

BERT

GRAMMARLY

REDA BENKIRANE



SOMMAIRE

- 01** BERT : TRANSFER LEARNING
- 02** DATASET
- 03** STRATÉGIE
- 04** EXPÉRIMENTATION
- 05** ÉVALUATION

BERT

Définition

BERT stands for Bidirectional Encoder Representations from Transformers :

- **Bidirectionnel** - pour comprendre le texte que vous recherchez, vous devrez regarder en arrière (les mots précédents) et en avant (les mots suivants).
- **Transformers** - L'article "Attention Is All You Need" a présenté le modèle Transformer. Le Transformer lit des séquences entières de Tokens à la fois.
- **Incorporation de mots contextualisée (pré-entraînée)** - Utilisation d'une méthode pour encoder les mots en fonction de leur sens/contexte.

BERT

TRANSFER LEARNING

- Le défi majeur en NLP est le manque de données pour des tâches spécifiques.
- BERT est pré-entraîné par Google sur un large ensemble de données (Wikipedia,..etc).
- BERT capte maintenant plusieurs caractéristiques linguistiques et est assez performant.
- Il suffit de l'ajuster (fine-tuning) à notre tâche.
- **Transfer Learning** est efficace, au lieu d'entraîner un modèle à zero !!

BERT

Pourquoi BERT

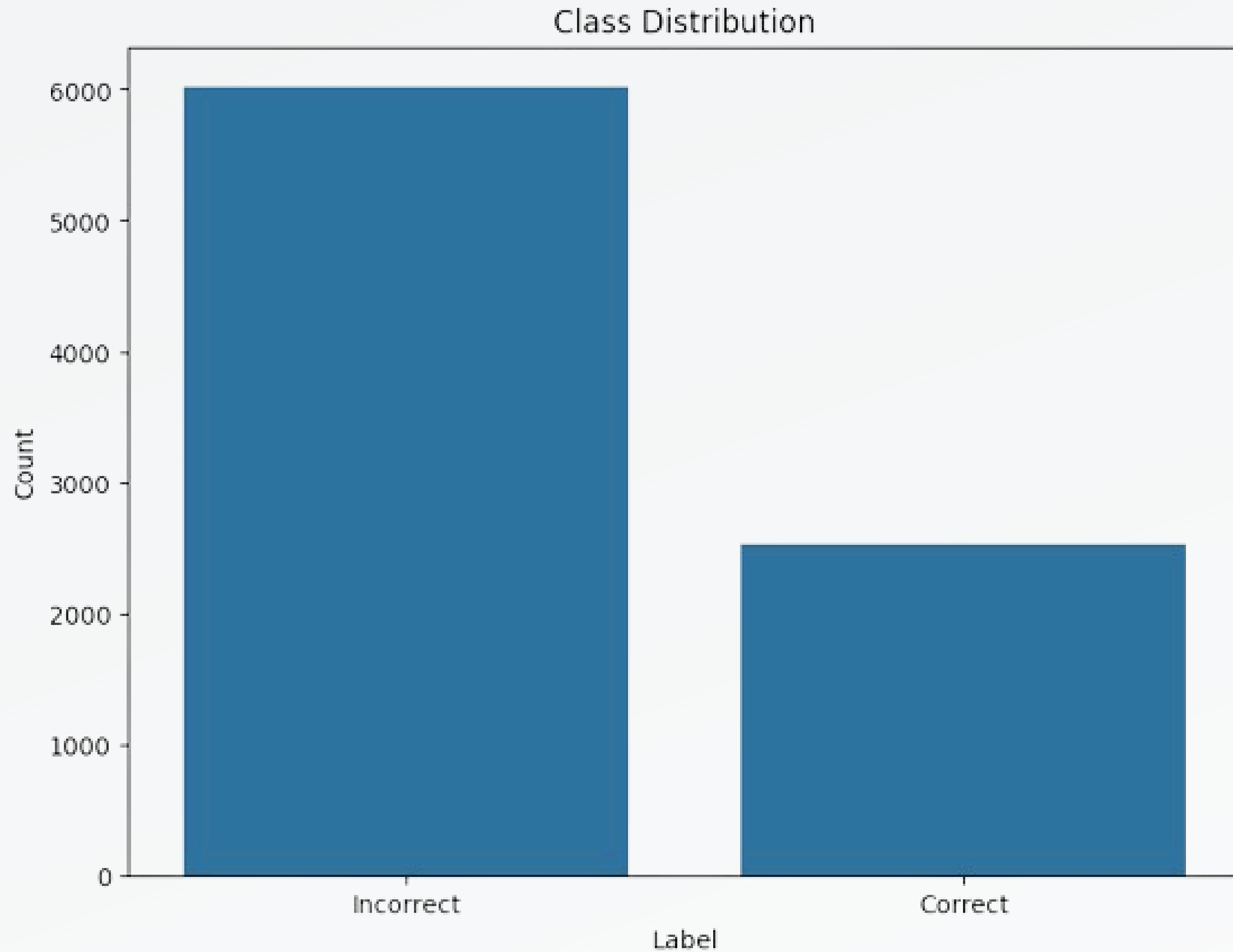
- Le modèle pré-entraîné prend moins de temps entrainer sur une tâche spécifique.
- L'entraînement nécessite moins de données que s'il soit entraîné à zero
- On a le privilège d'entrainer un grand modèle avec plus de 100 millions de paramètres avec moins de contraintes.
- Entrainer BERT sur notre tâche a prouvé avoir de très bons résultats .

BERT

Innovation

- BERT a été entraîné en masquant 15 % des Tokens dans le but de les deviner. Un objectif supplémentaire était de prédire la phrase suivante.
- Des modèles comme LSTM, traitent une phrase de manière uni-directionnelle, contrairement à BERT (bi-directionnel).
- Ceci lui permet de comprendre le contexte de la phrase pour prédire un mot (chaque mot a sa propre représentation contextuelle).

DATASET



STRATÉGIE

Problématique

- Le modèle s'entraînera sur une dataset déséquilibrée.
- Il a tendance à surajuster (overfit) en négligeant la classe minoritaire.
- On aura une mauvaise performance du modèle.

STRATÉGIE

Techniques

- Undersampling : réduit la taille de la classe majoritaire en supprimant des échantillons
- l'oversampling augmente la taille de la classe minoritaire en générant des données synthétiques ou en dupliquant des échantillons.

STRATÉGIE

Théorie et Techniques

- Supprimer des échantillons entrainera une très grande perte d'information
- Les grands modèles nécessitent une grande quantité de données
- Plus on a des données, plus on évitera le surajustement
- Les données synthétiques sont toujours bénéfiques, exemple : GAN, modèles de diffusion

EXPÉRIIMENTATION

Comparaison

- J'ai réussi à démontrer, expérimentalement, que l'oversampling est la meilleure méthode pour la performance du modèle.
- J'ai diminuer la fonction coût de 2% par rapport au cas de dataset déséquilibrée et de 3% par rapport au cas de dataset réduite.
- J'ai augmenté l'accuracy de 7% et la précision de 3% par rapport au cas de dataset déséquilibrée.
- J'ai augmenté l'accuracy de 11% et la précision de 10% par rapport au cas de dataset réduite.

ÉVALUATION

Modèle choisi

