Neural Language Models

Computational Linguistics: Jordan Boyd-Graber

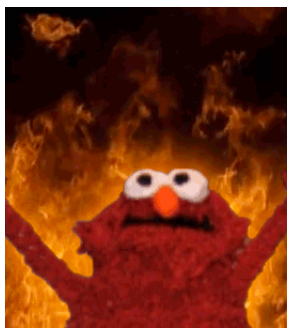
University of Maryland

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 imporant!

The power of neural language models

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

Why Muppets?

Deep contextualized word representations

Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

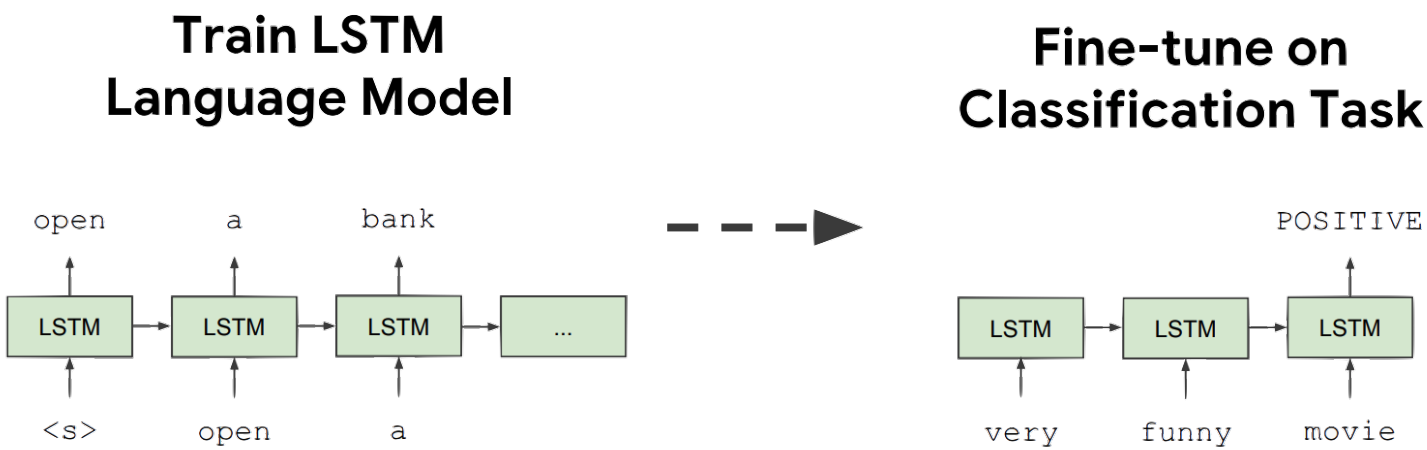
Abstract

We introduce a new type of *deep contextualized* word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). Our word vectors are learned functions of the internal states of a deep bidirectional language model (LM), which is pre-trained on a large text corpus. We show that these representations can be easily added to existing models and significantly improve the state of the art across six challenging NLP problems, including question answering, textual entailment and sentiment analysis. We also present an analysis showing that exposing the deep internals of the pre-trained network is crucial, allowing downstream models to mix different types of semi-supervision signals.

1 Introduction

Pre-trained word representations are a key component in many neural language understanding models. However, learning high quality representations can be challenging. They should ideally model both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). Our approach also benefits from scale, learning rich representations from a neural language 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



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?

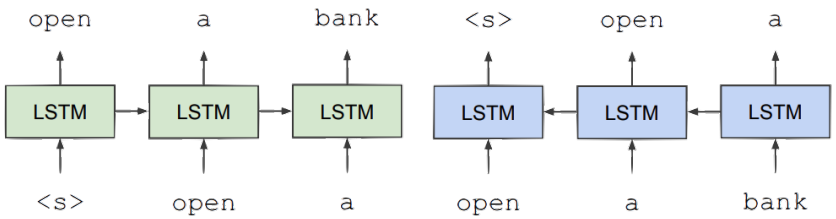
- Language models “fill in the blank” and learn representations to do that
- Other tasks can often be transformed implicitly or explicitly into fill in the blank tasks
  - Sentiment: “The burrito made me sick” (so I think it’s good/bad)
  - Entailment: “John married Lisa” (thus) “Lisa is John’s wife”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

Where are the innovations?

