

# Neural Language Models

Computational Linguistics: Jordan Boyd-Graber

University of Maryland

Neural Models

Adapted from material by Anna Rogers, Jacob Devlin, and Richard Socher

# Maryland and Muppets

- Kermit (Jim, BS 1960),  
1955
- ELMO (Mohit, PhD 2019)
- BERT (Jacob, MS 2009)



# The power of neural language models

- Not just for predicting words
- Representation is important!

# The power of neural language models

- Not just for predicting words
- Representation is important!
- **Fine tuning**



# Why Muppets?

## Deep contextualized word representations

**Matthew E. Peters<sup>†</sup>, Mark Neumann<sup>†</sup>, Mohit Iyyer<sup>†</sup>, Matt Gardner<sup>†</sup>,**

{matthewp, markn, mohiti, mattg}@allenai.org

**Christopher Clark\*, Kenton Lee\*, Luke Zettlemoyer<sup>†\*</sup>**

{csquared, kentonl, lsz}@cs.washington.edu

<sup>†</sup>Allen Institute for Artificial Intelligence

\*Paul G. Allen School of Computer Science & Engineering, University of Washington

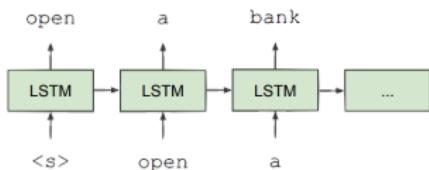
### Abstract

We introduce a new type of *deep contextualized* word representation that models both (1) complex characteristics of word use (e.g., syn-

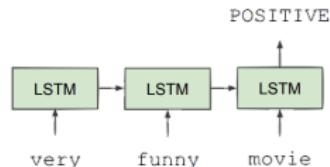
guage model (LM) objective on a large text corpus. For this reason, we call them ELMo (Embeddings from Language Models) representations. Unlike previous approaches for learning contextu-

# Fine tuning

## Train LSTM Language Model



## Fine-tune on Classification Task



From *Semi-supervised Sequence Learning* by Dai and Le, 2015

## Why does this work?

- Language models “fill in the blank” and learn representations to do that
- Other tasks can often be transformed implicitly or explicitly into fill in the blank tasks

## Why does this work?

- Language models “fill in the blank” and learn representations to do that
- Other tasks can often be transformed implicitly or explicitly into fill in the blank tasks
  - ▶ Sentiment: “The burrito made me sick” (so I think it’s good/bad)
  - ▶ Entailment: “John married Lisa” (thus) “Lisa is John’s wife”
  - ▶ Question Answering: “The first president of the United States was \_\_\_\_\_”

## Why does this work?

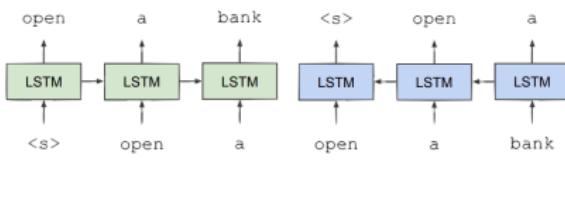
- Language models “fill in the blank” and learn representations to do that
- Other tasks can often be transformed implicitly or explicitly into fill in the blank tasks
  - ▶ Sentiment: “The burrito made me sick” (so I think it’s good/bad)
  - ▶ Entailment: “John married Lisa” (thus) “Lisa is John’s wife”
  - ▶ Question Answering: “The first president of the United States was \_\_\_\_\_”
- Other tasks not so obvious, but still seems to work!

