HW1: Classification

Jiafeng Chen Francisco Rivera

February 19, 2019

1 Introduction

In this write-up, our main focus is language modeling. That is, given words in a sentence, can we predict the word that follows? We implemented a Trigram model, a embedding neural network model, an LSTM, and a few extensions—including pre-trained embeddings, ensemble models, and multi-head attention decoders.

2 Problem Description

To tackle language modeling, we start with sequences of words $w \in \mathcal{V}$ in some vocabulary $\mathcal{V}$ and aim to predict the last word in the sequence which we cannot observe. We can do this probabilistically by attempting to estimate,

$$p(w_t \mid w_1, \ldots, w_{t-1}) \tag{1}$$

that is, the conditional distribution over the last word conditional on the words leading up to it.

In particular, there will be a special word, $w_{\text{unk}} \in \mathcal{V}$ which represent an unknown word; we use this whenever we encounter a token we have not previously seen.

In some models, we represent words with dense embeddings. That is, each word gets assigned a vector $v \in \mathbb{R}^d$ where d is the embedding dimension. These embeddings are trained as part of the model, but can also be initialized to pre-trained values.

3 Model and Algorithms

3.1 Trigram model

In our trigram model, we aim to estimate the probability written in Equation 1. This conditional probability is intractable itself because it's likely that we've never seen the exact sequence of words $w_1, \ldots, w_{t-1}$. However, we can gain tractability by dropping words toward the beginning of the sequence, hoping that they don't affect the probability too much. That is, we hope that,

$$p(w_t \mid w_1, \ldots, w_{t-1}) \overset{?}{\approx} p(w_t \mid w_{t-2}, w_{t-1}).$$

Having replaced our first probability with a simpler one which conditions on less information, we can estimate the latter by its empirical sample estimator. In other words, we can take all the times in our training set when we've seen words w_{t-2}, w_{t-1} adjacent to each other, and consider the empirical distribution of the word that follows them. We represent this sample approximation as $\widehat{p}$ and write,

$$p(w_t \mid w_{t-2}, w_{t-1}) \approx \widehat{p}(w_t \mid w_{t-2}, w_{t-1}).$$

By doing this, we've solved most of the intractability of conditioning on the entire sentence $w_1, \ldots, w_{t-1}$, but we still have some of the same problems. Namely, it's possible that in our training set, we either haven't seen words w_{t-2} and w_{t-1} together before, or we've seen them only a very small number of times such that the empirical probability distribution becomes a poor approximation. (To avoid division by zero errors, we adopt the convention that empirical probabilities are all 0 if we haven't seen the words being conditioned on before.) We can fix this by also considering the probabilities,

$$p(w_t) \text{ and } p(w_t \mid w_{t-1})$$

which give us the unconditional probability of a word and the probability conditional on only the previous word. These have the benefit of being more tractable to estimate and the drawback of losing information. In the end, we calculate a blend of these three approximations:

$$\alpha_1 \widehat{p}(w_t \mid w_{t-2}, w_{t-1}) + \alpha_2 \widehat{p}(w_t \mid w_{t-1}) + (1 - \alpha_1 - \alpha_2)\widehat{p}(w_t).$$

Training the weights (α_1, α_2) loads up most of our weight on α_1 which suggests the latter two probabilities are better used as "tie-breakers" when conditioning on the previous bi-gram yields a small number of possibilities. In our final model, we use $(\alpha_1, \alpha_2) = (0.9, 0.05)$.

3.2 Neural Network Language Model

Following Bengio et al. (2003), we implement a neural network language model (NNLM). We model (1) by first assuming limited dependence:

$$p(w_i \mid w_1, \ldots, w_{i-1}) = p(w_i \mid w_{i-1}, \ldots, w_{i-k}),$$

i.e., the current word only depends on the past k words, a useful restriction motivated by n-gram models. Next, we convert input tokens $w_{i-1}, \ldots, w_{i-k}$ into embedding vectors $\{v_{i-t}\}_{t=1}^{k}$ and concatenate them into a dk vector $v_{i-k:i-1}$. We then pass this vector into a multilayer perceptron network with a softmax output into $|\mathcal{V}|$ classes. We implement this efficiently across a batch by using a convolution operation, since convolution acts like a moving window of size k. This way we can generate $T - k + 1$ predictions for a sentence of length T. We depict the convolution trick in Figure 1.

3.3 LSTM

A concern with our trigram model is that it completely ignores words more than two positions before the word we wish to predict. To the extent we believe these words are predictive (personal experience with language suggest that they should be!), the trigram model has an inherent limitation in its ability to model them.