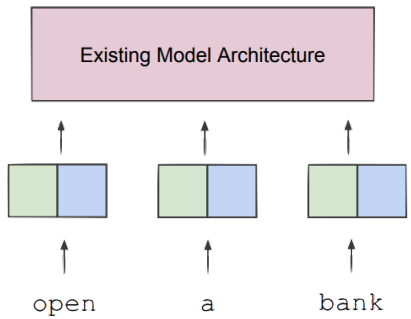
- Bidirectional (ELMO)
- Attention and Objective Tweaks (Transformers)
- Training objectives (BERT)
- Sequence encoding (Transformer + BERT)

Bidirectional (ELMO)

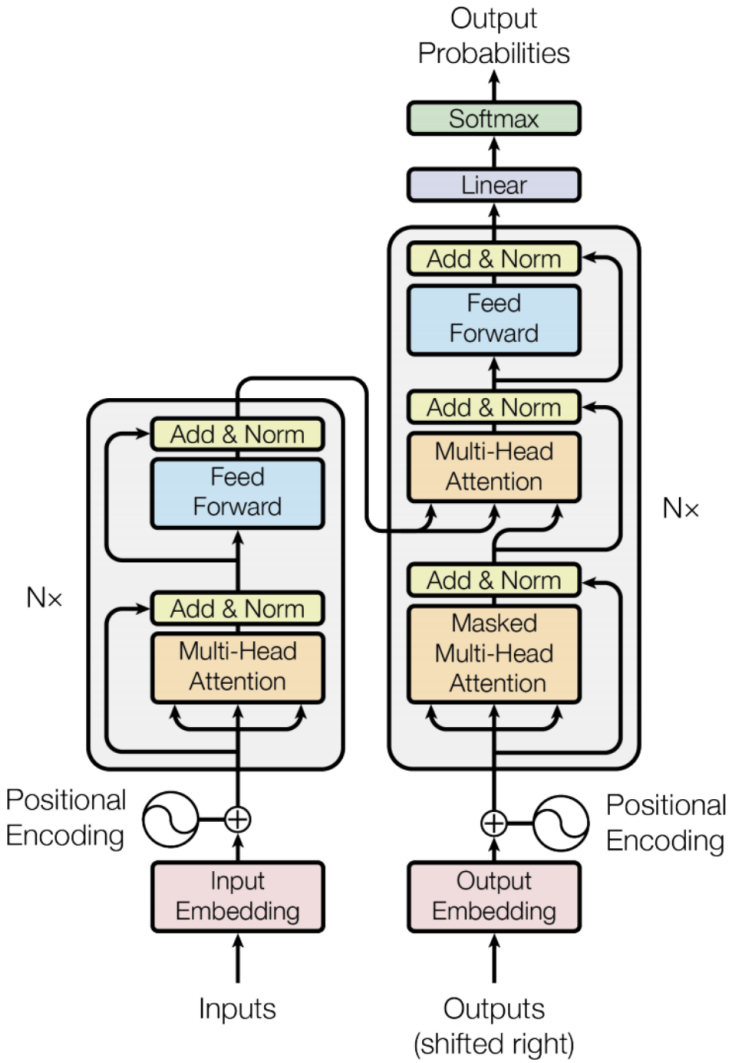
Train Separate Left-to-Right and Right-to-Left LMs



Apply as "Pre-trained Embeddings"

