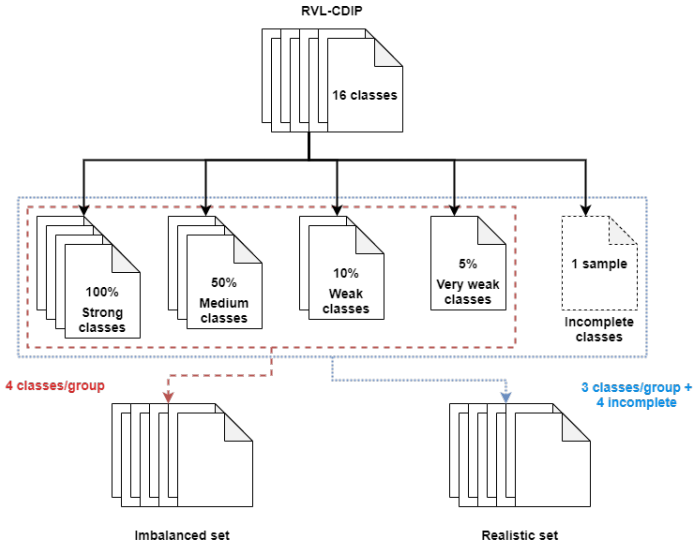

Fig. 5. Simulation protocol used to generate Imbalanced and Realistic datasets.

To assess methods, we use accuracy and precision (and recall when we study at class level). To gather accuracy and precision, we complete this selection with the $F_{0.5Measure}$, the calculation of which follows the equation below. This variation of the F_1 reinforces the importance of precision in the final score. This choice relies on the idea that precision is of greater interest for our industrial process (rejecting a document has less impact than predicting a wrong class).

$$(1 + 0.5^2)\frac{Precision \times Accuracy}{(0.5^2 \times Precision) + Accuracy} \quad (2)$$

All methods that do not have one are combined with a rejection system, which is a simple high threshold (around 0.99 on result within a range of [0:1]) applied on the confidence rate returned by the method for each document. It is used to maximize precision, as required for our context (refer to section 1).

The methods use in our experiments are biRNN [8], VGG16 [10], ProtoNet [12] (as the state of the art), our multimodal adapted architecture (Multi) [16] and our multiple cascade architecture (noted with system name of each stage separated by a symbol '+') [17].

4.2. Results

Table 1 introduces results on RVL-CDIP and YOOZDB. On YOOZDB, our solutions are slightly inferior to biRNN alone, especially for three stage cascades. It is because this dataset's strong classes rely mainly on text to be well classified, the contributions of image classifiers are just too low. Consequently, Multi remains inferior to textual approaches. On this dataset, the interest of three stage cascades is limited. They are close or inferior to two stage cascades because the number of documents is too low to properly train the third stage. Overall, the results indicate that our solutions may differentiate efficiently the YOOZDB classes. The best results here come from the biRNN.

On RVL-CDIP, the results are better for cascade and Multi. The three stages cascades are more interesting than two stages on a dataset of this size. As on YOOZDB, results show that the order of stages is primordial with cascade architecture (mainly the first stage modality). Multi show strong results but with a lower precision, that is an important backside. In this case, Multi seems to be the best.

Overall, the results show that our two methods return equivalent or better results than state of

Table 1. Results on original datasets.

Datasets	YoozDB				RVL-CDIP			
Methods / Measures	Accuracy	Precision	Reject R.	F0,5 M.	Accuracy	Precision	Reject R.	F0,5 M.
First stage modality : Image								
ProtoNet	63.56%	**97.93%**	35.10%	88.37%	63.68%	91.36%	30.30%	84.05%
VGG	84.70%	95.47%	11.28%	93.10%	75.14%	**94.41%**	20.41%	89.80%
VGG-ProtoNet	84.21%	95.61%	11.92%	93.09%	75.79%	94.04%	19.41%	89.72%
VGG-biRNN-ProtoNet	**89.62%**	94.64%	**5.31%**	**93.59%**	**83.22%**	92.06%	**9.60%**	**90.14%**
First stage modality : Text								
biRNN	93.57%	99.22%	5.69%	**98.04%**	77.95%	88.58%	12.01%	86.23%
biRNN-ProtoNet	**93.77%**	98.95%	5.23%	97.87%	80.72%	88.99%	**9.30%**	87.21%
biRNN-VGG-ProtoNet	93.75%	97.23%	**3.58%**	96.51%	**81.08%**	90.61%	10.52%	**88.53%**
First stage modality : Text & Image								
Multi	**95.53%**	**98.27%**	2.79%	**97.17%**	89.71%	**90.42%**	0.79%	**90.28%**
Multi-ProtoNet	95.51%	97.40%	**1.94%**	97.01%	**89.84%**	90.33%	**0.55%**	90.23%

Table 2. Results on datasets generated from RVL-CDIP.

