Lecture 12: Course Review

Andrei Arsene Simion

Columbia University

April 19, 2023

Outline

- Course Review

Outline

- Course Review

Word Embeddings

- Word2Vec
  - SkipGram vs CBOW: What's the difference?
  - Negative Sampling: Why is it important?
- GloVe: How is it different than Word2Vec?
  - What is an alternate formulation of Word2Vec that makes these 2 problems similar?
- FastText: What is the difference between FastText and the above two models?
- What is the universe of tokens for each of the models above?
- How is UNK (unknown) handled?

Neural Networks

- What loss is typically used for Classification?
- For Regression?
- Can you do back propagation by hand? If I give you a small network, can you draw out the computation graph?
- Why is ReLU much more popular (now) than Sigmoid or Tanh?
- What is BatchNorm? What is the role of BatchNorm?
- What is LayerNorm? What is the role of LayerNorm?
- What is a residual connection? What is the role of this?
- What is DropOut? What other methods do you know?
- What is the difference between SGD and Adam?
- What happens to SGD if the step size if too large? What if too small?

CNN

- How are CNN's used in NLP?
  - How can you use CNN for text classification?
  - How are CNN models used to get word embeddings?
  - Do you know how do compute the result of a filter applied to a sentence?
  - What if there is a stride?
  - What if there is padding?
- What are Gated CNN models? What were they used for?
- What is the Connection between CNN2D and CNN1D in PyTorch?

RNN

- What are the recursions of an RNN? What are the equations?
- What is the problem with RNN models?
- What is greedy decoding?
- How does Beam Search work?
- What is the benefit of a RNN vs a Bag of Words model?

LSTM

- What is an LSTM?
- What are the recursions of a LSTM?
- What is a GRU?
- What is the difference between an LSTM and a GRU / RNN?
- How does an LSTM try and solve the problems present in an RNN?

Attention

- What is Attention and how is it applied to an LSTM?
- How can we do SMT with an LSTM? What is the full architecture?
- How do we apply Attention to the SMT problem? What is the difference between this and the above?
- What is the main metric for SMT? Can you identify its components and do you know the formula for it?

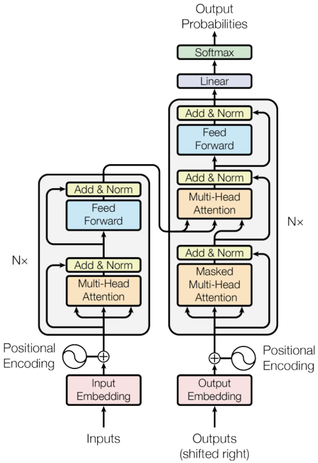
ELMo, UML-Fit

- What is ELMo?
- What is the difference between ELMo and Word2Vec?
- How is ELMo fit? How can you use it?
- What is CoVE?
- What is UML-Fit?
- What is BiDaf?
- What is the SQUAD dataset?

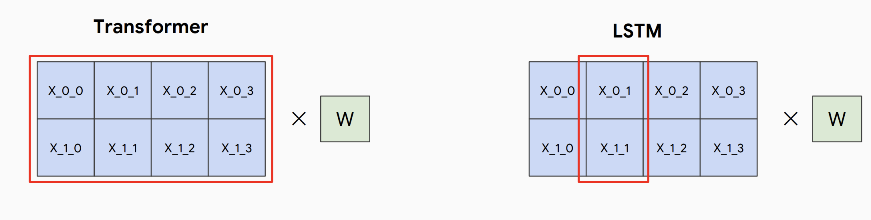
Transformer

- Why are transformers faster than LSTM models?
- Can you compute Self-Attention for a basic example?
- What is Masked Self-Attention? When is it used?
- What is Multi-Head Attention?
- Does the Transformer have Feed Forward Layers? What is the idea of these layers?
- What types of embeddings does a Transformer have?

Transformer



Transformer vs LSTM



Transformer

X_0_0	X_0_1	X_0_2	X_0_3
X_1_0	X_1_1	X_1_2	X_1_3

× W

LSTM

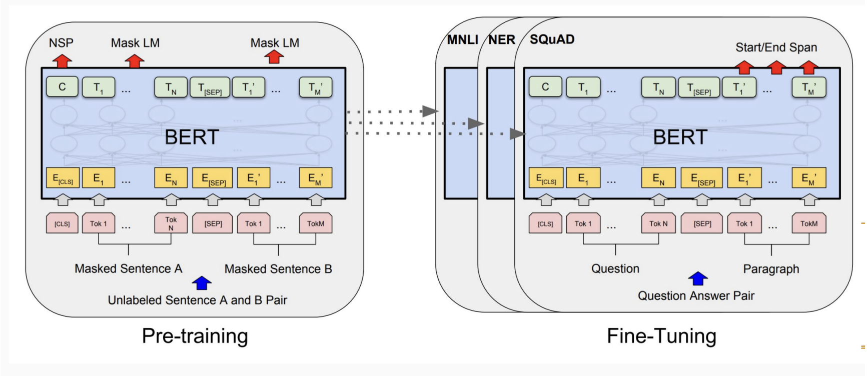
X_0_0	X_0_1	X_0_2	X_0_3
X_1_0	X_1_1	X_1_2	X_1_3

× W

BERT

- Can you describe how BERT is trained?
- What types of embeddings does BERT use? What types of tokens?
- Was BERT over or under trained? Why?
- Can you use BERT to generate text?
- What type of Transformer is BERT?
- What is the difference between BERT and ELMo?
- How do you use BERT for text classification?
- How do you use BERT for NER? For Question-Answering? For Tagging?
- What is DistilBERT? Why is it important?

BERT


NSP
Mask LM
Mask LM
C
T_1
...
T_N
$T_{[SEP]}$
T_1'
...
T_M'
BERT
$E_{[CLS]}$
E_1
...
E_N
$E_{[SEP]}$
E_1'
...
E_M'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
TokM
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
Pre-training
MNLI
NER
SQuAD
Start/End Span
C
T_1
...
T_N
$T_{[SEP]}$
T_1'
...
T_M'
BERT
$E_{[CLS]}$
E_1
...
E_N
$E_{[SEP]}$
E_1'
...
E_M'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
TokM
Question
Paragraph
Question Answer Pair
Fine-Tuning

GPT

- Can you describe how GPT is trained?
- What is the main difference between the GPT1 and GPT2 models + paper results?
- What is the main difference between the GPT2 and GPT3 models + paper results?
- Can you use GPT for text classification? How? Would you expect it to do better or worse than BERT?

Other Transformers

- What is BART?
- What is ALBERT? What are the key differences between this model and BERT?
- What is SentenceBERT? Why is it necessary?
- What is ROBERTA? What was the key result of ROBERTA?
- SpanBERT?
- Can you describe the objectives of the above and how the input x is corrupted?

Other topics

- Why is Knowledge Distillation important?
- What is the objective in Knowledge Distillation?
- What is temperature scaling? What is the effect of various T on different distribution functions?
- What is Top-K Sampling?
- What is Nucleus Sampling?
- How do you generally say a language model has generated "good" text? Lower perplexity?

References

- See Lectures 1 - Lecture 11