ML & Opt Lecture 23

Sequence-to-Sequence Modeling
- architecture using RNNs
- training & inference
- teacher forcing
- perplexity - performance metric
- tricks for better performance
- attention
Sequence-to-Sequence Modeling

"Sutskever et al.
Sequence to sequence learning
with neural networks"

Goal: given input-sequence of tokens \((x_1, \ldots, x_T)\)
predict matching output-sequence of tokens \((y_1, \ldots, y_{T'})\)

Canonical example: Machine Translation

Issue: \(T \neq T'\) in general so the many-to-many design paradigm for RNNs may not fit

Soln: encoder-decoder architecture
Sequence Encoder-Decoder architecture

Encoder:
- h_o
- muy
- bien
- te

Decoder:
- Very
- well
- you
- <END>
- <START>
- very
- well
- and
- you

(Teacher forcing; only possible during training; helps train faster)

for LSTM, really cell & hidden states
Training

- train the encoder & decoder simultaneously (using BPTT) so that given the input sequence \((x_1, \ldots, x_T)\) and output sequence \((y_1, \ldots, y_T)\)
- encoder learns an informative \(h_T\) for decoding
- decoder learns how to, given \(h_T\) and a special \(<\text{START}>\) token, generate the correct output followed by an \(<\text{END}>\) token.

- Since the output \(\hat{y}_t\) depends on \(\hat{y}_{t-1}\) we can have very slow training if \(\hat{y}_t\) is inaccurate at the start of the sequence, because the errors compound. To mitigate, during training we can use teacher forcing: use \(y_{t-1}\) (the ground truth) as the input to predict \(\hat{y}_t\)
Given input sequence \((X_1, \ldots, X_T)\) use encoder to obtain \(h_T\).

- Initialize the decoder w/ hidden state \(h_T\) and feed it the \(<\text{START}\>\) token.

- At each time, feed \(J_{t-1}\) into the decoder to predict the next token and take \(J_t\) to be the most likely token.
Practical tips

- use teacher forcing (sometimes)
- use deep LSTMs in the encoder and decoder
- reverse the order of the source sequence to introduce shorter length dependencies w/ the target sequence

Mué y bien y to Very well and you

tu y bien Mué Very well and you

- use "beam decoding" in practice (see Sutskever et al.)
Seq2Seq models w/ attention

Issue: regardless of length of input sequence, all context/historical dependence of the output sequence is contained in a d-dimensional vector \( h_t \). Leads to poor performance for long sequences.

A remedy introduced in Bahdanau et al. 2015, "Neural Machine Translation By Jointly Learning to Align and Translate":

- allow each output token to attend to each individual hidden state in the input sequence. Captures the idea of focusing on the relevant portion of input sequences. Avoids the need to capture everything relevant to all outputs \( \hat{y}_1, \ldots, \hat{y}_T \) in one vector \( h_T \).
**Seq2Seq w/ attention**

replace our previous hidden state update for the decoder

\[ s_{t-1} \rightarrow \text{[blank]} \rightarrow s_t \quad \text{i.e.} \quad s_t = g(s_{t-1}, y_t) \]

w/ first finding which historical hidden states from the input are most relevant

\[ s_{t-1} \rightarrow [h_1, \ldots, h_T] \rightarrow \alpha_t \in \mathbb{R}^T \quad \text{a probability distribution over} \ [h_1, \ldots, h_T] \]

then form a context vector

\[ c_t = \sum_{i=1}^{T} (\alpha_t)_i \cdot h_i \quad \text{and update with} \quad s_t = g(s_{t-1}, c_t) \]
Diagram of our Decoder using Attention

\[ h_T = S_0 \]

\[ c_1 = \sum_{i=1}^{T} (\alpha_{1})_i \cdot h_i \]

\[ c_2 = \sum_{i=1}^{T} (\alpha_{2})_i \cdot h_i \]

\[ c_{T'} = \sum_{i=1}^{T} (\alpha_{T'})_i \cdot h_i \]
Options for computing attention vectors $\alpha_t$

- **Option 1** (used in Bahdanau et al. 2015)

  Take
  $$\alpha_i^2 = \mathbf{v}^T \tanh(\mathbf{W} \begin{bmatrix} h_i \\ s_{t-1} \end{bmatrix})$$
  for $i = 1, \ldots, T$

  parameters to be learned

  $$\alpha_t = \text{softmax}(\alpha^2)$$

  "MLP attention"?

- **Option 2** (Dot product attention)

  Idea: generate keys and queries as linear transforms of $h_i$ and $s_{t-1}$, respectively, and take logits $\alpha_i^2$ to be their inner products

  $$\mathbf{w}_k h_i = k_i$$
  $$\alpha_i^2 = \langle k_i, q_{t-1} \rangle$$

  $$\mathbf{w}_q s_{t-1} = q_{t-1}$$

  $$\alpha_t = \text{softmax}(\alpha^2)$$
Complexity of using attention

- Clear we have introduced parameters

Quadratic dependence on the sequence lengths?

Effort/time to use a RNN seq2seq model:
1) Simple RNN models: $O(T+T')$
2) RNN attention models: $O(TT')$

because for each output token, must compute $\alpha_t \in RT$ by comparing to all inputs, then form a weighted average of all inputs.