



DÉVELOPPEZ UNE PREUVE DE CONCEPT

OPENCCLASSROOMS - INGÉNIEUR MACHINE LEARNING

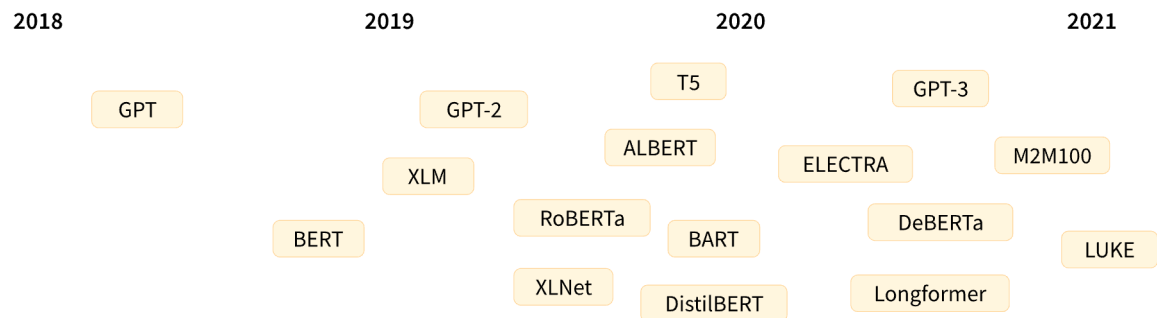
THIBAUD GROSJEAN - MAI 2022

OBJECTIFS

- Implémenter une technique de pointe
- Définir le plan de travail prévisionnel
- Mener le travail de recherche
- Réaliser un état de l'Art
- Comparer les performances à un modèle baseline

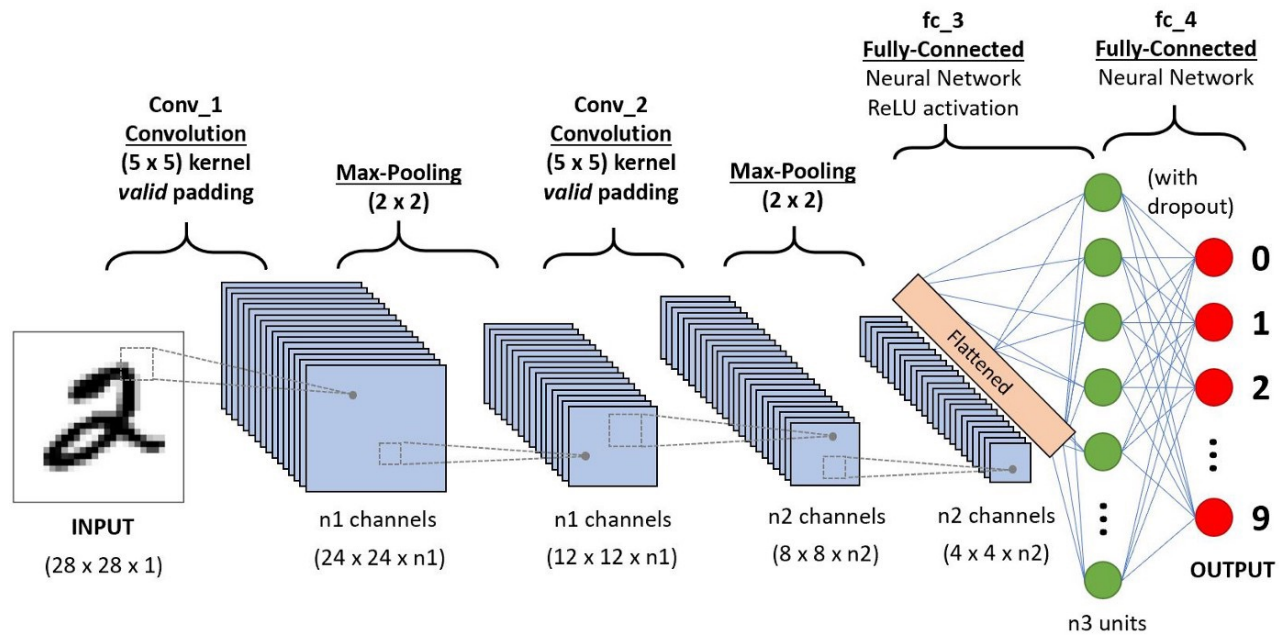
BIBLIOGRAPHIE

- Ashish Vaswani et al, Attention Is All You Need, 2017
- Kevin Clark et al., ELECTRA: Pre-Training Text Encoders As Discriminators Rather Than Generators, 2020
- Pengcheng He et al., DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021
- Francesco Barbieri et al., TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification, 2020
- Daniel Loureiro et al., TimeLMs: Diachronic Language Models from Twitter, 2022



