CamemBERT: a Tasty French Language Model

Présenté par :
Emir TAS, Shivamshan SIVANESAN, Kishanthan KINGSTON

Encadré par :
Prof. Nicolas OBIN, Prof. Bruno GAS, Prof. Alice COHEN-HADRIA

État de l’art :

CamemBERT: a Tasty French Language Model [1]:

Dans l’article, les auteurs ont utilisé RoBERTa comme architecture et Oscar corpus (partie français seulement) pour l’entraînement du modèle. Ils ont utilisé SentencePiece pour segmenter les mots.

Évaluation: Part-Of-Speech tagging, dependency parsing, Named Entity Recognition, Natural Language Inference.

RoBERTa: A Robustly Optimized BERT Pretraining Approach [2]:

RoBERTa est une architecture qui est basée sur l’architecture BERT (Bidirectionnel Encoder Representations from Transformers), qui à son tour est une variante des transformers.

Attention Is All You Need [3]:

Présentation des transformers basé sur le mécanisme de l’attention.

Architecture Transformers [3] :

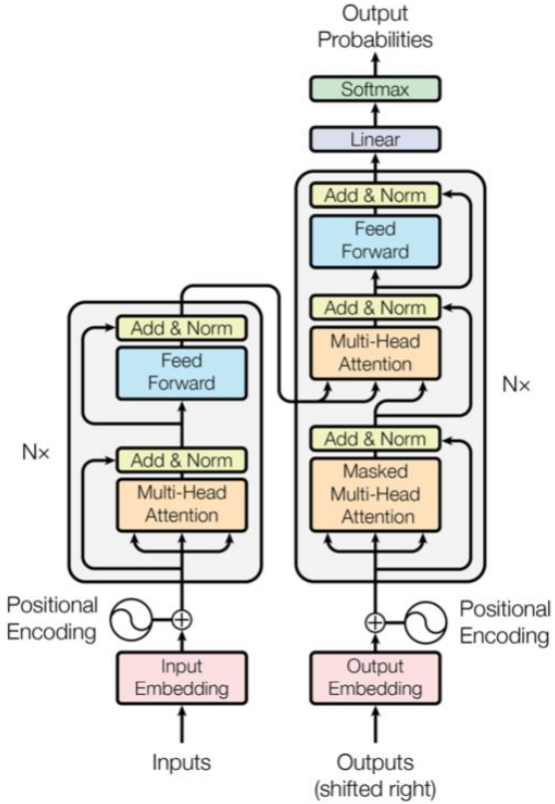


Figure 1: The Transformer - model architecture.

Hidden size : 768

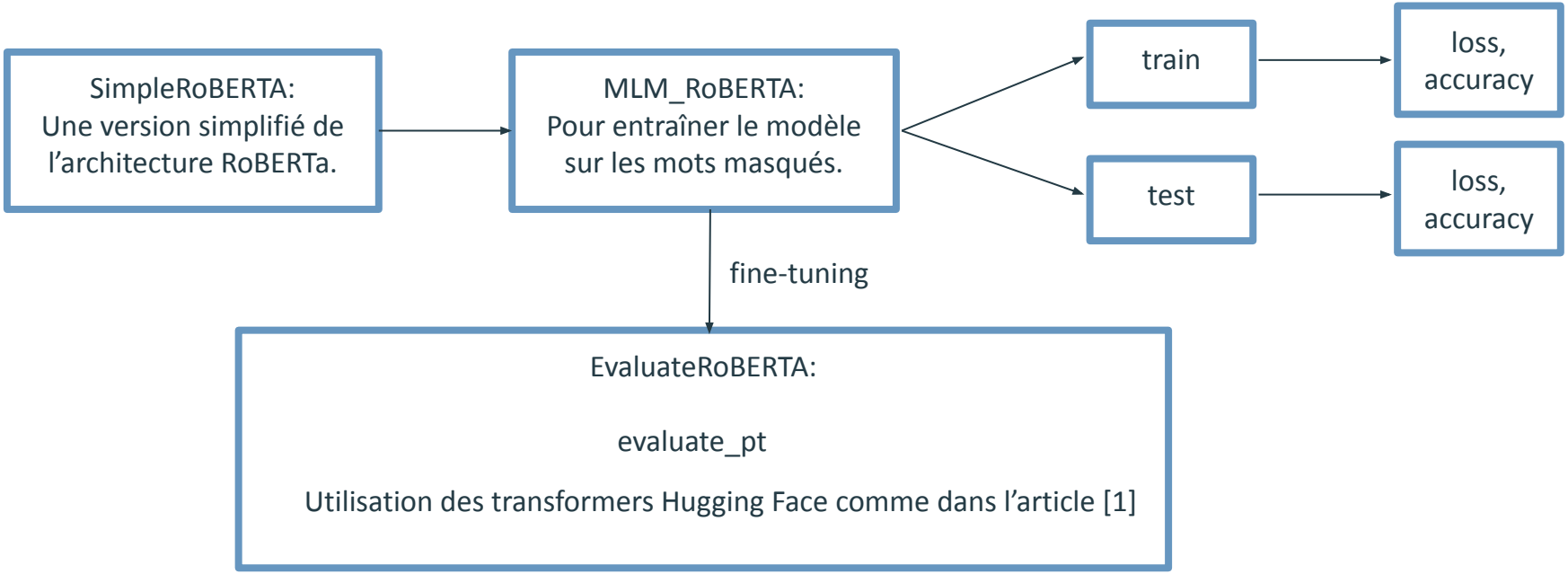
Attention layer : 12

Multi-head attention : 12

Input size : 512

Parameters : 110 M

Notre modèle :



