TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Pretraining for NLP

Pretraining for NLP

In NLP unsupervised pretraining is now required for strong benchmark performance.

Pretrained Word Embeddings

Advances in Pre-Training Distributed Word Representations, Mikolov et al., 2017

We want a mapping from a word w to a vector $e(w)$ — a word embedding.

fastText from Facebook is currently popular.

It provides both contextual bag of words (cbow) and byte pair encoding (BPE) word vectors.

cbow word vectors

We construct a population distribution on pairs (c, w) here c is a bag of word context and w is a word.

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, E_{c,w} \; -\ln P(w|c)$$

Φ consists of a matrix $e[w, i]$ where $e[w, I]$ is the word embedding of w, and a matrix $e'[w, i]$ giving the embedding of the word w when it appears in a context.

A score $s(w|c)$ is defined by

$$s(w|c) = \frac{1}{|c|} \sum_{w' \in c} e(w)^\top e'(w')$$

Negative Sampling in cbow

Rather than define $P_\Phi(w|c)$ by a softmax over w, one uses restricted negative sampling.

We construct a training set of triples (w, c, N_C)

$$\Phi^* = \underset{\Phi}{\text{argmin}}\ E_{w,c,N_c}\ \ln\left(1 + e^{-s(w,c)}\right) + \sum_{n \in N_C} \ln\ \left(1 + e^{s(n,c)}\right)$$

Byte Pair Encoding (BPE)

BPE constructs a set of character n-grams by starting with the unigrams and then greedily merging most common bigrams of n-grams.

Given a set of character n-grams each word is treated as a bag of character n-grams.

$$e[w] = \frac{1}{N} \sum_{n \in w} e(n)$$

Current systems use byte pairs but train the byte pair embeddings as part of transformer training.

BERT: Blank Languagage Modeling

We replace a random subset of the words with a blank token.

We run a transformer on a block of text containing some blanks.

For a blank occurring at position t we predict the word at position t:

$$P(w) = \underset{w}{\text{softmax}}\ h[t, J]e[w, J]$$

Blank language modeling outperforms language modeling when used for pretraining in classification tasks such as the GLUE tasks.

GLUE

GLUE: General Language Understanding Evaluation

ArXiv 1804.07461

GLUE Leader Board as of February 27, 2020

SuperGLUE Leader Board as of February 27, 2020

Fine Tuning on Question Answering

COMET: Busselut et al, June 2019.

Charlie is drifting though life:

The Chatbot Meena

The Chatbot Meena

END