Cross-lingual language model pretraining

Alexis Conneau and Guillaume Lample

Facebook AI Research
Why learning cross-lingual representations?

1

2

3

4

This is great.
C'est super.
Das ist toll.
Cross-lingual language models
Mult. Masked Language Modeling (MLM)

Similar to BERT, we pretrain a Transformer model with MLM but in many languages:

.. multilingual representations emerge from a single MLM trained on many languages.

Devlin et al. – BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding (+ mBERT)
Translation Language Modeling (TLM)

Multilingual MLM is unsupervised, but we leverage parallel data with TLM:

Translation language modeling (TLM) pretraining

.. to encourage the model to leverage cross-lingual context when making predictions.
Results on XLU benchmarks
Results on Cross-lingual Classification (XNLI)

The pretrained encoder is fine-tuned on the English XNLI(*) training data and then tested on 15 languages.

Average XNLI accuracy on the 15 languages of XNLI for zero-shot cross-lingual classification:

- XNLI baseline: 65.6
- mBERT: 66.3
- LASER: 70.2
- XLM (MLM): 71.5
- XLM (MLM+TLM): 75.1

(*) Conneau et al. – XNLI: Evaluating Cross-lingual Sentence Representations (EMNLP 2018)
Results on Unsupervised Machine Translation

Initialization is key in unsupervised MT to bootstrap the iterative BT process.

Initialization is essential for neural unsupervised MT (*)

---

Full Transformer model initialization significantly improves performance (+7 BLEU)

- Embeddings pretrained: 27.3 BLEU
- Full model pretrained (CLM): 30.5 BLEU
- Full model pretrained (MLM): 34.3 BLEU
- Supervised 2016 SOTA (Edinburgh): 36.2 BLEU

---

(*) Lample et al. – Phrase-based and neural unsupervised machine translation (EMNLP 2018)
Results on Supervised Machine Translation

We also show the importance of pretraining for generation

- Pretraining both the encoder and decoder improves BLEU score
- MLM better than LM pretraining
- Back-translation + pretraining leads to the best BLEU score
- Pretraining is more important when supervised data is small
Conclusion

• Cross-lingual language model pretraining is very effective for XLU

• New state of the art for cross-lingual classification on XNLI

• Reduces the gap between unsupervised and supervised MT

• Recent developments have improved XLM/mBERT models
Thank you!

Code and models available at github.com/facebookresearch/XLM

Lample & Conneau – Cross-lingual Language Model Pretraining (NeurIPS 2019)