RoBERTa MLM pré-entraîné

Dataset
GSD French

Prétraitement

Auto-Tokenizer
Roberta

Trainer
“roberta-base”

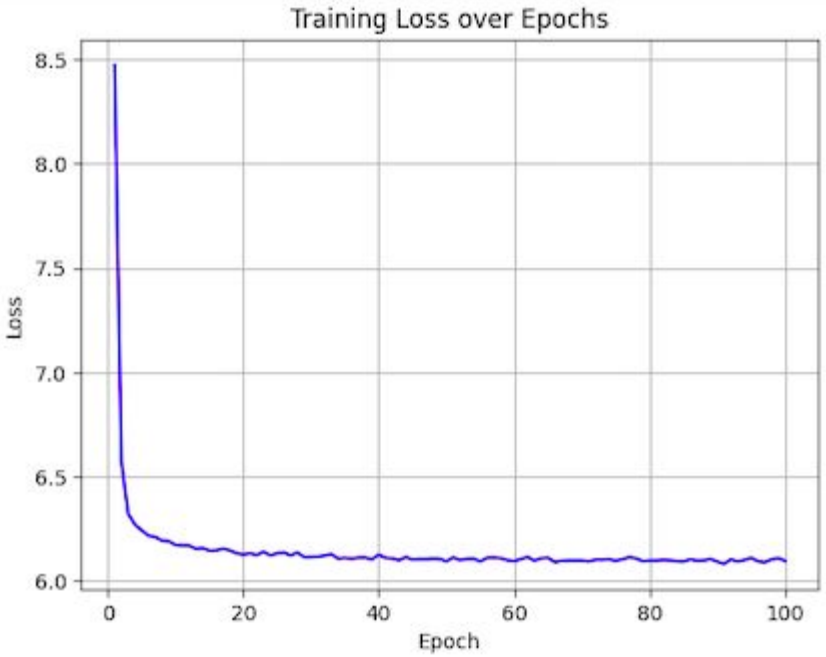
```
fill_mask("Je prépare un <mask> au chocolat.")
```

```
[{'score': 0.6465614438056946,
  'token': 17927,
  'token_str': ' dessert',
  'sequence': 'Je prépare un dessert au chocolat.'},
 {'score': 0.06801274418830872,
  'token': 2391,
  'token_str': ' restaurant',
  'sequence': 'Je prépare un restaurant au chocolat.'},
 {'score': 0.058270663022994995,
  'token': 5765,
  'token_str': ' menu',
  'sequence': 'Je prépare un menu au chocolat.'},
 {'score': 0.02214585803449154,
  'token': 4206,
  'token_str': ' excellent',
  'sequence': 'Je prépare un excellent au chocolat.'},
 {'score': 0.017763499170541763,
  'token': 11658,
  'token_str': ' satisfaction',
  'sequence': 'Je prépare un satisfaction au chocolat.'}]
```

Affichage des 5 tokens les plus probables :

- **“dessert” → le plus probable et le plus logique**
- **“restaurant”**
- **“menu”**
- **“excellent”**
- **“satisfaction”**

Résultats MLM_RoBERTa:



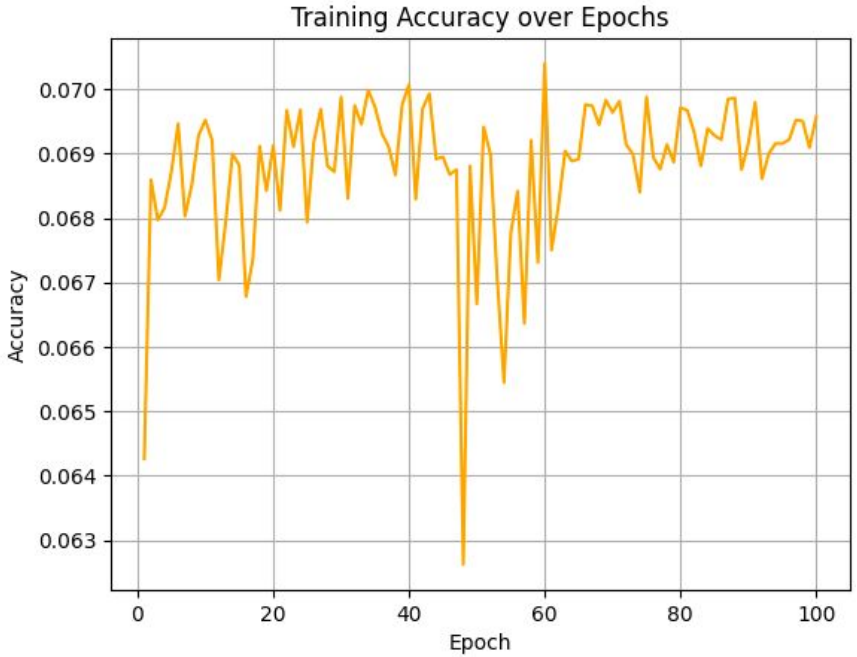


Figure : Graphique représentant la loss et l'accyracy du modèle que nous avons créé MLM_RoBERTa

