CS224N: Lecture 10 : Transformers and Pretraining

1. Subword modeling

   Finite vocabulary assuptions make even less sense in many languages.

      many languages exhibit complex morphology, or word structure

      the effect is more word types, each occurring fewer times

   Subword modeling in NLP encompasses a wide range of methods for reasoning about structure below the word level.

   Common words end up bing a part of the subword vocabulary, while rarer words are split into components. In the worst case, words are split into as many subwords as they have characters.

2. Motivating model pretraining from word embeddings

   The training data we have for our downstream task must be sufficient to teach all contextual aspects of language

   Most of the parameters in our network are randomly initialized

   In modern NLP

   -All parameters in NLP networks are initialized via pretraining

   -Pretraining methods hide parts of the input from the model, and train the model to reconstruct those parts.

   This has been exceptionally efffective at building strong

   -representations of language

   -parameter initializations for strong NLP models

   -Probability distributions over language that we can sample from

   Recall the language modeling task

   -Model p, the probability distribution over words given their past contets

   -There's lots of data for this

   Pretraining through language modeling

   -Train a neural network to perform language modeling on a large amount of text

   -Save the network parameters

3. Model pretraining three ways

   Decoders

   -Language models. What we've seen so far

   -Nice to generate from can't condition on future words

Encoders

-Gets bidirectional context-can condition on future

Encoder-Decoders

-Good parts of decoders and encoders

-What's the best way to pretrain them

Generative Pretrained Transformer(GPT)

-Natural Language Inference: Label pairs of sentences as entailing/contradictory/neural

4. Interlude: what do we think pretraining is teaching?

5. Very large models and in-context learning

   BERT: Bidirectional Encoder Representations from Tranformers

   -Predict a random 15% of subword tokens

   -Two models were released: BERT-base, BERT-large

   -Trained on: BooksCorpus, English Wikipedia

   -Pretraining is expensive and impractical on a single GPU

   -Finetuning is practical and common on a single GPU

   -QQP: Quora Question Pairs, QNLI: natural language inference over question answering data, SST-2: sentiment analysis, CoLA: corpus of linguistic acceptability, STS-B: semantic textual similarity, MRPC: microsoft paraphrase corpus, RTE: a small natural language inference corpus