Comparison study of BERT, RoBERTa and DistilBERT

Introduction

The BERT algorithm from Google and the transformation based NLP methods have been very popular recently. After BERT is published by Jacob Devlin et. el there was a big trend of exploring similar models based on the original transformation based idea and coming up with variations which are targeted at a certain category of NLP. RoBERTa and DistilBERT are flavors of BERT which improves the performance and inference drawing speed of the original BERT model.

BERT

This is a bi-directional transformer for pre-training on a lot of text data to learn language representation that can be then used to achieve specific natural language processing tasks using machine learning. While the BERT is better on various parameters, the main improvement is on performance which can be attributed to the bi-directional transformation, pre-training tasks of Masked Language Model and Next Sentence Prediction.

Using BERT in the analysis

BERT is open sourced at https://github.com/google-research/bert and pre-trained for 104 languages with implementations in TensorFlow and Pytorch. It can be fine-tuned for several types of tasks, such as text similarity, text classification, text labeling such as parts of speech, question and answer, named entity recognition etc. However, pre-training BERT can be computationally expensive unless you use a GPU.
BERT authors have also released a single multi-lingual model trained on an entire Wikipedia dump of 100 languages. Multilingual BERT has a lower performance compared to English.
According to the blog, below are the statistics of training the BERT model.

	TPU Pod	TPU Chips	TPU Cores*1	PFLOPS*2	GPU*3
$BERT_{BASE}$	4 x 4 days	16 x 4 days	32 x 4 days	0.7 x 4 days	2 x 50-70 days
$BERT_{LARGE}$	16 x 4 days	64 x 4 days	128 x 4 days	2.9 x 4 days	8 x 50-70 days

Dataset above 100K shows robust performance. BERT outperforms many algorithms including Pre-Open AI SOTA, BiLSTM and OpenAI GPT.

RoBERTa

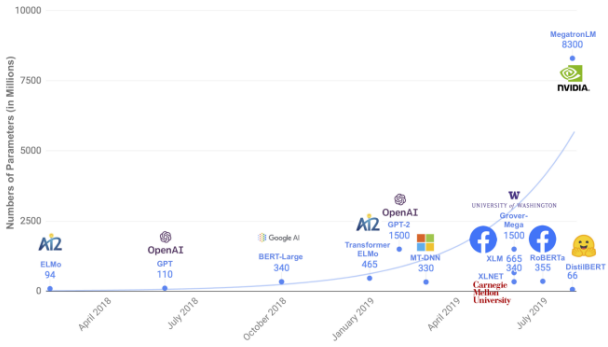
This algorithm was introduced by Facebook and very well optimizes original BERT. This is a retraining of BERT using the training data which is 10 times the original data.
To improve the training procedure, RoBERTa removes the Next Sentence Prediction (NSP) task from BERT's pre-training and introduces dynamic masking so that the masked token changes during the training epochs. Larger batch-training sizes were also found to be more useful in the training procedure.
Another important difference is RoBERTa uses additional web crawling data from stories and blogs.
Result of these retraining is RoBERTa outperforms both BERT and XLNet. As shown in the comparison table below.
The improvements in RoBERTa can easily be adapted as we see since they are mostly

related to changing the parameters in pretraining and using more data.

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$BERT_{LARGE}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$XLNet_{LARGE}$	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**

DistilBERT

DistilBERT is a distilled version of BERT. Smaller, faster, cheaper and lighter, Sanh et al.
The unique thing about this model is while other models aim to optimize BERT's performance, this model targets to reduce the large size and enhance speed while still keeping the accuracy and usability of the original model.



As shown in the above image from Sanh et al, the DistilBERT only has 66 parameters. It is much lower compared to RoBERTa's 355 and BERT-Large's 340.
Architecture of DistilBERT is similar to the original BERT model but has fewer encoder blocks. Also the token type embeddings and the pooler are removed.
DistilBERT also utilizes good practices from RoBERTa, that is a larger batch size, dynamic masking, and no Nest Sentence Prediction.
DistilBERT is pre-trained with the Masked Language Modeling tasks. Its objective is to optimize a triple loss:

- The language model loss, which is the same loss used in BERT.
- The distillation loss, which measures how similar the output of DistilBERT and of BERT.
- The cosine-distance loss, which measures how similar the hidden representation of DistilBERT and of BERT.
- The distillation loss and cosine-distance loss help train DistilBERT in a student-teacher way.

Conclusion

While the BERT created the revolution in the NLP and provided a new perspective to pretrain and keep the language model, it also changed the approach from traditional **BiLSTM** or the **Bag of Words** models. RoBERTa and DistilBERT are the variations of original BERT and improves performance and training time by utilizing the different techniques as we saw earlier. **XLNet** is also another popular variant of BERT and uses bi-directional transformation and autoregressive together. Overall the RoBERTa can be used for a vast variety of use cases, the DistilBERT can be used on the specific use cases where inference drawing is of key importance.

References

Cortiz, Diogo. “Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA.” *arxiv.org*, https://arxiv.org/ftp/arxiv/papers/2104/2104.02041.pdf.

Devlin, Jacob. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *arXiv.org*, https://arxiv.org/abs/1810.04805.

Khan, Suleiman. “BERT, RoBERTa, DistilBERT, XLNet — which one to use?” *Towards Data Science*, https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8.

Khan, Suleiman. “BERT Technology introduced in 3-minutes | by Suleiman Khan, Ph.D.” *Towards Data Science*, https://towardsdatascience.com/bert-technology-introduced-in-3-minutes-2c2f9968268c.

Sahn, Victor. “DistilBERT.” *arxiv*, https://arxiv.org/pdf/1910.01108.pdf.