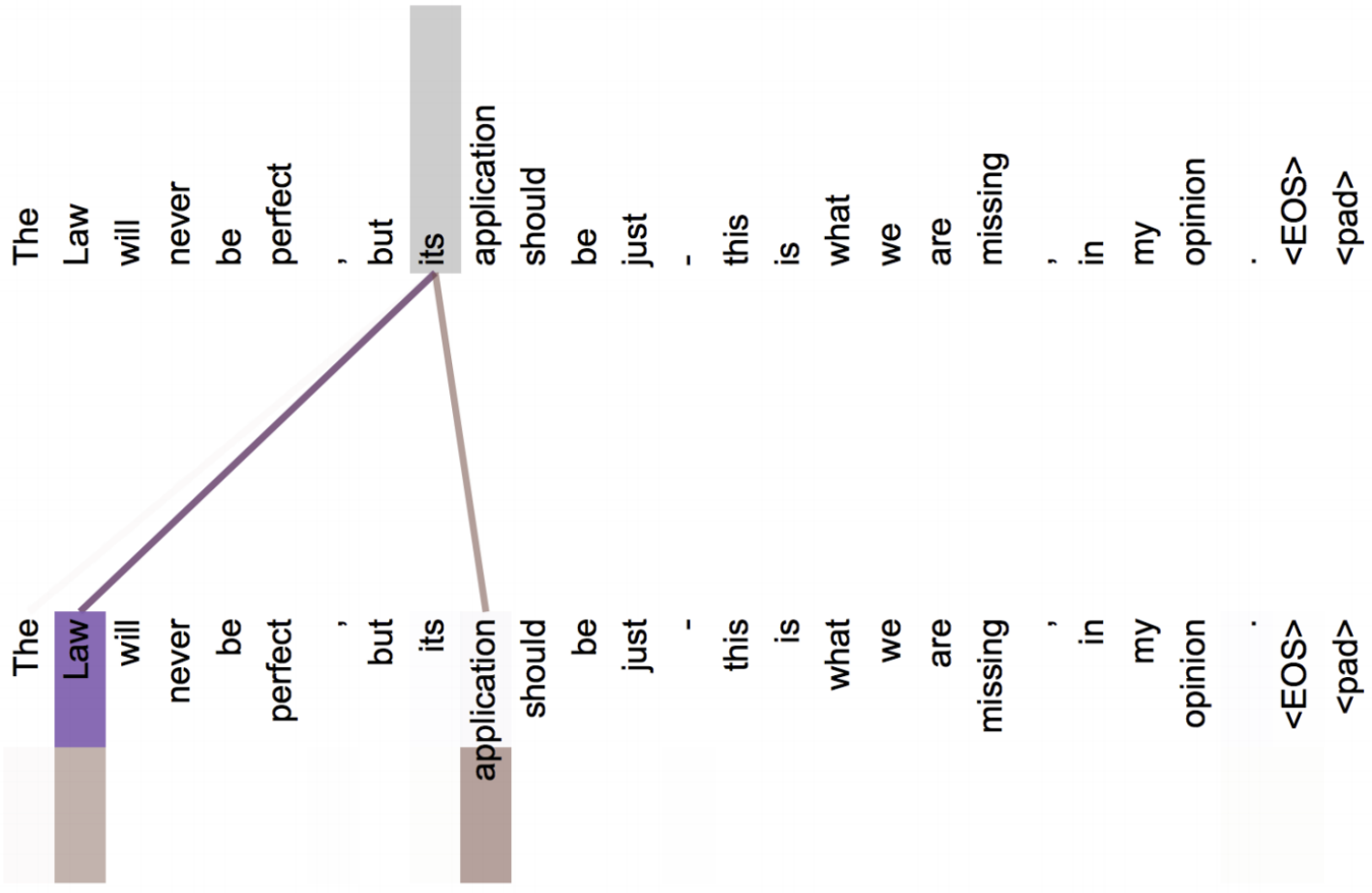


Attention (Transformers)