Corpus	Imbalanced				Realistic			
Methods / Measures	Accuracy	Precision	Reject R.	F0,5 M.	Accuracy	Precision	Reject R.	F0,5 M.
First stage modality : Image								
ProtoNet	62.55%	**90.60%**	30.96%	83.14%	56.76%	**78.16%**	27.38%	**72.68%**
VGG	59.17%	88.90%	33.44%	80.79%	51.28%	77.97%	34.24%	70.62%
VGG-ProtoNet	70.19%	87.67%	19.94%	83.51%	58.89%	73.28%	19.64%	69.68%
VGG-biRNN-ProtoNet	**82.45%**	86.54%	**4.73%**	**85.69%**	**65.75%**	70.94%	**7.30%**	69.83%
First stage modality : Text								
biRNN	68.37%	79.73%	14.25%	77.17%	50.71%	67.56%	24.94%	63.35%
biRNN-ProtoNet	**74.90%**	79.00%	**5.19%**	78.14%	**58.05%**	64.30%	**9.72%**	62.94%
biRNN-VGG-ProtoNet	74.48%	**81.95%**	9.11%	**80.34%**	56.11%	**70.81%**	20.72%	**67.28%**
First stage modality : Text & Image								
Multi	82.99%	**86.92%**	4.53%	**86.11%**	59.63%	**76.18%**	21.73%	**72.17%**
Multi-ProtoNet	**83.59%**	86.33%	**3.18%**	85.77%	**63.03%**	73.92%	**14.73%**	71.45%

the art. They are challenging them-self between Multi and the VGG+biRNN+ProtoNet cascade.

Table 2 introduces results in simulated cases. For the imbalanced case, our solutions clearly overtake results from state of the art methods. The combination between Multi and the Prototypical network does not work well. The problem came from Multi that has a lack of precision and a tendency to respond confidently even when wrong. For cascades, the approach based on image as the first stage remains better. In this context, three stages cascades are clearly better than two stages cascades, but the Multi model retains the best results. In this case, Multi is very close to the VGG+biRNN+ProtoNet that are our two best solutions. They overtake VGG with +5% to +6% of F0.5 Measure and ProtoNet with +2% to +3% of F0.5 Measure.

However, our methods' best results remain lesser than ProtoNet in a realistic case. The addition of the incompleteness reduces the result more than expected for the multi and the cascades. The cascade including Multi architecture does not work at all because Multi returns a very high confidence rate in almost all situations (it has a bad precision). For other cascades, the loss of precision increases more in contrast to the gain

Table 3. Comparison class by class between networks alone and two stages cascades, with mono-modal network and ProtoNet, on Imbalanced RVL-CDIP (Recall value calculation does not include rejects). The percentage groups correspond to those of the simulation protocol (see Fig 5).

Methods	Difference between biRNN-ProtoNet and biRNN							
Groups	100%				50%			
Classes	**Advert**	**File F**	**Handwr**	**Sc Report**	**Budget**	**Email**	**Invoice**	**Resume**
Precision	1.89%	2.77%	**8.73%**	-1.13%	-1.58%	-0.29%	-1.35%	-0.04%
Recall	-0.35%	4.53%	**5.54%**	-1.94%	-2.20%	-0.22%	-1.70%	-0.75%
Reject Rate	**-11.60%**	**-32.48%**	**-33.28%**	-2.52%	-2.88%	-1.52%	-2.08%	-0.76%
Groups	10%				5%			
Classes	**Form**	**Letter**	**Presen**	**Questi**	**Memo**	**News A**	**Sc Public.**	**Speci**
Precision	-1.73%	-1.02%	0.34%	-1.65%	0.12%	**-6.03%**	**-6.63%**	-0.43%
Recall	-3.22%	-2.42%	-2.99%	-3.04%	-2.95%	3.99%	3.65%	-2.42%
Reject Rate	**-5.84%**	**-5.36%**	**-7.56%**	-4.28%	-3.60%	**-15.36%**	**-13.24%**	-2.68%

Methods	Difference between VGG-ProtoNet and VGG							
Groups	100%				50%			
Classes	**Advert**	**File F**	**Handwr**	**Sc Report**	**Budget**	**Email**	**Invoice**	**Resume**
Precision	0.06%	**-5.97%**	**-5.80%**	0.69%	-1.45%	-0.25%	-0.06%	-0.23%
Recall	-3.91%	-0.07%	-0.70%	**-7.20%**	-1.95%	-0.28%	-1.84%	-1.22%
Reject Rate	-4.60%	**-7.00%**	-2.12%	**-5.00%**	**-6.52%**	-1.72%	-3.20%	-3.44%
Groups	10%				5%			
Classes	**Form**	**Letter**	**Presen**	**Questi**	**Memo**	**News A**	**Sc Public.**	**Speci**
Precision	**-8.16%**	-1.98%	**-5.48%**	-2.02%	-1.27%	-4.05%	-3.51%	-1.10%
Recall	**14.16%**	-1.63%	**10.21%**	0.72%	-0.37%	3.35%	**8.56%**	0.71%
Reject Rate	**-18.40%**	**-17.76%**	**-22.88%**	**-20.56%**	**-13.00%**	**-29.16%**	**-43.52%**	**-17.16%**

in accuracy (although it remains very high). In a case where the importance of accuracy would be equivalent or higher than that of precision, the cascade model would perform better than any other model.

