RoBERTa

Qu'est ce que BERT

BERT est un modèle de langage, un modèle qui modélise la distribution de séquences de mots ou de lettres. Il a par exemple été utilisé dans la prédiction d'un mot en fonction des mots qui le précédent.

Et RoBERTa (Robustly optimized BERT approach)

RoBERTa est une méthode d'apprentissage basée sur BERT mais améliorée, qui atteint des résultats similaires voire supérieurs à bon nombre de méthodes.
Ces améliorations consistent notamment en l'agrandissement des données d'entraînement (et donc en entraînant le modèle plus longtemps), en enlevant l'objectif initial qui était la prédiction des mots suivants, en l'entraînant sur des phrases plus longues, et en changeant dynamiquement les motifs (les patterns) des données d'entraînement.

L'entraînement sur des données plus longues au lieu de données plus quantitatives a permis à ce modèle d'atteindre **un score de 88.5 sur GLUE**

En pratique

Les données utilisées ici sont les données CC-NEWS (CommonCrawl News dataset), qui est un ensemble de données constituées d'articles de news, d'informations, sur des sites de news du monde entier. D'autres sources de données sont également utilisées comme BOOKCORPUS, English Wikipedia, OPENWEBTEXT ou encore STORIES.

BERT prend en entrée 2 segments (chaque segment constitué de plus d'une phrase) séparés par des tokens spécifiques :
Les segments étant x1,...,xn et y1,...,ym : [CLS],x1,...,xn,[SEP],y1,...,ym,[EOS]
BERT aura alors 2 objectifs pendant l'entraînement :

- Masked Language Model (MLM) qui consiste à remplacer aléatoirement des tokens dans la donnée d'entrée par le token [MASK] afin d'ensuite prédire quel était le token d'origine
- Next Sentence Prediction (NSP) qui consiste à prédire si deux segments se suivent ou non dans le texte d'origine

Différents tests ont ensuite été réalisés afin d'étudier l'importance du type de données d'entrée ou de l'objectif NSP, ce qui a donné RoBERTa.

Résultats

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$\text{XLNet}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

et en utilisant spécifiquement les paramètres de la 4e ligne du tableau ci-dessus, afin d'avoir des résultats sur GLUE :

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-
Ensembles on test (from leaderboard as of July 25, 2019)										
ALICE	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8	86.3
MT-DNN	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2/89.8	98.6	90.3	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	88.4
RoBERTa	**90.8/90.2**	**98.9**	90.2	**88.2**	96.7	92.3	67.8	**92.2**	89.0	**88.5**