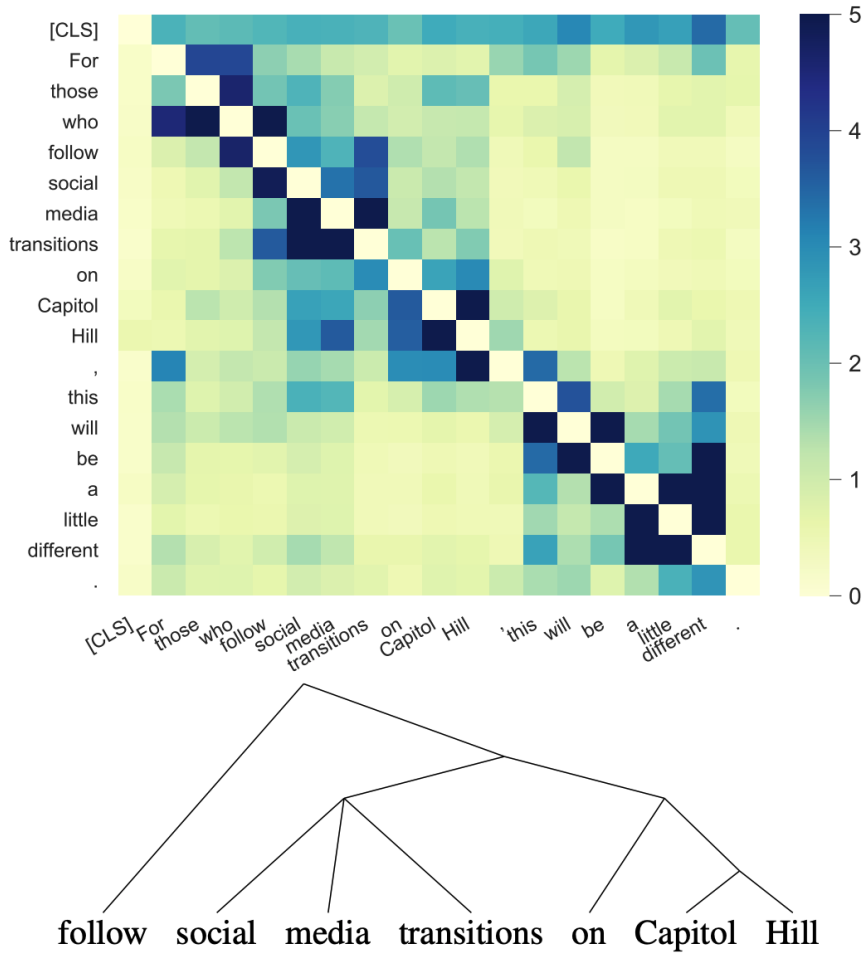


- Attention lets one word affect any other word
- BERT is stack of Transformer (Vaswani et al., 2017) encoders with multiple attention heads
  - Head computes key, value, and query vectors
  - Create weighted representation
  - All outputs in layer goes into fully-connected layer

Attention is task specific

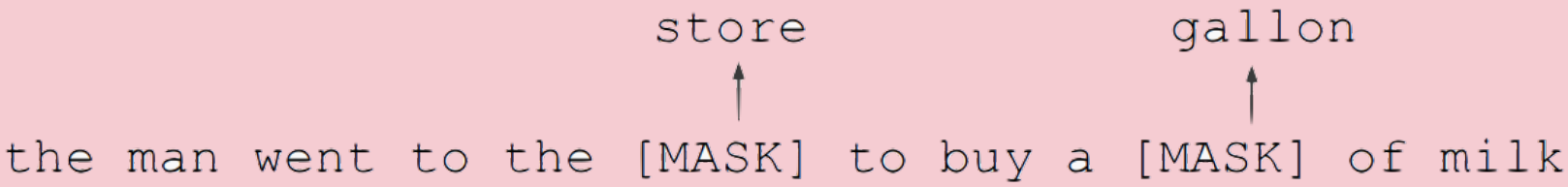


Attention is task specific



Training Objectives (BERT)

Masked Word (Pieces)

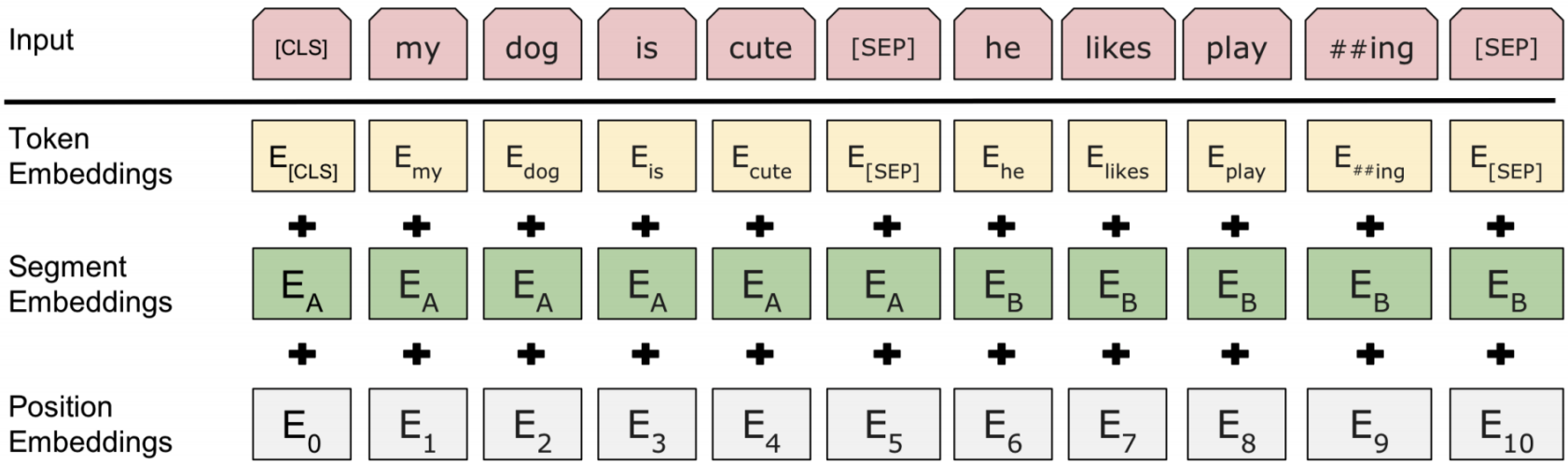


Next Sentence Prediction

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)