## Where are the innovations?

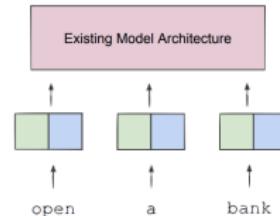
- Bidirectional (ELMO)
- Attention and Objective Tweaks (Transformers)
- Training objectives (BERT)
- Sequence encoding (Transformer + BERT)

# Bidirectional (ELMO)

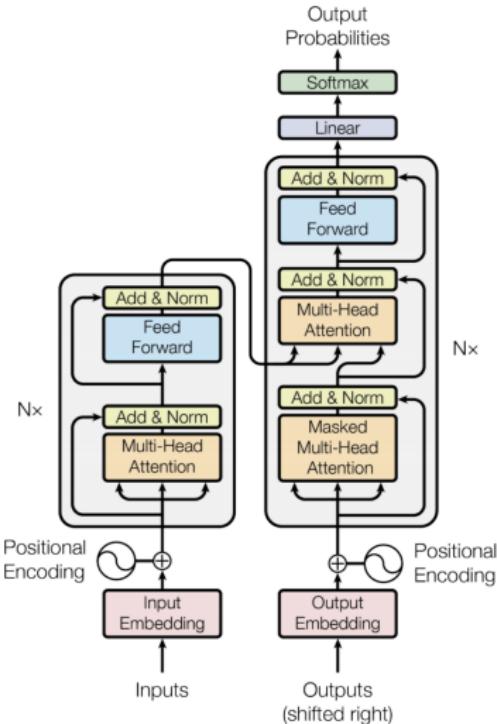
**Train Separate Left-to-Right and Right-to-Left LMs**



