

Problems For Language Modeling, Translation and Attention.

[illegible]

(a) Let k range over the set of possible labels in a dependency parse. There is typically a small fixed set of such labels. If we interpret k has a transformer head, what attention $\alpha(\mathbf{dobj}, \mathbf{prefer}, w)$ over the words w attended from the word **prefer** by the head **dobj** corresponds to the above dependency parse?

$$\alpha(\text{dobj}, \text{prefer}, w) = \begin{cases} 1 & \text{if } w \text{ is } \textbf{flight} \\ 0 & \text{otherwise} \end{cases}$$

Solution:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{focus}} \\ \mathcal{L}_{\text{focus}} &= \frac{1}{KTL} \sum_{\ell, k, t_1} H(k, t_1, T_2) \\ &= \frac{1}{KTL} \sum_{\ell, k, t_1} \sum_{t_2} \alpha[\ell, k, t_1, t_2] (-\ln \alpha[\ell, k, t_1, t_2])\end{aligned}$$

The factor of $\frac{1}{KTL}$ can be incorporated into λ which is also a correct solution although it introduces a stronger coupling of the hyper-parameter λ with the size of the transformer model.

(c) Dependency edges tend to be between nearby words. Repeat part (b) but for a loss $\mathcal{L}_{\text{near}}$ which encourages the attention $\alpha[\ell, k, t_1, T_2]$ to be focused near t_1 . The loss $\mathcal{L}_{\text{near}}$ should be “robust” in the sense that it has a maximum value that is independent of the length T of the transformer window. This should allow some “outlier” long distance attentions which are needed for coreference.

Solution:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{near}}$$

$$\mathcal{L}_{\text{near}} = \frac{1}{KT(L-1)} \sum_{k, t_1, 0 \leq \ell \leq L-2} \min \left(L_{\text{max}}, \sum_{t_2} \alpha[\ell, k, t_1, t_2] (t_1 - t_2)^2 \right)$$

Other solutions are also possible.

(d) State the “universality assumption” under which the “loss shaping” terms of (b) and (c) above only hurts the language modeling performance. Also give a plausibility argument that these terms might help in practice.

Solution: The universality assumption is that the model parameters will be optimized so as to find the actual minimum of the loss on the population distribution. In that case the loss shaping terms can only hurt performance. If, on the other hand, the training does not find the optimal parameters on the population then it seems plausible that the loss shaping terms might help with the optimization or the generalization.

Problem 2. The Adjusting Temperature for Dimension. For a typical language model the softmax operation defining the probability $P(w_{t+1} \mid w_1, \dots, w_t)$ has the form

$$\alpha[\ell, t, w] = \text{softmax}_w h[\ell, t, I] e[w, I]$$

We now consider adding a “temperature parameter” β to this softmax.

$$\alpha[\ell, t, w] = \text{softmax}_w \beta h[\ell, t, I] e[w, I] \quad (1)$$

(a) Assume that the components of the vector $h[t, I]$ are independent with zero mean and unit variance. Also assume that the word vectors have been initialized so that the components of the vector $e[w, I]$ are zero mean and unit variance. What initial value of β gives the result that the inner product $\beta h[t, I] e[w, I]$ has zero mean and unit variance. Explain your answer. (Use I to denote the dimension of the vectors $h[t, I]$ and $w[w, I]$.)

Solution: The mean of a sum of zero mean variables has zero mean. The standard deviation of a sum of I independent variables each with unit variance is \sqrt{I} . Therefore we get unit variance by setting $\beta = \frac{1}{\sqrt{I}}$.

(b) Relate your answer to (a) to the equation used for the self attention $\alpha(k, t_1, t_2)$ computed in the transformer.

Solution: The transformer attention is given by

$$\alpha[\ell, k, t_1, t_2] = \text{softmax}_{t_2} \frac{1}{\sqrt{I}} \text{Query}[\ell, k, t_1, I] \text{Key}[\ell, k, t_2, I]$$

This has the same “temperature parameter” as the answer to part (a).

Problem 3. Parameterizing Inner-Product This problem is on transformer self-attention. Modern classification problems tend to use a softmax operation of the form

$$P(y|h) = \text{softmax}_y h^\top e(y) \quad (2)$$

where h is a vector computed by the neural network and $e(y)$ is a vector embedding for the label y . Many early systems would insert a parameter matrix so that we have

$$P(y|h) = \text{softmax}_y h^\top W e(y) \quad (3)$$

However, it was generally observed that additional parameterization of the inner product operation does not improve the results. The vector h and the embedding $e(y)$ can be learned to be such that the standard inner product works well. However, the attention softmax of the transformer (3) does not use a naive inner product.

(a) Explain why we cannot replace (3) with a naive inner product of $L[\ell, t_1, J]$ and $L[\ell, t_2, J]$ as in (2).

Solution: The attention $\alpha[\ell, k, t_1, t_2]$ need to depend on the particular head k . Different heads compute different information.

(b) Rewrite the transformer self-attention equation (2) in the form of (3) where the matrix W in (3) is replaced by by a matrix defined in terms of W^K and W^Q .