TRANSFORMEURS

- Transductif
- Parallelisable
- Auto-Attention
- Auto-Supervisé
- Compréhension statistique



TRANSFORMEURS

- CNN
- Séquentialité

Focus

The → The big red dog
big → The big red dog
red → The big red dog
dog → The big red dog

Attention Vectors

$[0.71 \quad 0.04 \quad 0.07 \quad 0.18]^T$

$[0.01 \quad 0.84 \quad 0.02 \quad 0.13]^T$

$[0.09 \quad 0.05 \quad 0.62 \quad 0.24]^T$

$[0.03 \quad 0.03 \quad 0.03 \quad 0.91]^T$

ATTENTION

- Qu'est-ce que l'attention ?
- Attention par produit scalaire
- Optimisation
- Encodeur
- Encodeur-Décodeur

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

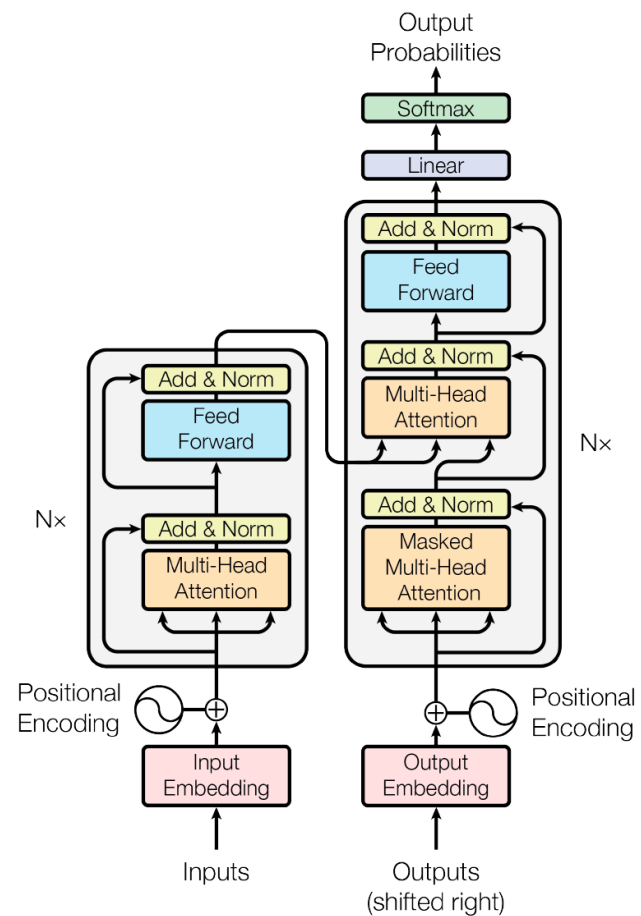
ATTENTION

- Décodeur



ATTENTION

- Attention à têtes multiples
- Parallélisation
- Attention Démêlée



ARCHITECTURE

- 6+6 layers d'attention
- 768 neurones de préclassification
- Drop-out (0.3)
- Neurones de classification

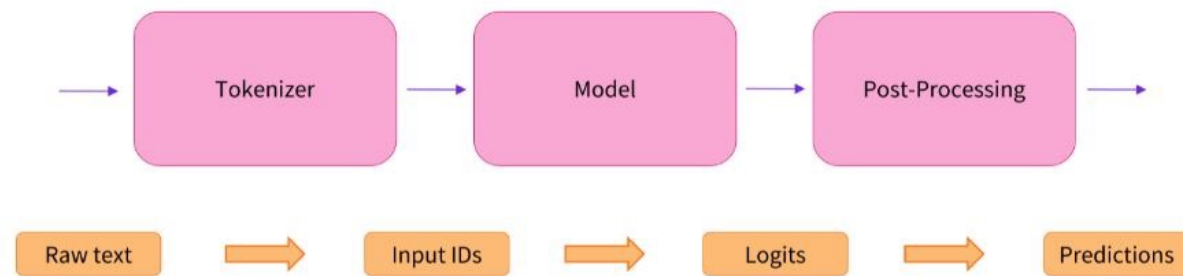
SÉQUENCES

- Vecteurs

- *Input ids* (tokens)
- *Input types*
- *Attention masks*
- *Labels* (cibles)

