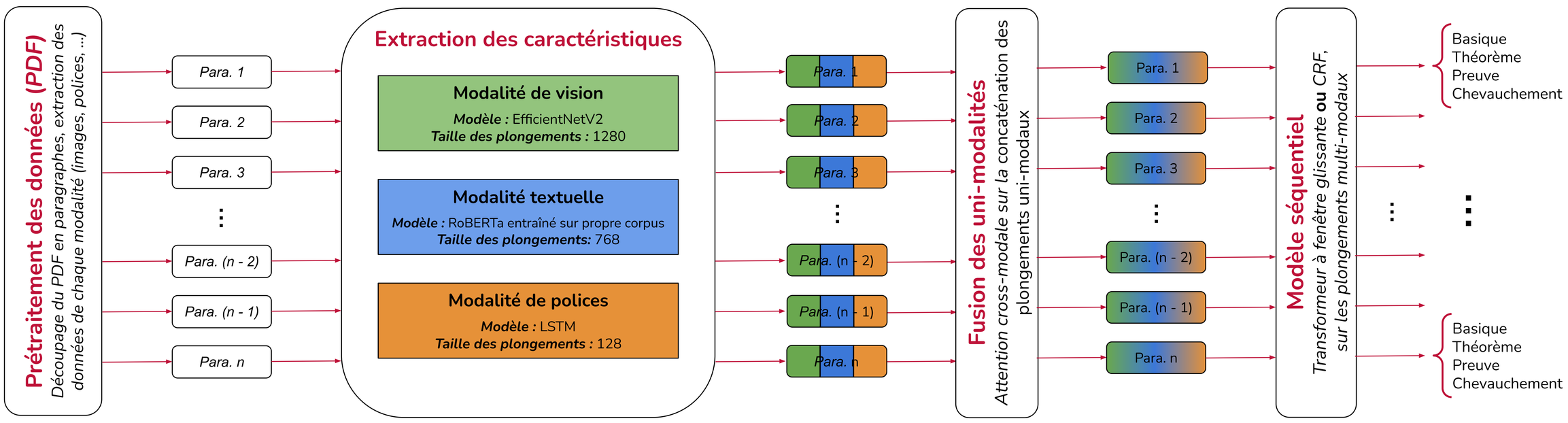
Apprentissage multimodal modulaire pour l'extraction de théorèmes et de preuves dans des documents scientifiques longs

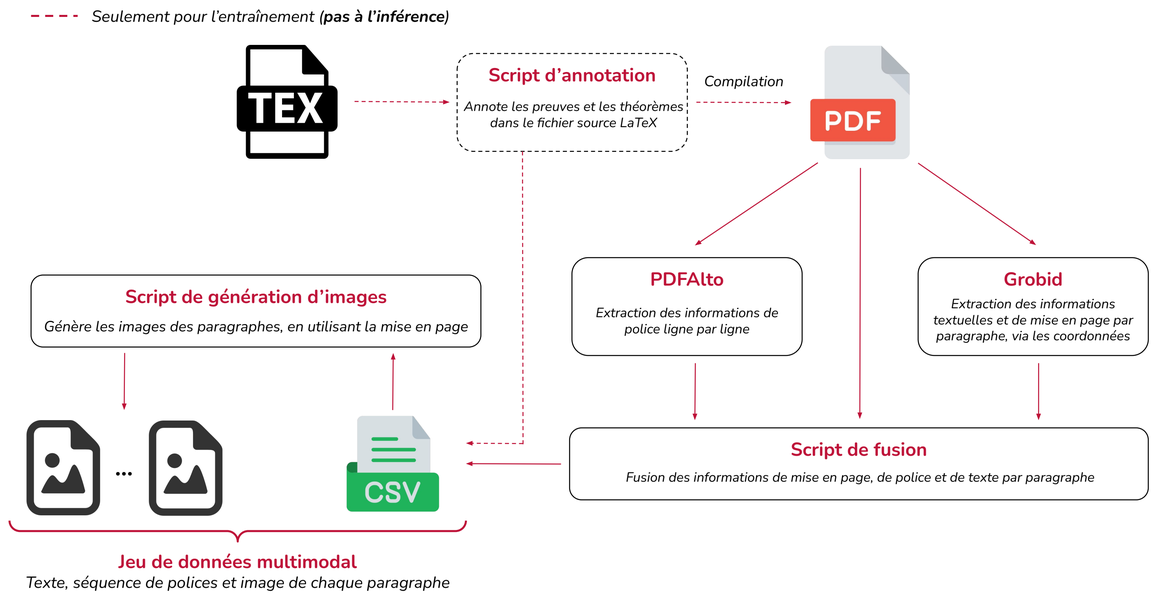
Shrey Mishra **Antoine Gauquier** Pierre Senellart