One way to combat this is with an LSTM.

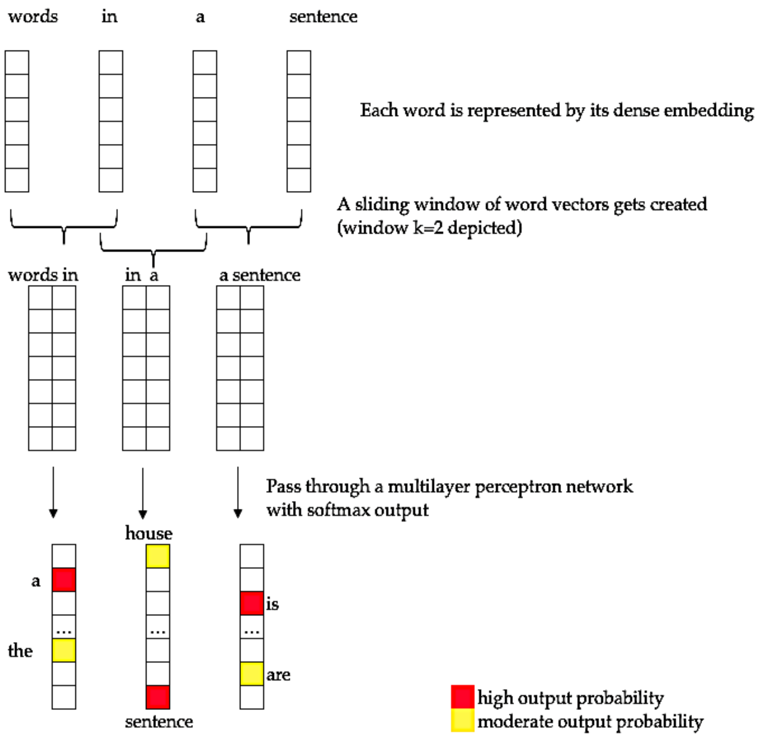

words
in
a
sentence
Each word is represented by its dense embedding
A sliding window of word vectors gets created (window k=2 depicted)
words in
in a
a sentence
Pass through a multilayer perceptron network with softmax output
house
a
the
is
are
sentence
high output probability
moderate output probability

Figure 1: Diagram for Section 3.2

3.4 Multihead Attention

Following Vaswani et al. (2017), we implement a variant of the multi-head attention decoder network. Instead of passing the last hidden state from an LSTM h_t to predict h_{t+1}, we use a concatenation of h_t and a *context vector* $c_t = a_1 h_1 + \cdots + a_{t-1} h_{t-1}$ for *attention values* $a_1, \ldots, a_{t-1}$ residing in the unit simplex. Following Vaswani et al. (2017), we use the *scaled attention* mechanism, where

$$\boldsymbol{a} = \mathsf{Softmax}\left(\left\{\frac{h_i^T h_t}{\sqrt{\dim(h_t)}}\right\}_{i=1}^{t-1}\right).$$

In *multi-head attention*, we repeat the attention mechanism above on different linear projections of $h_1, \ldots, h_t$, with the motivation being that we wish to capture similarity on different projections of words—one attention layer could be capturing grammar, another semantics, a third pronoun association, etc. We depict the architecture in Figure 3, for $t = 3$ and predicting $t + 1$.

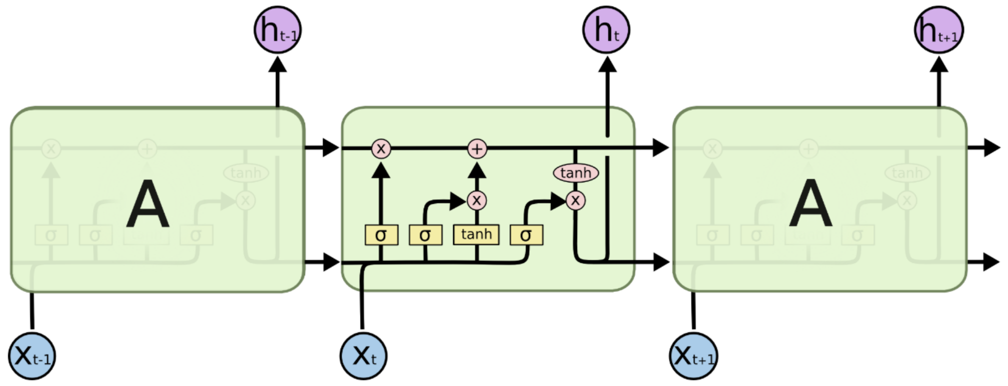


Figure 2: Depiction of LSTM working

Certain computational tricks need to be employed for efficient utilization of the GPU. Unlike the encoding attention network, the decoder cannot condition on future information when predicting the future. As a result, each attention layer can only look at past hidden states. We parallelize the procedure and take advantage of GPU hardware by applying attention as usual, computing every inner product $h_i^T h_j$ for all i, j, and use a mask that sets entries with $h_i^T h_j$ to $-\infty$ if $j \leq i$ (which correspond to the forbidden attention links by looking ahead) before applying the softmax.