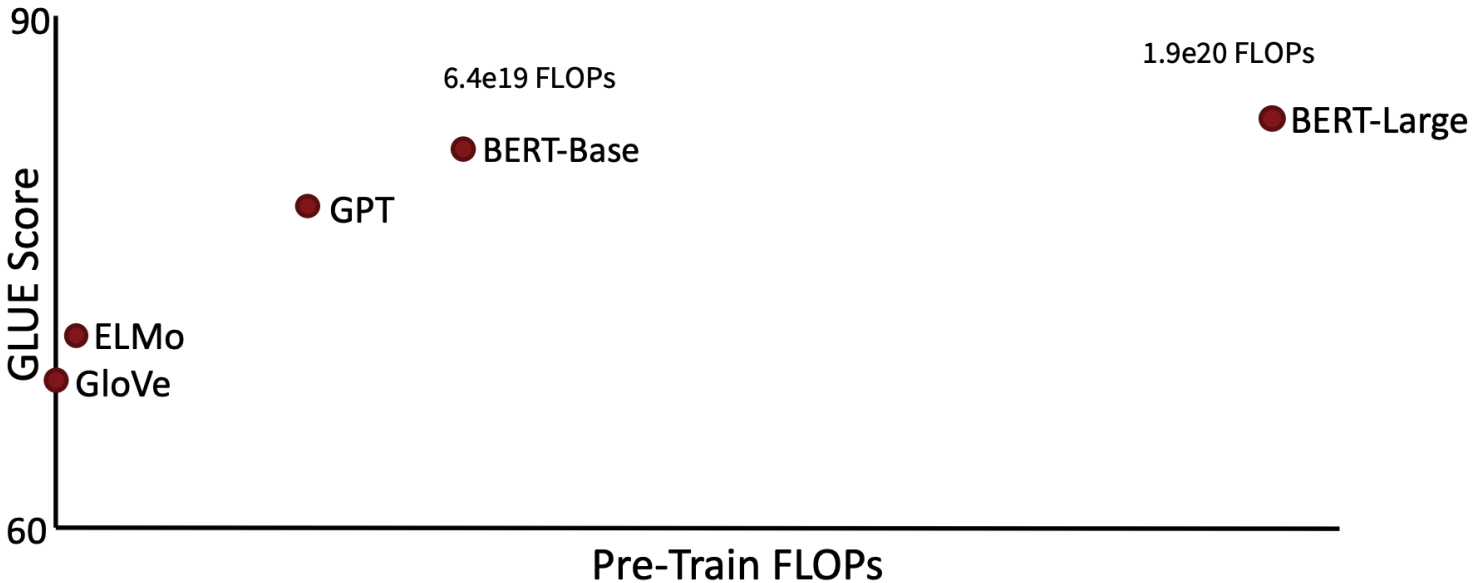
What's not to love?