- Longueur de séquence: 100

Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[\text{CLS}]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[\text{SEP}]}$	E_{he}	E_{likes}	E_{play}	$E_{\text{##ing}}$	$E_{[\text{SEP}]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}



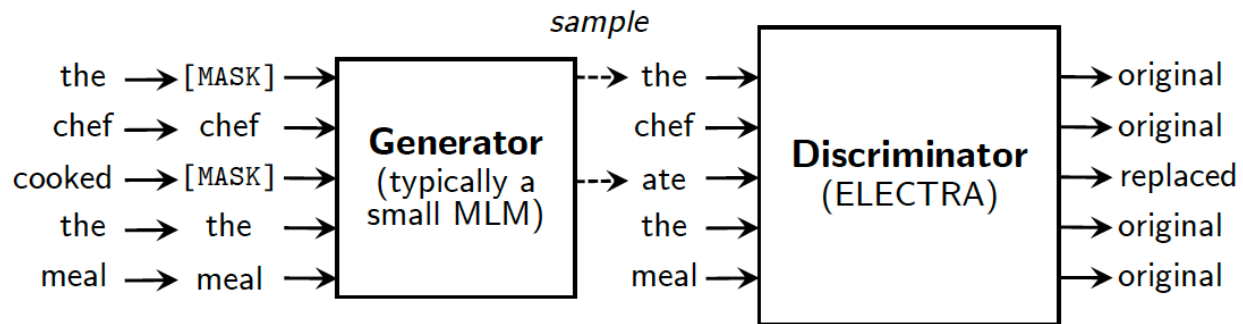
PIPELINE

- Simple en apparence...
- Mais complexe
- Le *Tokenizer* permet d'inverser les *logits* du modèle

Model	Wiki+Book 16GB	OpenWebText 38GB	Stories 31GB	CC-News 76GB	Giga5 16GB	ClueWeb 19GB	Common Crawl 110GB	CC100 2.5TB
BERT	✓							
XLNet	✓				✓	✓	✓	
RoBERTa	✓	✓	✓	✓				
DeBERTa	✓	✓	✓					
DeBERTa _{1.5B}	✓	✓	✓	✓				
DeBERTaV3	✓	✓	✓	✓				
mDeBERTa _{base}								✓

PRÉ-ENTRAÎNEMENT

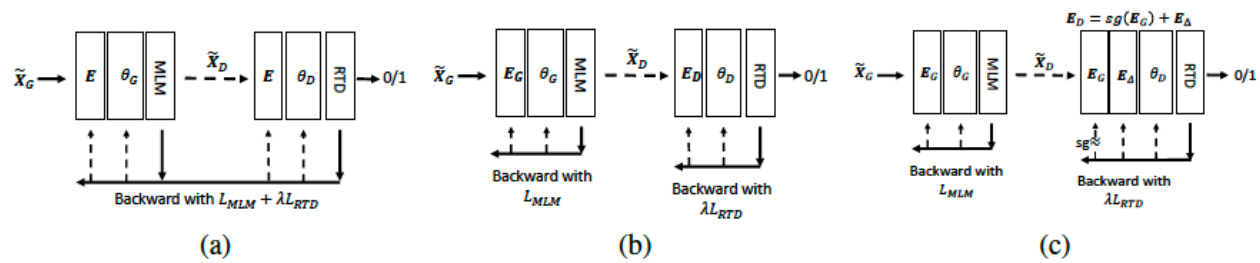
- Représentations de langage
- **161GB**
- Puissance de calcul massive



