CS224n: NLP with Deep Learning **Winter 2019**

Lecture 8

Course Coordinator: Prof. Chris Manning *Scribes: Akash Gupta*

Machine Translation - The task of translating a sentence x from one language (source) to a sentence y in another language (target).

Statistical Machine Translation (SMT) -

- Learn a probabilistic model from the data.
- Formally,

  $$argmax_y P(y|x) = argmax_y P(x|y)P(y)$$

  $P(x|y)$ - Translation model which keeps account how words/phrases should be translated.
  $P(y)$ - Language model which helps to generate meaningful English sentences.
- Need large amount of parallel data for SMT. E.g.- Rosetta Stone.
- Break $P(x|y)$ to $P(x, a|y)$, a - alignment.
- Alignment is the correspondence between particular words in the translated sentence pair.
- Alignment - one-to-one, many-to-one, one-to-many, many-to-many
- Decoding for SMT using some heuristic search algo.
- Disadvantages - Systems are extremely complex, Lots of feature engineering required, extra resources, human effort, etc.

Neural Machine Translation (NMT) - Machine Translation using a single neural network and the NN is called seq2seq involving 2 RNNs - Encoder RNN and Decoder RNN.

- Need 2 sets of word embedding for each language.
- Seq2Seq models can be used for other tasks as well - Summarization, Dialogue, Parsing, Code generation.
- Seq2Seq - Conditional LM. Predicting the next word in target sentence with conditioning over source sentence.
- Seq2Seq is an example of learning system end-to-end since backprop goes all the way back to encoder RNN.
- Practically, you pad short sentences to max example in the batch so that batch is even sized tensor. Mathematically, the source are target sentences can be of different length.
- Greedy decoding - Taking argmax at each time-step. Problem - Taking argmax over each step isn't gonna give argmax over entire sentence.

Beam Search decoding - On each step of decoder, keep track of the k most probable partial translations (or hypotheses).

- k $\sim$ 5 to 10
- A hypotheses has a score which is it's log probability.

$$scores(y_1, ..., y_t) \ = \ logP_{LM}(y_1, ..., y_t|x) \ = \ \sum_{i=1}^{t} logP_{LM}(y_i|y1, ..., y_{i-1}, x)$$

  Scores are all negative, and higher score is better
  We search for higher-scoring hypotheses, tracking top k on each step.
- Beam Search is not guaranteed to find the optimal solution.
- In Beam search decoding, different hypotheses may produce $< END >$ tokens at different timesteps. Solution - Compute all those completed hypotheses separately and do this until some predefined threshold.
- Directly choosing highest score hypotheses is wrong since longer hypotheses will have lower scores. Solution - Normalize by length.

$$\frac{1}{t} \sum_{i=1}^{t} logP_{LM}(y_i|y1, ..., y_{i-1}, x)$$

Advantages of NMT - Better performance, much less human effort, single NN to be optimized whereas SMT has many subcomponents, Same method for different sentence pairs.

Disadvantages of NMT - less interpretable, difficult to control

Evaluation of Machine Translation -

- BLEU - Bilingual Evaluation Understudy.
- It compares machine written translation with human written translation and computes a similarity score based on n-gram precision, brevity penalty (for too short system translations)
- BLEU is useful but imperfect - good translation can have low n-gram precision.

Machine Translation problems - OoV words, Domain mismatch, Maintaining context over long sentences, Lower resources for a language.

Attention -

- Need - With just passing last hidden state to the decoder RNN in Seq2Seq. This could act as informational bottleneck.
- Attention is a general technique and is a like a selective summary of the information contained in the input values and where the output query determines which one to focus on.
- How to compute? Compute attention scores $- >$ Take softmax to get attention dist. $- >$ Using attention to take weighted sum of values to obtain attention output.(sometimes called context vector) $- >$ Concatenate with current hidden state in the decoder RNN to compute the output
- Advantages - improves NMT performance, solves bottleneck problem, helps with vanishing gradient problem since now we have direct connections, Provides some interpretability (Free alignment system)
- Many types of attention - Basic dot pdt, Multiplicative, Additive.