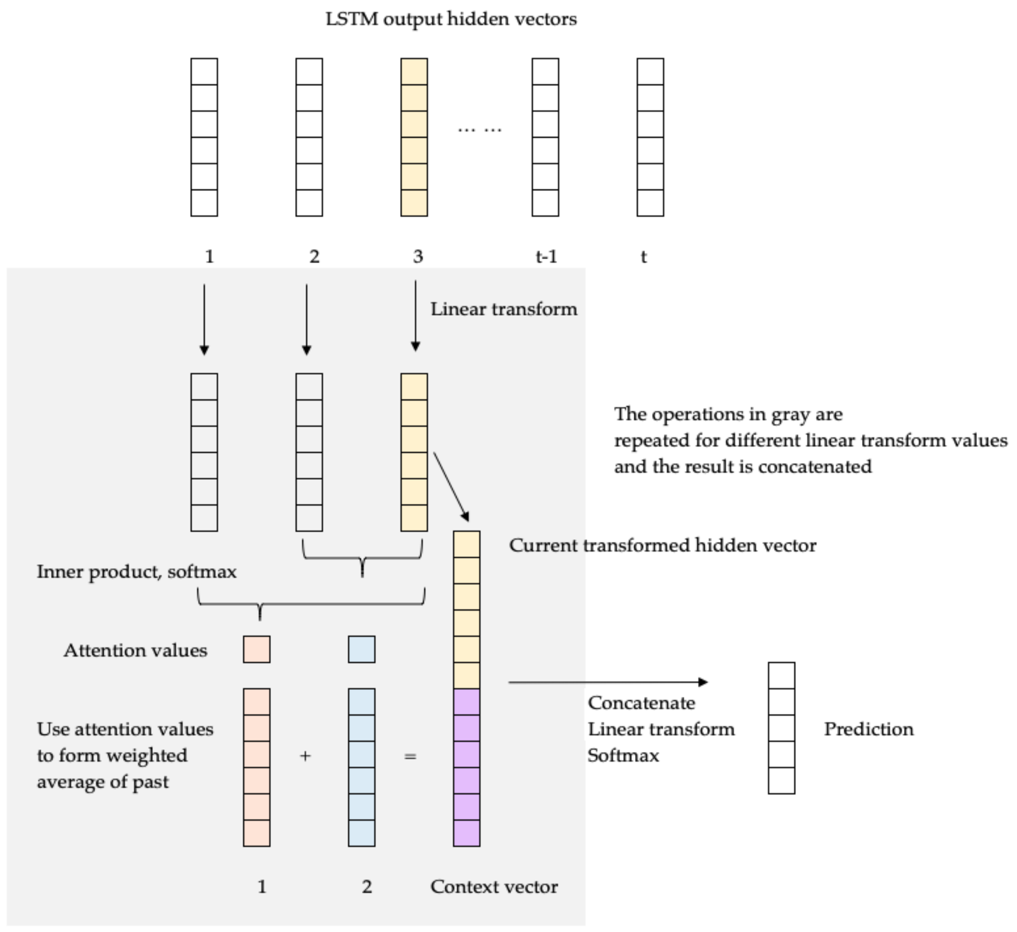
4 Experiments

5 Conclusion

References

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.



LSTM output hidden vectors
... ...
1
2
3
t-1
t
Linear transform
The operations in gray are repeated for different linear transform values and the result is concatenated
Current transformed hidden vector
Inner product, softmax
Attention values
Use attention values to form weighted average of past
+
=
Concatenate
Linear transform
Softmax
Prediction
1
2
Context vector

Figure 3: Diagram for Section 3.4. In this diagram, we predict the fourth word in the sentence by conditioning on the first three. We compute attention values of h_3 with h_2 and h_1 and concatenate h_3 with a context vector $a_1 h_1 + a_2 h_2$, where a_i are the attention values. We pass the resulting vector through a linear transformation and softmax output. In multi-head attention, we repeat the process for different projections of h_1, h_2, h_3 and concatenate the results.