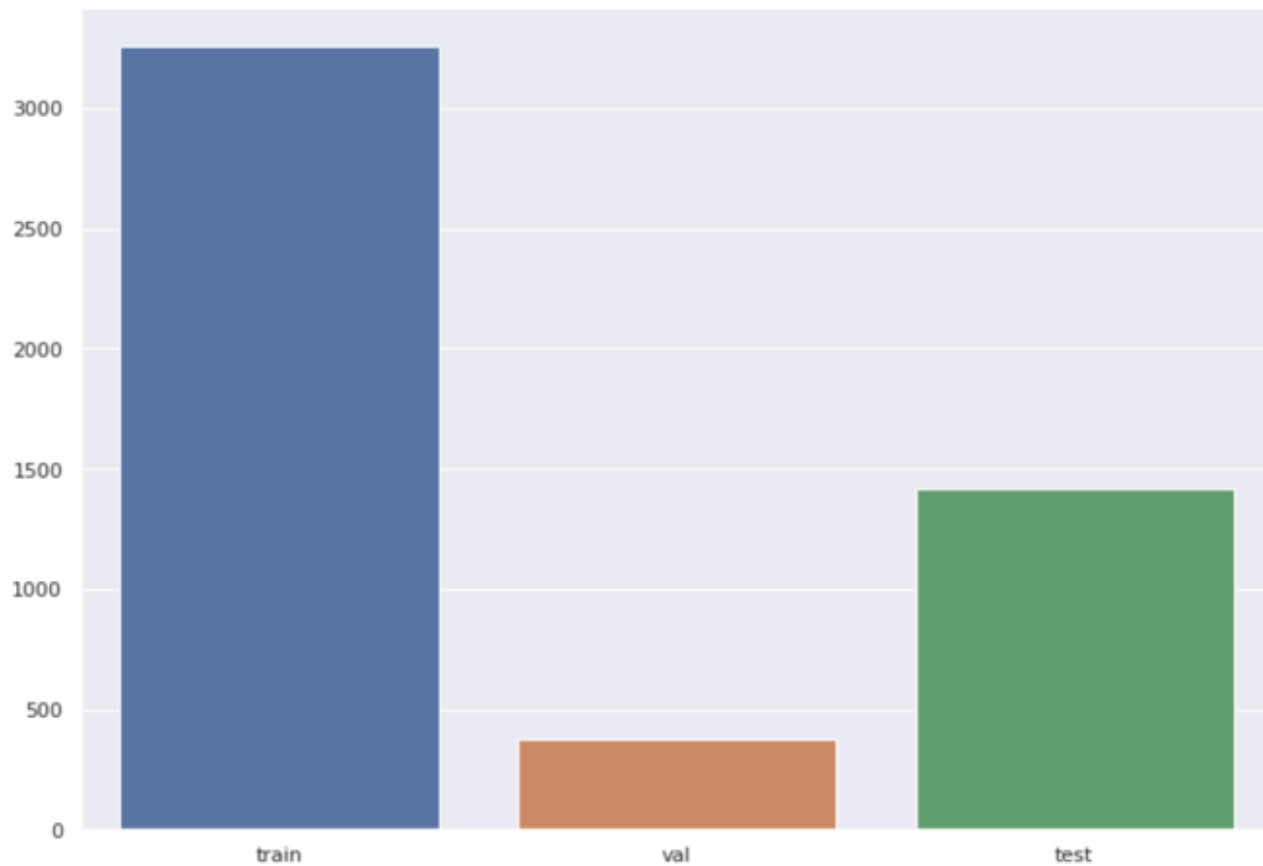
PRÉ-ENTRAÎNEMENT

- NSP
- MLM
- Tokens corrompus "[MASK]"
- 15%
- RTD (ELECTRA)

