Deep Learning for NLP 2020
Home Exercise 11 Solution

August 2, 2020

Miscellaneous questions

(a) **Multi-task learning**: What is the main difference between multi-task learning and transfer learning? Explain in at most three sentences. (2P)

(b) **Transformers**: What are positional encodings and why do transformers need them? Explain in at most three sentences. (2P)

(c) **GANs**: Which two of the following three high-level tasks are part of the conception of Generative Adversarial Networks: (i) classification, (ii) sequence tagging, (iii) generation? Which of those two tasks is more central to the main idea of GANs? (2P)

(d) **Evaluation**: Explain the main difference between reference-based and reference-free text generation evaluation. Name one advantage and one disadvantage of reference-free text generation evaluation over reference-based evaluation. (2P)

(e) **Zero-shot cross-lingual transfer**: what is meant by zero-shot cross-lingual transfer? (1P) Name two approaches to improve cross-lingual transfer using representations based on multilingual BERT. (1P)

Solutions

(a) Transfer learning transfers knowledge from one task to another. For example, BERT is first trained on the self-supervised task of language modeling and then fine-tuned on POS tagging. In contrast, multi-task learning learns several tasks jointly at the same time.

(b) Transformers have no notion of word order of the input, in contrast to RNNs. To remedy, a position-dependent signal is added to each word-embedding to help the model incorporate the order of words. See also: `https://datascience.stackexchange.com/questions/51065/what-is-the-positional-encoding-in-the-tra`

(c) Classification and generation. Generation is more central. The idea behinds GANs is to generate (text or images) in such a way that the classifier (discriminator) cannot distinguish real (= training data distribution) from fake.

(d) Reference-based compares a human reference to a system output. Reference-free compares the input to the system output. Advantage: no need for human references (cheaper). Disadvantage: Harder, and lower correlation with humans.

(e) Cross-lingual transfer: transfer across languages (e.g., train a classifier in English, apply to German). Zero-shot: no training data at all in the target language (e.g., German).

About multilingual BERT: mBERT uses no parallel data at all (such approaches are also called unsupervised methods for inducing multilingual vector spaces; see Lecture 6). To improve mBERT one can, for example, re-map it using a bit of parallel data (then it becomes supervised) or normalize the embeddings across languages.