XLM — Enhancing BERT for Cross-lingual Language Model

Cross-lingual Language Model Pretraining

Rani Horev Feb 12, 2019 · 5 min read

Attention models, and BERT in particular, have achieved promising results in Natural Language Processing, in both classification and translation tasks. A new paper by Facebook AI, named XLM, presents an improved version of BERT to achieve state-of-the-art results in both types of tasks.

XLM uses a known pre-processing technique (BPE) and a dual-language training mechanism with BERT in order to learn relations between words in

different languages. The model outperforms other models in a cross-lingual classification task (sentence entailment in 15 languages) and significantly improves machine translation when a pre-trained model is used for initialization of the translation model.

Background

XLM is based on several key concepts:

Transformers, invented in 2017, introduced an attention mechanism that processes the entire text input simultaneously to learn contextual relations between words (or sub-words). A Transformer includes two parts — an encoder that reads the text input and generates a lateral representation of it (e.g. a vector for each word), and a decoder that produces the translated text from that representation. A great in-depth review of Transformers can be found here.

While the vanilla Transformer has only limited context of each word, i.e. only the predecessors of each word, in 2018 the BERT model took it one step forward. It uses the Transformer's encoder to learn a language model by masking (dropping) some of the words and then trying to predict them,

allowing it to uses the entire context, i.e. words to the left and right of a masked word.

Due to the concurrent processing of all tokens in the attention module, the model needs more information about the position of each token. By adding a fixed value to each token based on its position (e.g. sinusoidal function) — a step named Positional Encoding — the network can successfully learn relations between tokens. Our summary of BERT can be found here.

In 2018, Lample et al. presented a translation model that combines Transformers and statistical phrase-based model (PBSMT). The latter is a probabilities table for pairs of phrases in different languages. An important concept in the paper is Back-Translation, in which a sentence is translated to the target language and back to the source. This concept enables using monolingual datasets, which are bigger and more common than bilingual datasets, in a supervised manner. One of the conclusions of Lample et al. is that initialization of the token embeddings is of high importance for the success of the model, especially when using Back-Translation. While the authors used a "simple" word embeddings using FastText, they suggest that "more powerful language models may further improve our results".

How XLM works

The paper presents two innovative ideas — a new training technique of BERT for **multilingual classification tasks** and the use of BERT as **initialization of machine translation models**.

Cross-lingual BERT for classification

Tough BERT was trained on over 100 languages, it wasn't optimized for multi-lingual models — most of the vocabulary isn't shared between languages and therefore the shared knowledge is limited. To overcome that, XLM modifies BERT in the following way:

First, instead of using word or characters as the input of the model, it uses Byte-Pair Encoding (BPE) that splits the input into the most common sub-words across all languages, thereby increasing the shared vocabulary between languages. This is a common pre-processing algorithm and a summary of it can be found here.

Second, it upgrades the BERT architecture in two manners:

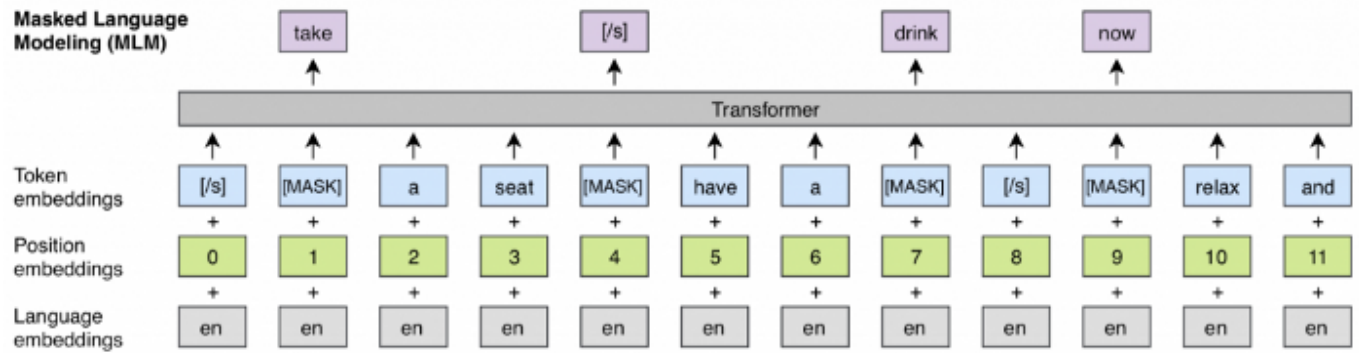
1. Each training sample consists of the same text in two languages, whereas in BERT each sample is built from a single language. As in

