BERT spécialisés

Sous-titre

Andrei Barbu et Théo Molfessis

25 mars 2025

Plan de la présentation

Objectif et Motivation

Bert spécialisé pour des tâches précises.

Pourquoi spécialiser ?

- Permettre une distinction plus fine entre différents concepts et sujets traités.
- Réduire l'empreinte énergétique en utilisant des modèles plus compacts, comme BERT tiny, spécialement entraînés sur des données spécialisées.

Exemples de spécialisation :

- LegalBert[2], Juribert [4], BioBert[5], FinBert [6].
- Tâches : NER, Classification, Question Answering ...

Données utilisées

Dataset ECHR : données juridiques en anglais
11 500 cas extraits de la base publique de la Cour Européenne des Droits de l'Homme.

Tâche :
Tâche binaire : une violation des droits est présente ou non pour chaque cas.

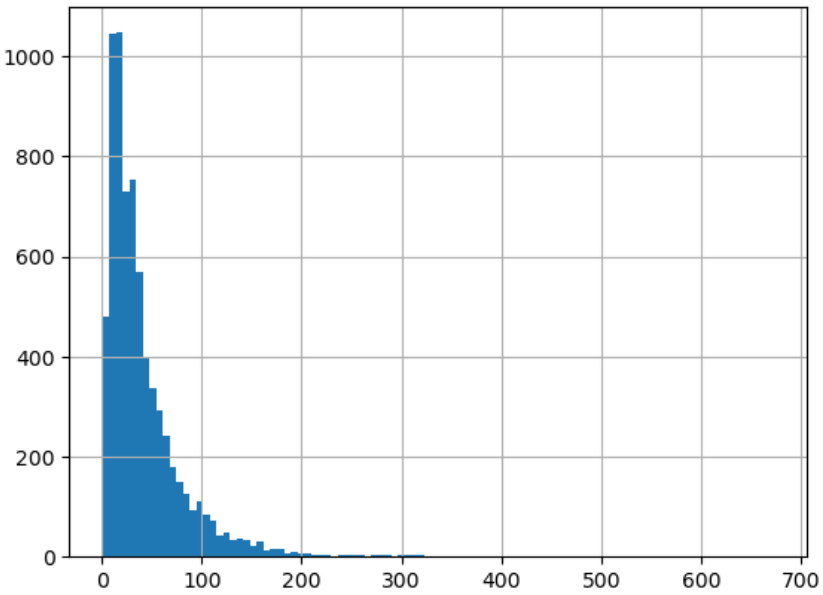
Structure des données :
Liste de faits extraits du dossier, ainsi que l'indication des articles de la Convention éventuellement violés.

Modèles considérés :

- **BERT tiny** : Un modèle compact adapté aux ressources limitées.
- **LegalBert Small** : Un modèle entraîné from scratch sur EURLEX

Approches de fine tuning : full, freeze et LoRA.

Analyse des données



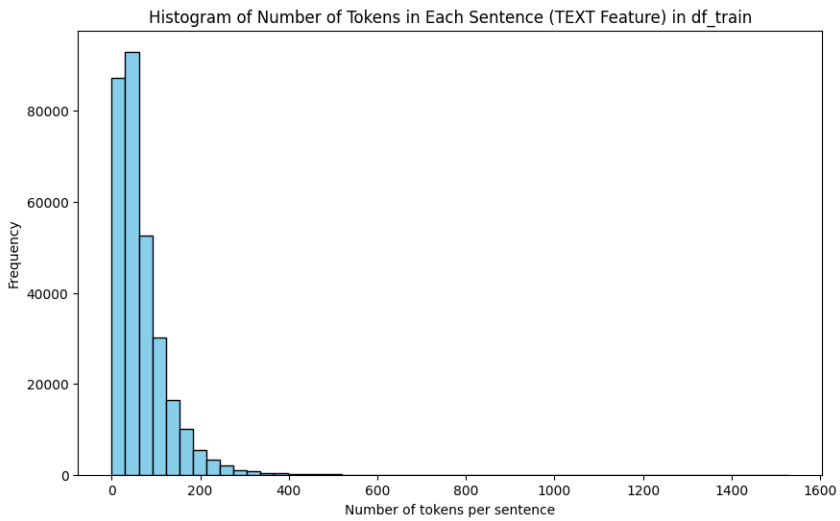


Figure – Histograms of the number of sentences per row and of the number of tokens per sentence

We therefore chose to pad/truncate at $N_s = 64$ sentences and $N_{tok} = 258$ tokens.

