01. Word2Vec
02. Attention / Transformers
03. GPT and BERT
04. Simplicity, ALBERT and SHA-RNN
05. Student presentations
06. Student project presentations
07.
Theory of Simplicity

Last two lectures: The bigger the better!

This lecture: The smaller the smaller the better!
LECTURE FOUR
Theory of Simplicity, ALBERT and SHA-RNN
Plan

1. Why simpler?
2. ALBERT
3. SHA-RNN
The Importance of being small

1. Occam’s Razor: Entities should not be multiplied beyond necessity.
2. I. Newton: Nature is pleased simplicity
One example: Inferring a DFA

- Given data that a DFA accepts: 1, 111, 11111, 1111111; and rejects: 11, 1111, 1111111. What is it?

- There are actually infinitely many DFAs satisfying these data.
- The first DFA makes a nontrivial inductive inference,
- The 2nd does not. It “over fits” the data, can’t make further predictions.
Maxwell's (1831-1879)'s equations say that in 1865:

• (a) An oscillating magnetic field gives rise to an oscillating electric field;
• (b) an oscillating electric field gives rise to an oscillating magnetic field.

Item (a) was known from M. Faraday's experiments. However (b) is a theoretical inference by Maxwell and his aesthetic appreciation of simplicity. The existence of such electromagnetic waves was demonstrated by the experiments of H. Hertz in 1888, 8 years after Maxwell's death, and this opened the new field of radio communication. Maxwell's theory is even relativistically invariant. This was long before Einstein’s special relativity. As a matter of fact, it is even likely that Maxwell's theory influenced Einstein’s 1905 paper on relativity which was actually titled ‘On the electrodynamics of moving bodies’.
Bayesian Inference

Bayes Formula:
\[ P(H|D) = \frac{P(D|H)P(H)}{P(D)} \]

Take -log, maximize \( P(H|D) \) becomes minimize:
\[ -\log P(D|H) - \log P(H) \quad \text{(modulo } \log P(D), \text{ constant).} \]

Where, by Shannon-Fano Theorem,
- \( -\log P(D|H) \) is the coding length of \( D \) given \( H \).
- \( -\log P(H) \) is the coding length of model \( H \)

Thus, to maximize the probability is the same as minimizing the model length (and error description length).
PAC Learning theory / Statistical Inference

Given a set of data, if you have a model to fit the data, then the smaller the model is, the more likely it is to be correct. Such a statement can be proved formally, but it is not our focus here.

The key message I wish to deliver is: if you can do the same work with a smaller (neuron network) model, it will be most likely better.
We have studied Transformer, GPT-2, and especially BERT.

Input/output vector size 768 or 1024
We have observed: The bigger model we get better results.

However, we have just learned the theory: the smaller the model is, the more likely we are close to ground truth.

So, what is the problem?

Could it be that “the ground truth model” is large?

But how much data have we used to train a human kid?

At 150 wpm (average speech speed), a human child of 10 year old has heard:

\[ 150 \times 30 \text{ (min)} \times 5 \text{ (hours)} \times 365 \text{ (days)} \times 10 \text{ (years)} \approx 80M \text{ words} \]

BERT used: 2500M words from Wikipedia, and 11,038 books (1000M words) = 3500M words, 43 times of human kid. At least, this shows the ground truth model is not that large. We should search for smaller models. Research project: what is the size of human language model? (Create a pseudo model to test BERT vs SHA-RNN)
ALBERT

1. Separate word embedding size (128 now) from hidden layer vector size (768). This saves $Vx(768-128)$ weights.
2. Cross-layer parameter sharing: including attention part and feed forward network. I.e. all layers share same parameters. Doing this costs 1.5% loss of accuracy, but significantly saved parameters (so that ALBERT can add more layers)
3. Inter-sentence coherence loss: Replace “next sentence” prediction by “sentence order prediction”.