Résultats RoBERTa pré-entraîné - POS-tagging:

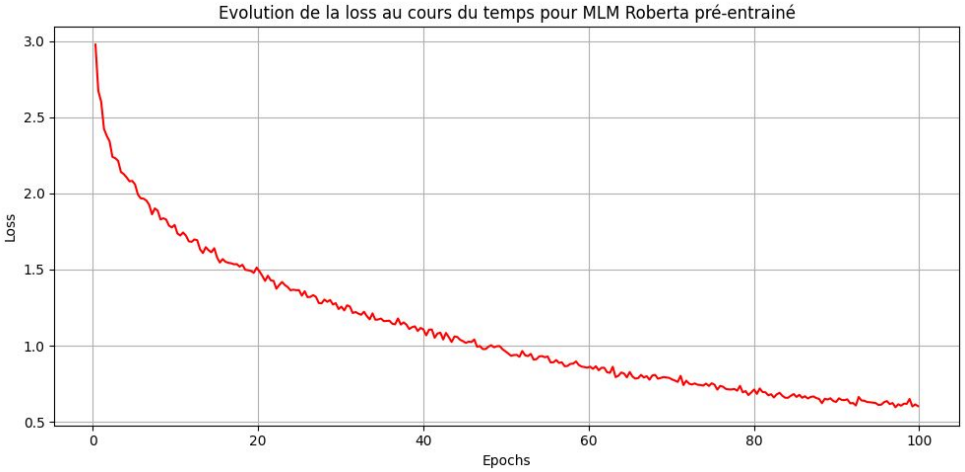


Figure : Graphique représentant la loss du modèle pré-entraîné

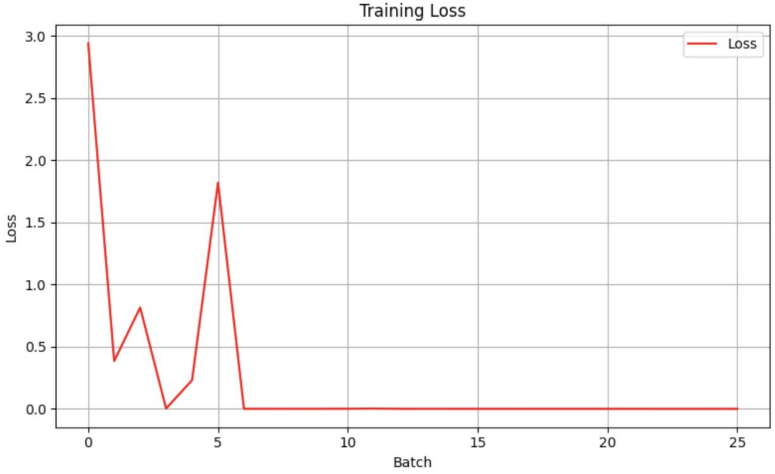


Figure : Graphique représentant la loss de la tâche POS-tagging

Conclusions :

- **Implémentation de l’architecture RoBERTa depuis zéro:**
  - Mise en place réussie de la classe MLM_RoBERTa avec l'utilisation de l’architecture SimpleRoBERTa
- **Entraînement sur MLM:**
  - Utilisation de la tâche MLM pour l'entraînement comme dans l’article CamemBERT [1]
  - Cependant, les résultats montrent des problèmes de convergence.
- **Loss très élevé et accuracy très faible:**
  - Les résultats d'entraînement révèlent une perte élevée et une exactitude très faible.
  - Possibles raisons : gestion incorrecte des masques, hyperparamètres inappropriés, complexité du modèle.
- **Over-fitting lorsqu’on affine le modèle pré-entraîné pour la tâche POS-tagging**
- **À faire:**
  - Affiner notre modèle MLM_RoBERTa en utilisant des données de pos-tagging pour l'évaluation finale.

Bibliographie:

[1] Martin, Louis, et al. « CamemBERT: A Tasty French Language Model ». Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, p. 7203‑19. DOI.org (Crossref), https://doi.org/10.18653/v1/2020.acl-main.645.

[2] Y. Liu et al., « RoBERTa: A Robustly Optimized BERT Pretraining Approach ». arXiv, 26 juillet 2019. Consulté le: 22 novembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/1907.11692.

[3] A. Vaswani et al., « Attention Is All You Need ». arXiv, 1 août 2023. Consulté le: 22 novembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/1706.03762.