Approches de fine tuning

Rajout de couches supplémentaires pour la *downstream task*

1. **Approche MLP :**
   - BERT embeddings des phrases (taille h_{bert}) concaténés.
   - Un MLP est utilisé pour classifier le document.
   - Dropout, Lin($N_s \times h_{bert}$), *tanh*, Lin(2), softmax, selon [3].
2. **Approche par attention hiérarchique :**
   - BERT embedding de chaque phrase.
   - Mécanisme d'attention hierachique.
   - Q constant taille h_{bert}, K, V de taille h_{bert}^2 , Dropout , Lin(h_{bert}), *tanh*, Lin(2), softmax.

Ces deux approches permettent d'exploiter différemment les informations contenues dans le champ `TEXT`.

Stratégies de Fine Tuning

1. **Freeze (gel) des poids BERT :**
   - Seules les couches ajoutées en tête sont entraînées.
   - Avantage : entraînement rapide et moins de sur-apprentissage
   - Inconvénient : jugé moins efficace par SciBert[1].
2. **Fine tuning complet :**
   - Mise à jour de l'ensemble des paramètres du modèle.
   - Adaptation plus fine mais nécessite plus de données et un risque de sur-apprentissage plus élevé.
3. **Utilisation de LoRA :**
   - Au lieu de mettre à jour tous les poids linéaires ($\approx 400k$), on ajoute des modules d'adaptation (de faible rang : $8 \rightarrow 6.5\%$).
   - Petit nombre de paramètres, bonnes performances.
   - Version LoRA en encapsulant les couches linéaires de BERT.

Performances sur la tâche de classification

Modèle et Approche	Accuracy	Observations sur la loss
Tiny BERT		
Full (MLP)	0.881	Forte diminution entre l'époque 2 et 12.
Full (Attention)	**0.884**	Forte diminution entre l'époque 9 et 20.
Freeze (MLP)	0.827	Forte diminution entre l'époque 2 et 10.
Freeze (Attention)	–	Erreur, pas de CV ; loss stagne sur 12 époques.
LoRA rank 8 (MLP)	0.852	Forte diminution entre l'époque 2 et 10.
LoRA rank 8 (Attention)	0.834	Forte diminution entre l'époque 9 et 15.
Legal BERT Small (nombre de phrases et tokens divisés par 2)		
Freeze (MLP)	0.879	Diminution de 0.68 à 0.48 entre l'époque 0 et 4, puis diminution lente.
Full (MLP)	0.885	Diminution entre l'époque 0 et 6.
Full (Attention)	**0.889**	Diminution entre l'époque 2 et 6.
Résultats de référence LegalBert [2]		
BERT (base) - full, attention	0.826	
Legal BERT Small - full, attention	0.826	

Conclusion et Perspectives

Conclusion :

- Version Freeze MLP de Bert Tiny a le même écart de performance que dans [1]
- Freeze un model pré entraîné sur du droit est largement satisfaisant.
- Bert Tiny avec full fine tuning attention démontre qu'un petit modèle peut rapidement se spécialiser.

Références I

[1] Iz Beltagy, Kyle Lo et Arman Cohan. « SciBERT : A Pretrained Language Model for Scientific Text ». In : *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Sous la dir. de Kentaro Inui et al. Hong Kong, China : Association for Computational Linguistics, nov. 2019, p. 3615-3620. doi : 10.18653/v1/D19-1371. url : https://aclanthology.org/D19-1371/.

[2] Ilias Chalkidis et al. « LEGAL-BERT : The Muppets straight out of Law School ». In : *Findings of the Association for Computational Linguistics : EMNLP 2020*. Sous la dir. de Trevor Cohn, Yulan He et Yang Liu. Online : Association for Computational Linguistics, nov. 2020, p. 2898-2904. doi : 10.18653/v1/2020.findings-emnlp.261. url : https://aclanthology.org/2020.findings-emnlp.261/.

Références II

[3] Jacob Devlin et al. « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*. Sous la dir. de Jill Burstein, Christy Doran et Thamar Solorio. Minneapolis, Minnesota : Association for Computational Linguistics, juin 2019, p. 4171-4186. doi : 10.18653/v1/N19-1423. url : https://aclanthology.org/N19-1423/.

Références III

[4] Stella Douka et al. « JuriBERT : A Masked-Language Model Adaptation for French Legal Text ». In : *Proceedings of the Natural Legal Language Processing Workshop 2021*. Sous la dir. de Nikolaos Aletras et al. Punta Cana, Dominican Republic : Association for Computational Linguistics, nov. 2021, p. 95-101. doi : 10.18653/v1/2021.nllp-1.9. url : https://aclanthology.org/2021.nllp-1.9/.

[5] Jinhyuk Lee et al. « BioBERT : a pre-trained biomedical language representation model for biomedical text mining ». In : *Bioinformatics* 36.4 (sept. 2019), p. 1234-1240.

[6] Yi Yang, Mark UY et Allen Huang. *FinBERT : A Pretrained Language Model for Financial Communications*. Juin 2020. doi : 10.48550/arXiv.2006.08097.