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

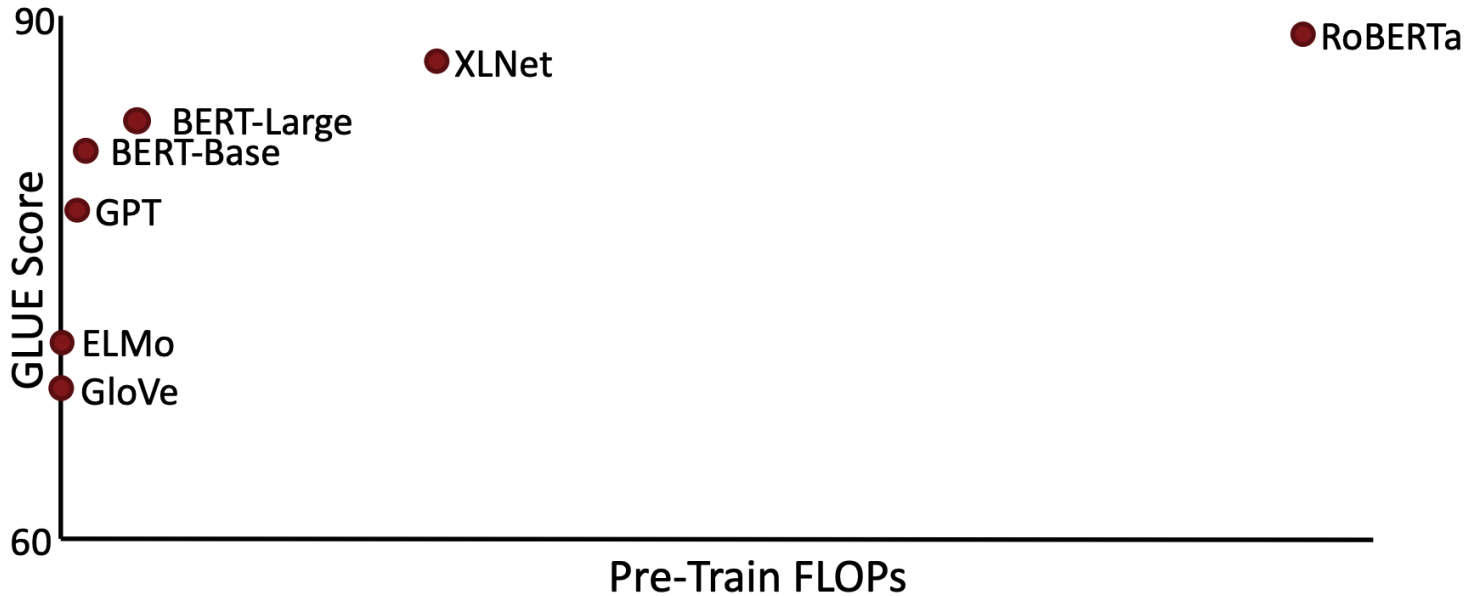
What's not to love?

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

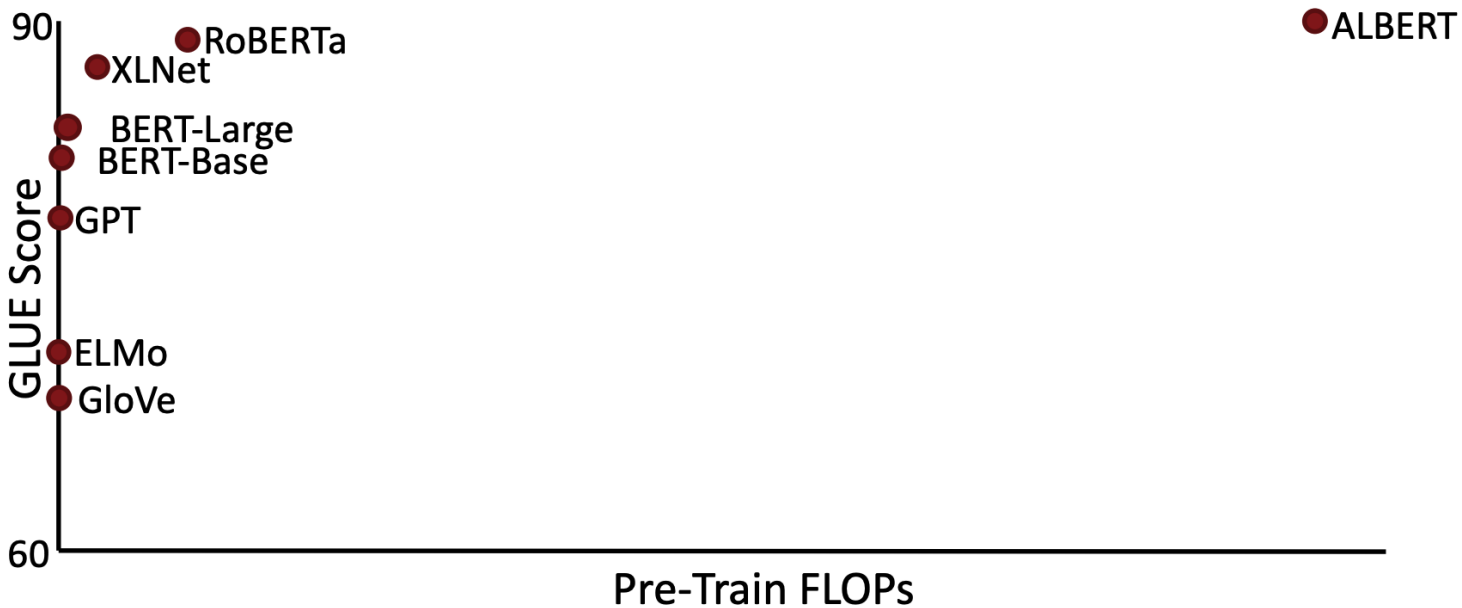
Computational (climate) cost



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