To conclude, cascade models show very interesting results in an imbalanced context where they allow a very strong increase in accuracy for a little loss of precision. Multi do the same but slightly better. In the realistic case, the best solution depends on what measure is your references. If you need accuracy, it will be VGG-biRNN-ProtoNet and if you prefer precision, it will be Protonet. Multi is more balanced and win with F1-measure.

4.3. Cascade improvement analysis

To make a deeper analysis of cascade models, we made a comparison between two stages cascades and networks alone with one for each modality on the Imbalanced dataset. Table 3 gathers all differences between the two stage model and the original, showing what is improved or reduced by the integration to the cascade.

For the comparison biRNN-ProtoNet/biRNN, the results are a bit surprising. The impact of multimodality is stronger than we expected. The Prototypical Network train was more focused (through targeted cascade training) on reinforcement of complex classes for text classification than under-represented classes (mainly on the classes "File Folders" and "Handwritten", as well as "Advertisement" and "Presentation" to a lesser extent). This effect makes the gains on the 10% and 5% groups exist but lesser than expected.

For VGG16-ProtoNet/VGG16, the results are more consistent with those initially expected. This shows that previously the disturbances were indeed due to the multimodal impact. The gains on the 10% classes are a large increase in re-

call against a much smaller but still significant amount of precision, with a reduction in the rejection rate around 20%. For the classes in the 5% group it is even more striking, as the losses are lower in precision for an even higher reduction in rejection rate (over 40% for the "Scientific Publication" class). Overall, these results show that cascade models allow improvement on underrepresented classes and cases of modality complementarity.

5. conclusion

To conclude, in this particularly difficult context we need to adapt state of art methods and we have proposed multiple options to do so (that we have already published): a multimodal adapted architecture with attention systems and a combination of systems like a cascade to process document stream progressively. These two options are efficient in ideal and imbalanced contexts. But, the results on realistic context remain less than required so we consider improving our two methods. The first issue to resolve is the combination between multimodal architecture and the cascade (address the precision issue). The cascade seems to be the solution with the most possible improvement, including the use of other methods as stages and the improvement of the selection function with better or more criterions.

References

[1] D. Schuster, K. Muthmann, D. Esser, A. Schill, M. Berger, C. Weidling, K. Aliyev, A. Hofmeier, Intellix–end-user trained information extraction for document archiving, in: 2013 12th ICDAR, IEEE, 2013, pp. 101–105.

[2] V. P. d'Andecy, A. Joseph, J.-M. Ogier, Indus: Incremental document understanding system focus on document classification, in: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), IEEE, 2018, pp. 239–244.

[3] H. Tan, A brief history and technical review of the expert system research, in: IOP Conference Series: Materials Science and Engineering, Vol. 242, IOP Publishing, 2017, p. 012111.

[4] C. Grosan, A. Abraham, Rule-based expert systems, in: Intelligent systems, Springer, 2011, pp. 149–185.

[5] T. M. Mitchell, et al., Machine learning, Vol. 1, McGraw-hill New York, 2007.

[6] M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: Psychology of learning and motivation, Vol. 24, Elsevier, 1989, pp. 109–165.

[7] J. Voerman, A. Joseph, M. Coustaty, V. P. d'Andecy, J.-M. Ogier, Evaluation of neural network classification systems on document stream, in: International Workshop on Document Analysis Systems, Springer, 2020, pp. 262–276.

[8] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing 45 (11) (1997) 2673–2681.

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

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

[11] M. Z. Afzal, A. Kölsch, S. Ahmed, M. Liwicki, Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification, in: 2017 14th IAPR ICDAR, Vol. 1, IEEE, 2017, pp. 883–888.

[12] J. Snell, K. Swersky, R. S. Zemel, Prototypical networks for few-shot learning, arXiv preprint arXiv:1703.05175 (2017).

[13] S. Bakkali, Z. Ming, M. Coustaty, M. Rusinol, Visual and textual deep feature fusion for document image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 562–563.

[14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

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

[16] I. S. Mahamoud, J. Voerman, M. Coustaty, A. Joseph, V. P. d'Andecy, J.-M. Ogier, Multimodal attention-based learning for imbalanced corporate documents classification, in: ICDAR, Springer, 2021, pp. 223–237.

[17] J. Voerman, I. S. Mahamoud, A. Joseph, M. Coustaty, V. P. d'Andecy, J.-M. Ogier, Toward an incremental classification process of document stream using a cascade of systems, in: ICDAR, Springer, 2021, pp. 240–254.

[18] A. W. Harley, A. Ufkes, K. G. Derpanis, Evaluation of deep convolutional nets for document image classification and retrieval, in: 2015 13th ICDAR, IEEE, 2015, pp. 991–995.