SÉQUENCES & GRADIENTS

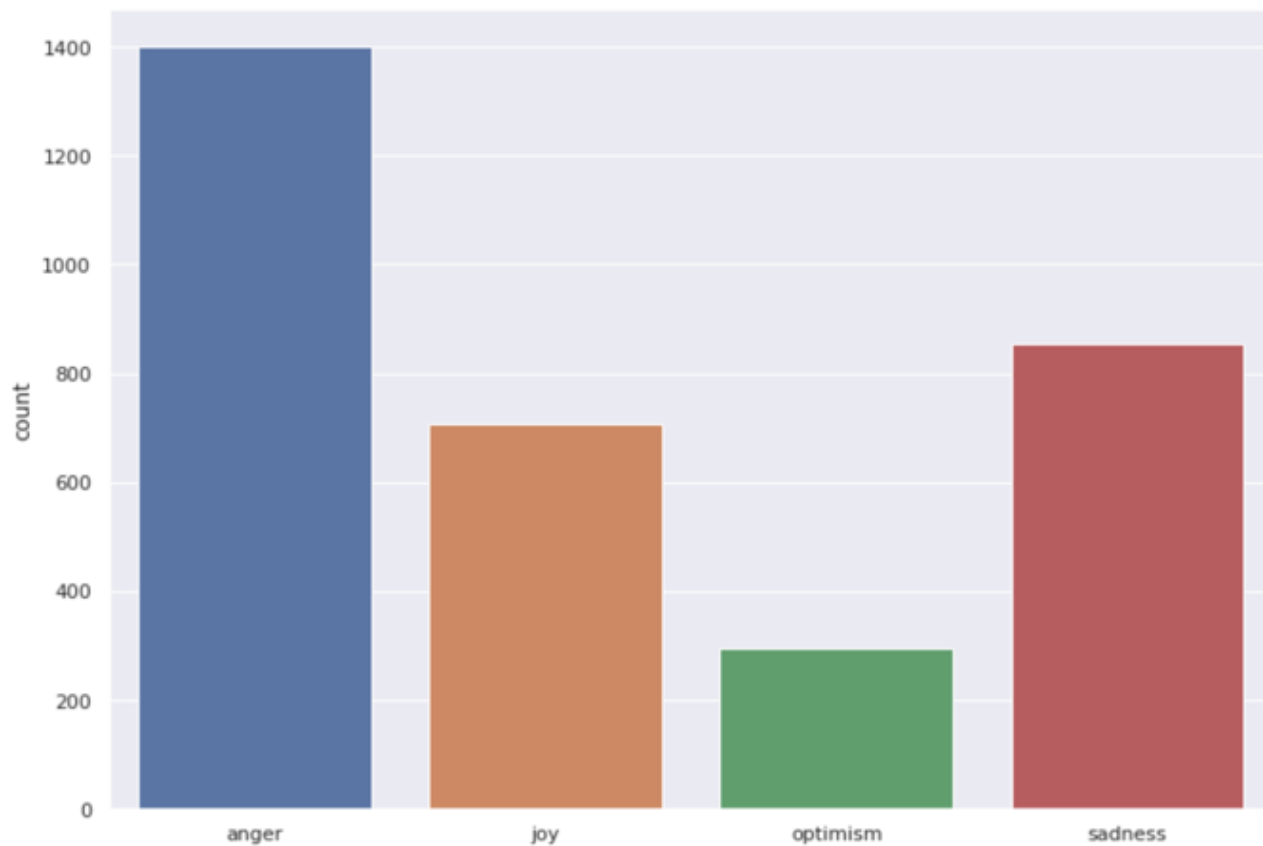


- Tug of war
- Partage de séquences (ES)
- Non partage de séquences (NES)
- **GDES**



JEU DE DONNÉES

- TWEETEEVAL Emotion
- 5052 tweets
 - Entrainement : 3257
 - Validation : 374
 - Test : 1421
- 4 classes
 - Anger
 - Joy
 - Optimism
 - Sadness



JEU DE DONNÉES

- Distributions déséquilibrée
- La classe modale “anger” représente 43% de jeu d’entraînement
- Score : F1 macro-moyenné

IMPLÉMENTATION

Classe
Modale

Régression
Logistique
(weighted)

Universal
Sentence
Encoder

BERT (base)

DeBERTaV3
(base)

IMPLÉMENTATION

- Sickit Learn
- HuggingFace
- PyTorch
- Google Colab (TPU)
- Weights and Biases
- `transformers_sentiment_imdb.ipynb`

RÉSULTATS

- Jeu de validation

<i>Validation</i>	CM	LR	USE	BERT (base)	DeBERTaV3 (base)
F1 macro	14.9	30.5	45.1	71.4	<u>76.8</u>
Accuracy	42.7	43.8	60.6	77.5	<u>83.6</u>

RÉSULTATS

- Jeu de test

<i>Test</i>	CM	LR	USE	BERT (base)	TimeLMS	DeBERTaV3 (base)
F1 macro	14.0	28.2	54.4	77.2	80.2	<u>82.4</u>
Accuracy	39.2	41.8	65.1	80.5	-	<u>84.7</u>



CONCLUSION

- Exploration technique transversale
- Découverte de Pytorch et Weights & Biases
- Classification multiple déséquilibrée
- 4 algorithmes implémentés
- Score supérieur au benchmark (jeu de test)

The background is a dark blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural networks, with lines and small circles representing nodes.

ÉCHANGE & QUESTIONS

Merci de votre attention !