Solution:

$$\begin{aligned} \alpha[\ell, k, t_1, t_2] &= \text{softmax}_{t_2} \frac{1}{\sqrt{I}} L[\ell, t_1, J] W^{QK}[\ell, k, J, J'] L[\ell, t_2, J'] \\ w^{QK}[\ell, k, j_1, j_2] &= W^Q[\ell, k, I, j_1] W^K[\ell, k, I, j_2] \end{aligned}$$

Problem 4. Repetition at Low Temperatures. For low temperatures and modest-sized language models we tend to generate infinitely repeating infinite sentences. We can get insight into this phenomenon by considering a trigram model where each word is predicted from the two preceding words using a conditional probability $P_{\Phi}(w_{t+2}|w_t, w_{t+1})$. We will assume trained word embeddings $e(w)$ for each word w and a neural network predictor of the form

$$P(w_{t+2}|w_t, w_{t+1}) = \underset{w_{t+2}}{\text{softmax}} \beta h_{\Phi}(e(w_t), e(w_{t+1}))^{\top} e(w_{t+2})$$

where h_{Φ} is some arbitrary neural network returning a query vector and β is a temperature parameter. We will assume that the model has been trained with β held fixed at 1 but that we will generate from this trigram model with different values of β . What degenerate behavior are we guaranteed to see if we sample at zero temperature? Explain your answer.

Solution: We are guaranteed to eventually repeat a bigram (a pair of words). At zero temperature the generation is deterministic and hence when a pair of words is repeated we must then see exactly the sequence following the first occurrence of that bigram and the generation enters a deterministic loop. This happens in practice with many transformer models when sampling at low temperatures.

Problem 5. Eliminating the Key Matrix. The self-attention in the transformer is computed by the following equations.

$$\text{Query}_{\ell+1}[k, t, i] = W_{\ell+1}^Q[k, i, J] L_{\ell}[t, J]$$

$$\text{Key}_{\ell+1}[k, t, i] = W_{\ell+1}^K[k, i, J] L_{\ell}[t, J]$$

$$\alpha_{\ell+1}[k, t_1, t_2] = \underset{t_2}{\text{softmax}} \left[\frac{1}{\sqrt{I}} \text{Query}_{\ell+1}[k, t_1, I] \text{Key}_{\ell+1}[k, t_2, I] \right]$$

Notice that here the shape of W^Q and W^K are both $[K, I, J]$. We typically have $I < J$ which makes the inner product in the last line an inner product of lower dimensional vectors.

(a) Give an equation computing a tensor $\tilde{W}^Q[K, J, J]$ computed from W^Q and W^K such that the attention $\alpha(k, t_1, t_2)$ can be written as

$$\alpha_{\ell+1}(k, t_1, t_2) = \underset{t_2}{\text{softmax}} \left[L_{\ell}[t_1, J_1] \tilde{W}^Q[k, J_1, J_2] L_{\ell}[t_2, J_2] \right]$$

For a fixed k we have that $W^Q[k, I, J]$ and $W^K[k, I, J]$ are matrices. We want a matrix $\tilde{W}^Q[k, J, J]$ such that the attention can be written in matrix notation as $h_1^{\top} \tilde{W}^Q h_2$ where h_1 and h_2 are vectors and \tilde{W}^Q is a matrix. You need write

this matrix \tilde{W}^Q in terms of the matrices for W^Q and W^K . But write your final answer in Einstein notation with k as the first index.

Solution: This is easier to do in vector-matrix notation for a fixed k . But it can also be done entirely in Einstein notation:

$$\begin{aligned}
&= \operatorname{softmax}_{t_2} \left[\frac{1}{\sqrt{I}} (L_\ell(t_1, J_1) W^Q[k, I, J_1]) (W^K[k, I, J_2] L_\ell(t_1, J_2)) \right] \\
&= \operatorname{softmax}_{t_2} \left[\frac{1}{\sqrt{I}} \sum_{j_1, j_2, i} L_\ell(t_1, j_1) W^Q[k, i, j_1] W^K[k, i, j_2] L_\ell(t_1, j_2) \right] \\
&= \operatorname{softmax}_{t_2} \left[\sum_{j_1, j_2} L_\ell(t_1, j_1) \left(\frac{1}{\sqrt{I}} \sum_i W^Q[k, i, j_1] (W^K[k, i, j_2]) \right) L_\ell(t_1, j_2) \right] \\
&= \operatorname{softmax}_{t_2} \left[\sum_{j_1, j_2} L_\ell(t_1, j_1) \tilde{W}^Q[k, j_1, j_2] L_\ell(t_1, j_2) \right] \\
&= \operatorname{softmax}_{t_2} \left[L_\ell(t_1, J_1) \tilde{W}^Q[k, J_1, J_2] L_\ell(t_1, J_2) \right] \\
\tilde{W}^Q[k, j_1, j_2] &= \frac{1}{\sqrt{I}} W^Q[k, I, j_1] W^K[k, I, j_2]
\end{aligned}$$

(b) Part (a) shows that we can replace the key and query matrix with a single query matrix without any loss of expressive power. If we eliminate the key matrix in this way what is the resulting number of query matrix parameters for a given layer and how does this compare to the number of key-query matrix parameters for a given layer in the original transformer version.

Solution: The original version uses $2(I \times J)$ key-query matrix parameters for each head. If we use only the single query matrix we use J^2 parameters for each head. These are the same for $I = J/2$.