ENS PSL | cnrs | Inria | institut universitaire de France



Pré-traitement des données

Chaque PDF est pré-traité afin d'en extraire différentes modalités. Le code source LaTeX est uniquement nécessaire dans les phases d'entraînement afin de générer automatiquement des données annotées. Seul le PDF est requis à l'inférence.



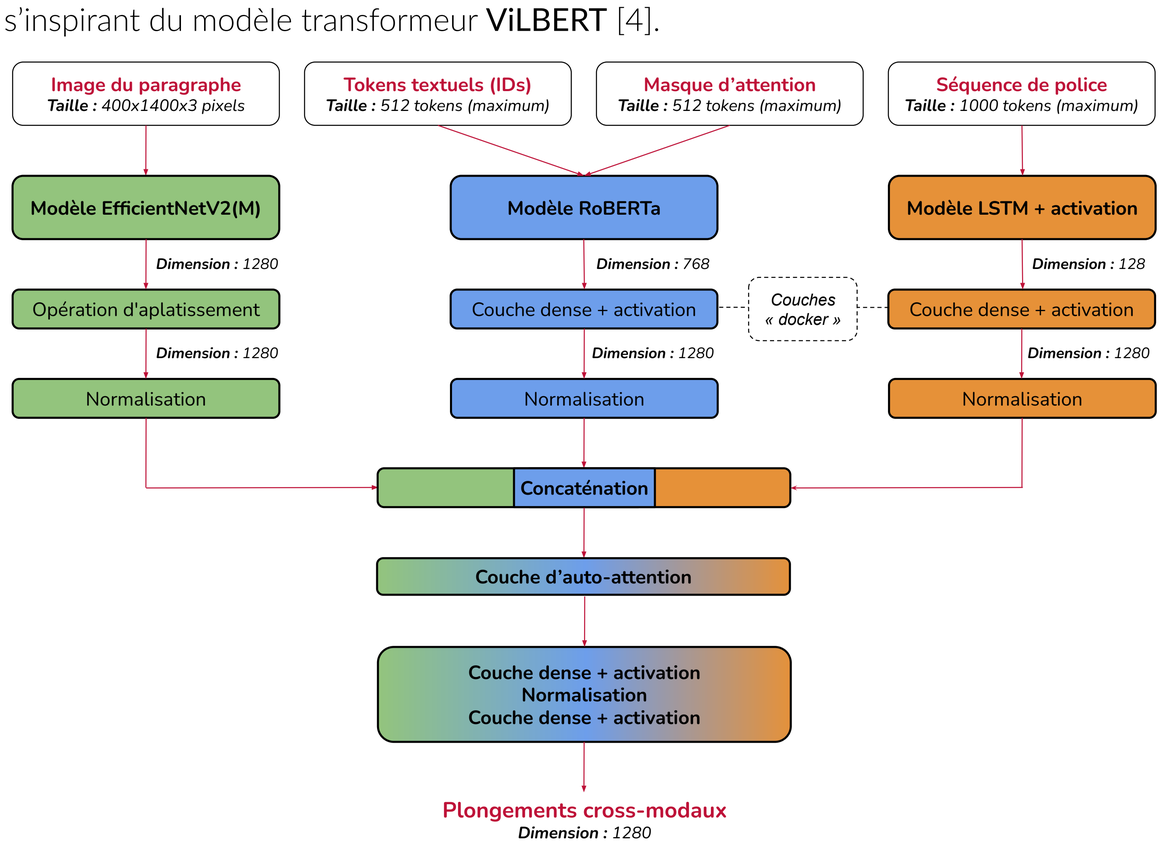
Modèles unimodaux

Un modèle par modalité est entraîné :

- **Texte.** Un modèle **RoBERTa** [3] est entraîné de zéro sur un nouveau corpus, constitué du texte des paragraphes
- **Vision.** Ré-entraînement d'**EfficientNet V2** [6] sur le rendu *bitmap* des PDF des paragraphes
- **Séquence de polices.** Un modèle **LSTM** [1] est entraîné sur les séquences de polices de caractère de chaque paragraphe (un token de police est donné pour chaque caractère du paragraphe)

Modèle d'attention cross-modale

Le modèle d'attention cross-modale permet de fusionner les trois unimodalités, en s'inspirant du modèle transformeur **ViLBERT** [4].