**Apply as “Pre-trained Embeddings”**

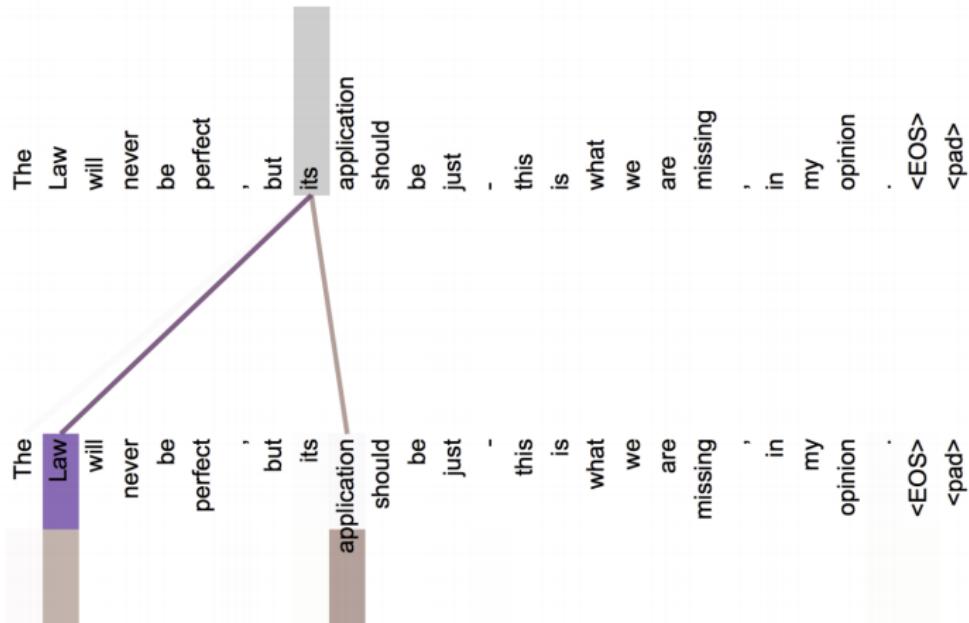


# Attention (Transformers)

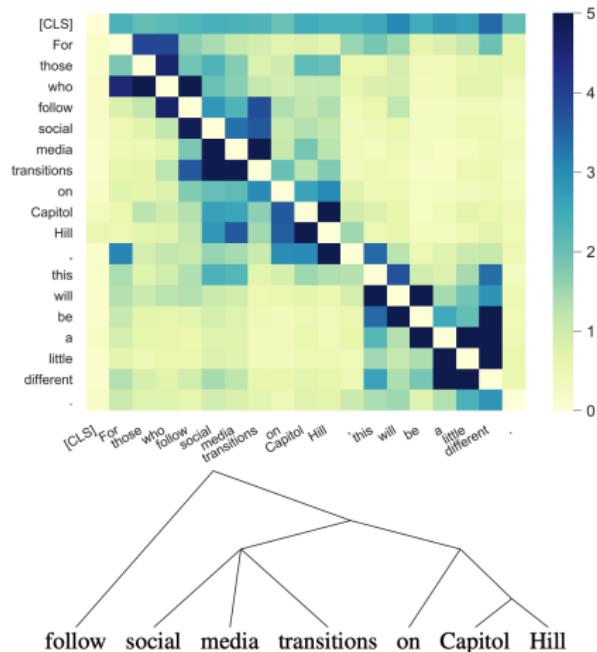


- Attention lets one word affect any other word
- BERT is stack of Transformer (Vaswani et al., 2017) encoders with multiple attention heads
  - ▶ Head computes key, value, and query vectors
  - ▶ Create weighted representation
  - ▶ All outputs in layer goes into fully-connected layer

# Attention is task specific



# Attention is task specific



## Training Objectives (BERT)

## Masked Word (Pieces)

the man went to the [MASK] to buy a [MASK] of milk

↑                      ↑  
store                  gallon

## Next Sentence Prediction

**Sentence A** = The man went to the store.

**Sentence B** = He bought a gallon of milk.

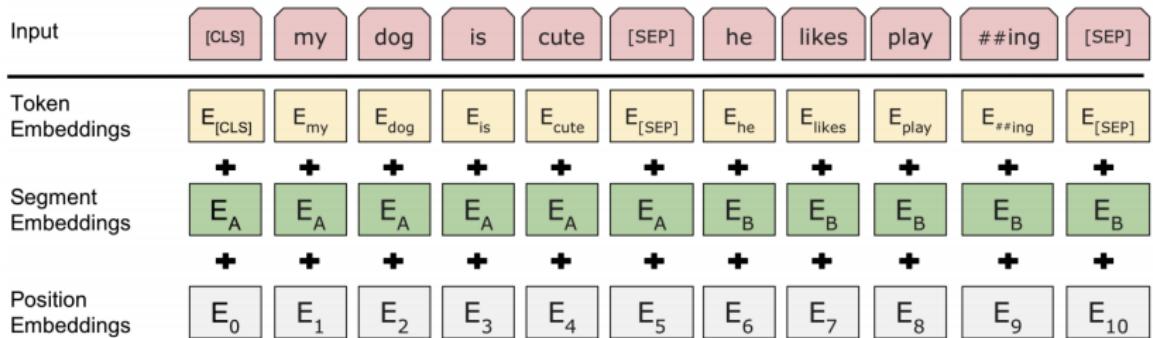
**Label** = IsNextSentence

**Sentence A** = The man went to the store.

**Sentence B =** Penguins are flightless.

**Label** = NotNextSentence

# Encoding (BERT)



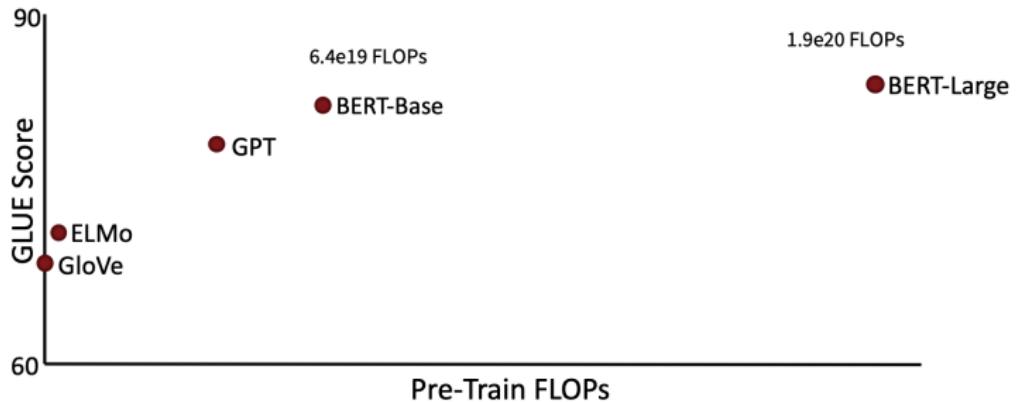
## What's not to love?

