CS6101-1910 / DYC1401 Deep Unsupervised Learning - Scribe Notes from Week 12

Language Models

Presenters: Lee Xin Jie, Si Chenglei, Li Xueqi, Liu Ziyang, Wang Wenjie
Authors: ***Hitoshi Iwasaki,*** Wu Zhaomin

Part 1: Language model and Their Uses

- Statistical / Probabilistic language modeling

  -- How should we compute the probability of a string?
  Some degree of flexibility should be preferred:
  grammar, word knowledge, sentiment analysis

  probability of being recognized as a sentence

  i. grammar
  ii. word knowledge
  iii. sentiment analysis

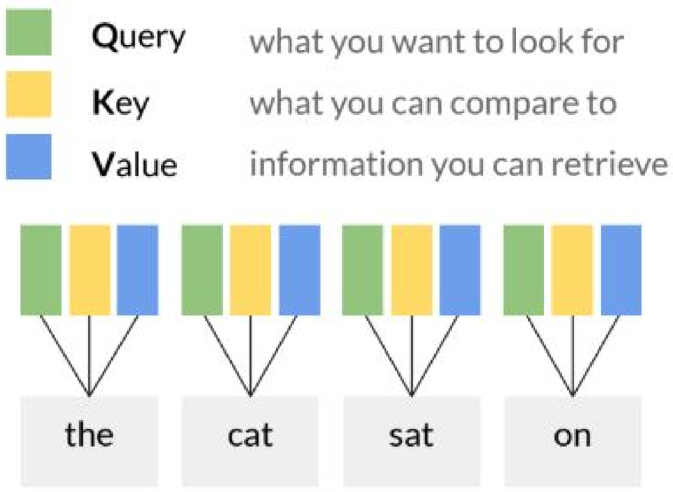
- How to compute probabilities
  b. Assume a uniform prior over tokens
  c. Assume all tokens are independent

  -- Infinite loss (probability = 0) should be avoided.

- Evaluation type 1: character level bits per character, word level perplexity
- Evaluation type 2: variety of ways of evaluation
  WER for speech recognition
  BLEU for translation
  F1 for POS tagging
  ACC for document classification
  - function of the length of the string
  - based on their usefulness for a downstream task

- GPT-1
  - LM-based
  - Fine-tune on supervised tasks
  - Analysis of BERT's Attention
  - Are Sixteen Heads Really Better than One?

  An attention layer has an access to all the hidden time steps.
  Attention block… Transformer.
  Dot product measure the closeness of the two vectors.


Query what you want to look for
Key what you can compare to
Value information you can retrieve
the
cat
sat
on

Q. Is it possible that many attention layers end up paying the same attention?
A. The idea is the same as the CNN's filters. They are designed to pick up differnt types of signals such as edges, brightness, and colors… Multi-heads attention layers should work similarly.

Part 2: Interpretability and Analysis of NLP Models and Datasets

- 1. Adversarial Attacks

  Modify the original tasks (documents or questions) not in favor of the model.
  How? -- change the original question or insert a fake sentence
  Humans are immune against these kinds of attacks. Humans are better at generalization.
  Even BERT is highly vulnerable to these kinds of tricks.

Adversial Attacks

- AddSent:
  - the adversarial sentence does not change the original one
- AddAny
  - only cares about whether the words decrease the sentence

- 2. Partial Training
  For reading comprehension task, a model can do a good job with documents and answers only (without questions). This is not supposed to be the case.

- 3. Probing Tasks
  What does BERT learn from multiple choice reading comprehension?
  BERT relies on certain types of key words.. Is it just learning BoW types of features?
  Partial training and shuffle training
  BERT still can learn and achieve good accuracy…

Part 3: Language Models and Their Developments

- GPT-2

  More data. Higher capacity model. No need for fine-tuning.

Language Models are Unsupervised Multitask Learners. Alec Radford et al., 2019

- T-5 (Text To Text Transfer Transformer)
  Language models are supposed to be evaluated by multi-tasks, no by a single task.
  The paper is a long experiment paper, consisting of 53 pages.

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Colin Raffel et al., 2019

Cool Links:

- GPT-2: 1.5B Release by OpenAI
- ULM FiT model: https://arxiv.org/abs/1801.06146
- AddSent and AddAny: http://stanford.edu/~robinjia/pdf/emnlp2017-adversarial.pdf
- HotFlip https://arxiv.org/abs/1712.06751
- Universal Adversarial Triggers (Hot off the press; just published a few days ago) https://arxiv.org/pdf/1908.07125.pdf
- To play with GPT