Modèles séquentiels

La prise en compte de l'*ordre* dans lequel apparaissent les paragraphes aide les modèles unimodaux ou multimodal à assigner la bonne classe à chaque paragraphe.

Deux architectures sont testées : un **Conditional Random Field** (CRF) [2] et un modèle transformeur à **fenêtre glissante** (TFG). Chacun d'entre eux intègre, en plus des plongements multimodaux, les informations de séquence suivantes :

- Le **numéro de la page** dans laquelle apparaît le paragraphe
- Les **distances horizontales et verticales** du paragraphe courant avec son prédécesseur, en utilisant les *bounding boxes de chacun*
- Un booléen décrivant si le pragraphe courant est **sur la même page que son prédécesseur**, afin de donner un sens aux distances mentionnées ci-dessus

Jeu de données

Nos jeux de données sont construits à partir de tous les articles présents sur arXiv jusqu'en mai 2020 (environ 1,7 millions). Plusieurs filtres sont appliqués successivement sur cet ensemble pour obtenir le jeu d'entraînement :

- On ne conserve que les **articles en anglais**.
- On ne conserve que les articles qui contiennent au moins un environnement de preuve ou de théorème dans leur fichier source LaTeX afin de pouvoir précisément les identifier dans leur rendu PDF, **et qui compilent correctement**.

Ceci réduit l'ensemble à un peu moins de 500k articles. Enfin, on ne conserve que les articles pour lesquels tout le processus de pré-traitement s'est déroulé sans erreur. Il en résulte un ensemble de 197k articles.

Le jeu de validation de notre modèle est constitué d'un peu plus de 3,6k documents PDFs regroupant plus de 500k paragraphes.

Résultats

Comparaison des performances (précision et F_1 moyen sur les trois classes) des modèles unimodaux et multimodaux, avec ou sans approche séquentielle ; 1 lot = 1 000 documents

Modalité	Modèle	Approche séq.	#Lots	#Paramètres (M)	Précision (%)	$\mathbf{F}_1$ moyen (%)
Basique	Prédit toujours *Basique*	–	–	–	59,41	24,85
Top-k premier mot	Utilise le premier mot	–	–	–	52,84	44,20
Ligne à ligne	BERT (affiné) [5]	–	–	110	57,31	55,71
Polices	LSTM 128 cellules	-	11	2	64,93	45,48
		CRF	11+8	2	71,50	64,51
		TFG	11+8	2	76,22	71,77
Vision	EfficientNetV2m	-	9	53	69,44	60,33
		CRF	9+8	53	74,63	70,82
		TFG	9+8	65	79,59	77,66
Texte	RoBERTa entraîné sur propre corpus	-	20	124	76,45	72,33
		CRF	20+8	124	83,10	80,99
		TFG	20+8	129	87,50	86,67
Multimodal	Attention cross-modale	-	2	185	78,50	75,37
		CRF	2+8	185	84,39	82,91
		TFG	2+8	198	**87.81**	**87.18**

Références

[1] Sepp Hochreiter et Jürgen Schmidhuber : Long short-term memory. *Neural computation*, 9(8), 1997.

[2] John D. Lafferty, Andrew McCallum et Fernando C. N. Pereira : Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In ICML*, 2001.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer et Veselin Stoyanov : RoBERTa : A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.

[4] Jiasen Lu, Dhruv Batra, Devi Parikh et Stefan Lee : ViLBERT : Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[5] Shrey Mishra, Lucas Pluvinage et Pierre Senellart : Towards extraction of theorems and proofs in scholarly articles. *In Proc. DocEng*, Limerick, Irlande, août 2021.

[6] Mingxing Tan et Quoc Le : EfficientNetv2 : Smaller models and faster training. *In ICML*, 2021.