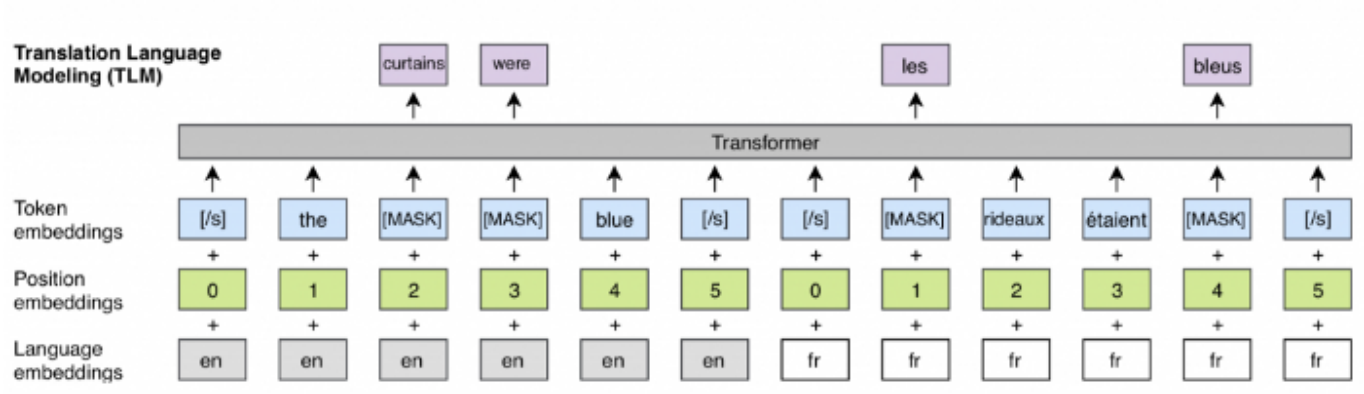
BERT, the goal of the model is to predict the masked tokens, however, with the new architecture, the model can use the context from one language to predict tokens in the other, as different words are masked words in each language (they are chosen randomly).

2. The model also receives the language ID and the order of the tokens in each language, i.e. the Positional Encoding, separately. The new metadata helps the model learn the relationship between related tokens in different languages.

The upgraded BERT is denoted as Translation Language Modeling (TLM) while the “vanilla” BERT with BPE inputs is denoted as Masked Language Modeling (MLM).

The complete model was trained by training both MLM and TLM and alternating between them.





Comparison of a single language modeling (MLM) similar to BERT, and the proposed dual-language modeling (TLM). Source: XLM

To assess the contribution of the model, the paper presents its results on sentence entailment task (classify relationship between sentences) using XNLI dataset that includes sentences in 15 languages. The model significantly outperforms other prominent models, such as Artetxe et al. and BERT, in all configurations — train only on English and test on all (Zero-Shot), train on translated data to English (Translate-Train), train on English, and test on translated data (Translate-Test). These results are considered state-of-the-art.

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2

Evaluation of cross-lingual sentence encoders															
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Comparison of XNLI results (accuracy) of prominent models in different training and testing configurations. Each column represents a language. Source: XLM

Initialization of translation models with MLM

The paper presents another contribution of BERT, and more precisely of the MLM model — as a better initialization technique for Lample et al. translation model. Instead of using FastText embeddings, the initial embeddings of the tokens are taken from a pretrained MLM and fed into the translation model.

By using these embeddings to initialize the tokens of both the encoder and the decoder of the translation model (which uses Transformer), the translation quality improves by up to 7 BLEU as shown in the table below.

Encoder	Decoder	en-fr	fr-en	en-de	de-en	en-ro	ro-en
FastText		29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3

CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

Translation results with different initialization techniques. CLM stands for Causal Language Modeling in which a given word is trained based only on the previous words and not using the masking technique. Source: XLM

Note: The paper also shows that training a cross-lingual language-model can be very beneficial for low-resource languages, as they can leverage data from other languages, especially similar ones mainly due to the BPE pre-processing. This conclusion is similar to the one from Artetxe et al. (Our summary can be found here).

Compute considerations

The models are implemented in PyTorch and can be found here, including pretrained models. The training was done with 64 Volta GPUs for the language modeling tasks and 8 GPUs for the translation tasks, though the duration isn't specified. Exact implementation details can be found in section 5.1 and 5.2 of the paper.

Conclusion

As in many recent studies, the paper shows the power of language models and transfer learning, and BERT in particular, to improve performance in many NLP tasks. By using simple but smart tweaks of BERT it can outperform other cross-lingual classification models and significantly improve translations models.

Interestingly, the translation model used in the paper and the MLM model that was used for initialization are both based on Transformer. It'd be safe to assume that we'll see more combinations of this kind, such as using the new Transformer-XL for initialization.

To stay updated with the latest Deep Learning research, subscribe to my newsletter on LyrnAI