- If you're not implementing, no more difficult than RNN/LSTM
- Much higher accuracies

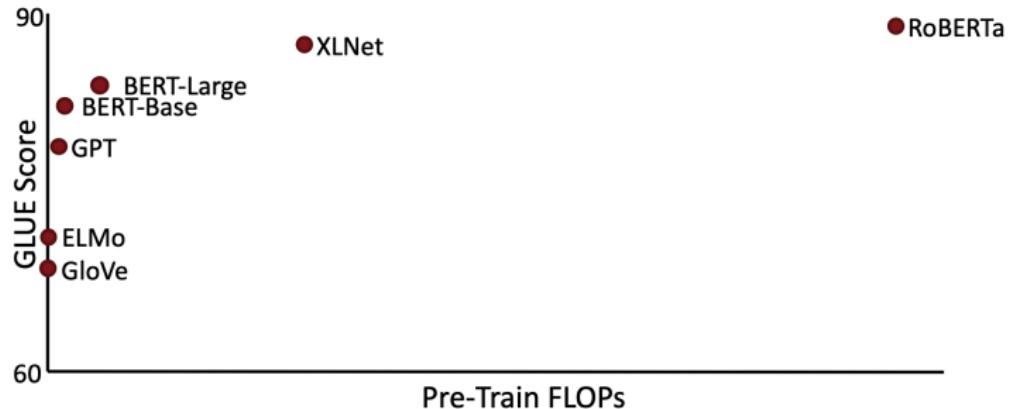
## What's not to love?

- If you're not implementing, no more difficult than RNN/LSTM
- Much higher accuracies
- Complicated!
  - ▶ Hard to understand what's going on
  - ▶ Expensive compute

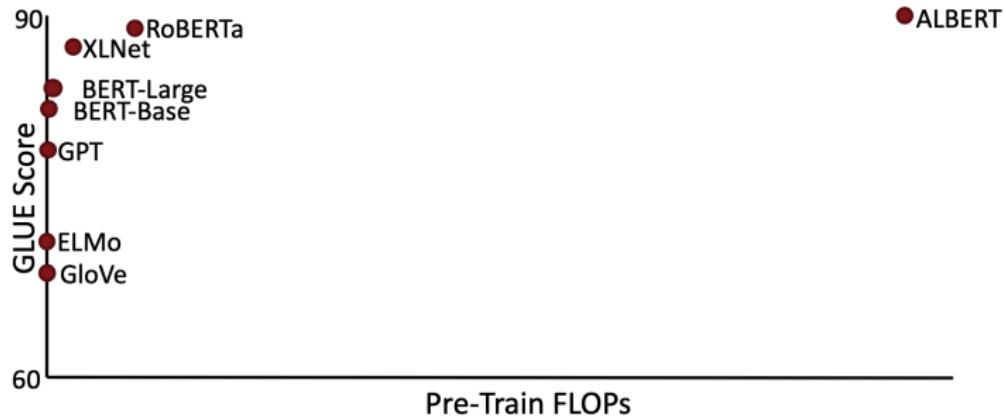
## Computational (climate) cost



## Computational (climate) cost



## Computational (climate) cost



## Not a panacea

- You still need to understand the data!
- Basic problems can (and should) be resolved with logistic regression
- BERT is so good it can hid your mistakes