## ALBERT Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>SQuAD1.1</th>
<th>SQuAD2.0</th>
<th>MNLI</th>
<th>SST-2</th>
<th>RACE</th>
<th>Avg</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BERT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>base</td>
<td>108M</td>
<td>90.5/83.3</td>
<td>80.3/77.3</td>
<td>84.1</td>
<td>91.7</td>
<td>68.3</td>
<td>82.1</td>
<td>17.7x</td>
</tr>
<tr>
<td>large</td>
<td>334M</td>
<td>92.4/85.8</td>
<td>83.9/80.8</td>
<td>85.8</td>
<td>92.2</td>
<td>73.8</td>
<td>85.1</td>
<td>3.8x</td>
</tr>
<tr>
<td>xlarge</td>
<td>1270M</td>
<td>86.3/77.9</td>
<td>73.8/70.5</td>
<td>80.5</td>
<td>87.8</td>
<td>39.7</td>
<td>76.7</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>ALBERT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>base</td>
<td>12M</td>
<td>89.3/82.1</td>
<td>79.1/76.1</td>
<td>81.9</td>
<td>89.4</td>
<td>63.5</td>
<td>80.1</td>
<td>21.1x</td>
</tr>
<tr>
<td>large</td>
<td>18M</td>
<td>90.9/84.1</td>
<td>82.1/79.0</td>
<td>83.8</td>
<td>90.6</td>
<td>68.4</td>
<td>82.4</td>
<td>6.5x</td>
</tr>
<tr>
<td>xlarge</td>
<td>59M</td>
<td>93.0/86.5</td>
<td>85.9/83.1</td>
<td>85.4</td>
<td>91.9</td>
<td>73.9</td>
<td>85.5</td>
<td>2.4x</td>
</tr>
<tr>
<td>xxlarge</td>
<td>233M</td>
<td><strong>94.1/88.3</strong></td>
<td><strong>88.1/85.1</strong></td>
<td><strong>88.0</strong></td>
<td><strong>95.2</strong></td>
<td><strong>82.3</strong></td>
<td><strong>88.7</strong></td>
<td><strong>1.2x</strong></td>
</tr>
</tbody>
</table>
1. Author’s motivation: Alternative route of research?
2. My motivation: We should always look for simplicity.
3. Here, we go back to the old approach of LSTM to raise sufficient doubt that Transformer is the only way.
The core idea is this cell state $C_t$, it is changed slowly, with only minor linear interactions. It is very easy for information to flow along it unchanged.

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t$$
Theory of Simplicity

Bits Per Character (BPC)

1. BPC is average cross-entropy
2. T is length of string. \( P_t \) is true distribution. \( n \) is (character) alphabet size. \( P_{i}(x_{j})=1 \) iff \( i=j \).

\[
bpc\text{(string)} = \frac{1}{T} \sum_{t=1}^{T} H(P_t, \hat{P}_t) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{c=1}^{n} P_t(c) \log_2 \hat{P}_t(c),
\]

\[
= -\frac{1}{T} \sum_{t=1}^{T} \log_2 \hat{P}_t(x_t).
\]
Table 1. Bits Per Character (BPC) on Enwik8. The single attention SHA-LSTM has an attention head on the second last layer and had batch size 16 due to lower memory use. Directly comparing the head count for LSTM models and Transformer models obviously doesn’t make sense but neither does comparing zero-headed LSTMs against bajillion headed models and then declaring an entire species dead. The hyper-parameters for the fully headed SHA-LSTM were used for the other SHA-LSTM experiments with zero tuning.
1. The Boom layer is a fully connected NN that maps \( v \) of size \( H \) to \( u \) of size \( N \times H \), then break \( u \) into \( N=4 \) vectors and sum them together to produce \( w \) of size \( H \) (1024).

2. The Attention mechanism is very similar to Attention we learned in Lecture 2, except softmax is replaced by sigmoid.
Deep RNN

\[ h', y = f_1(h, x), \quad g', z = f_2(g, y) \]
Theory of Simplicity

4 layer SHA RNN

1. Each layer gives the next layer sequential output.
2. Name: Stacked LSTM or Deep LSTM.

* Single Attention Layer
A theory of simplicity

Literature & Resources


S. Merity, Single hearded Attention RNN: Stop thinking with your head, 2020

Can somebody present Sukhbaatar et al 2019 Adaptive Transformer?

Can somebody present the Sparse Transformer by Child et, 2019.