Using BERT in downstream tasks

CS 585, Fall 2019

Mohit Iyyer
College of Information and Computer Sciences
University of Massachusetts Amherst

many slides from Jacob Devlin

stuff from last time

- slides ahead of time?
- can you repeat questions asked during class?
- project??? (milestone 1 due Oct 24)
- HW2??? (due Oct 18)
- midterm??? (Oct 31)

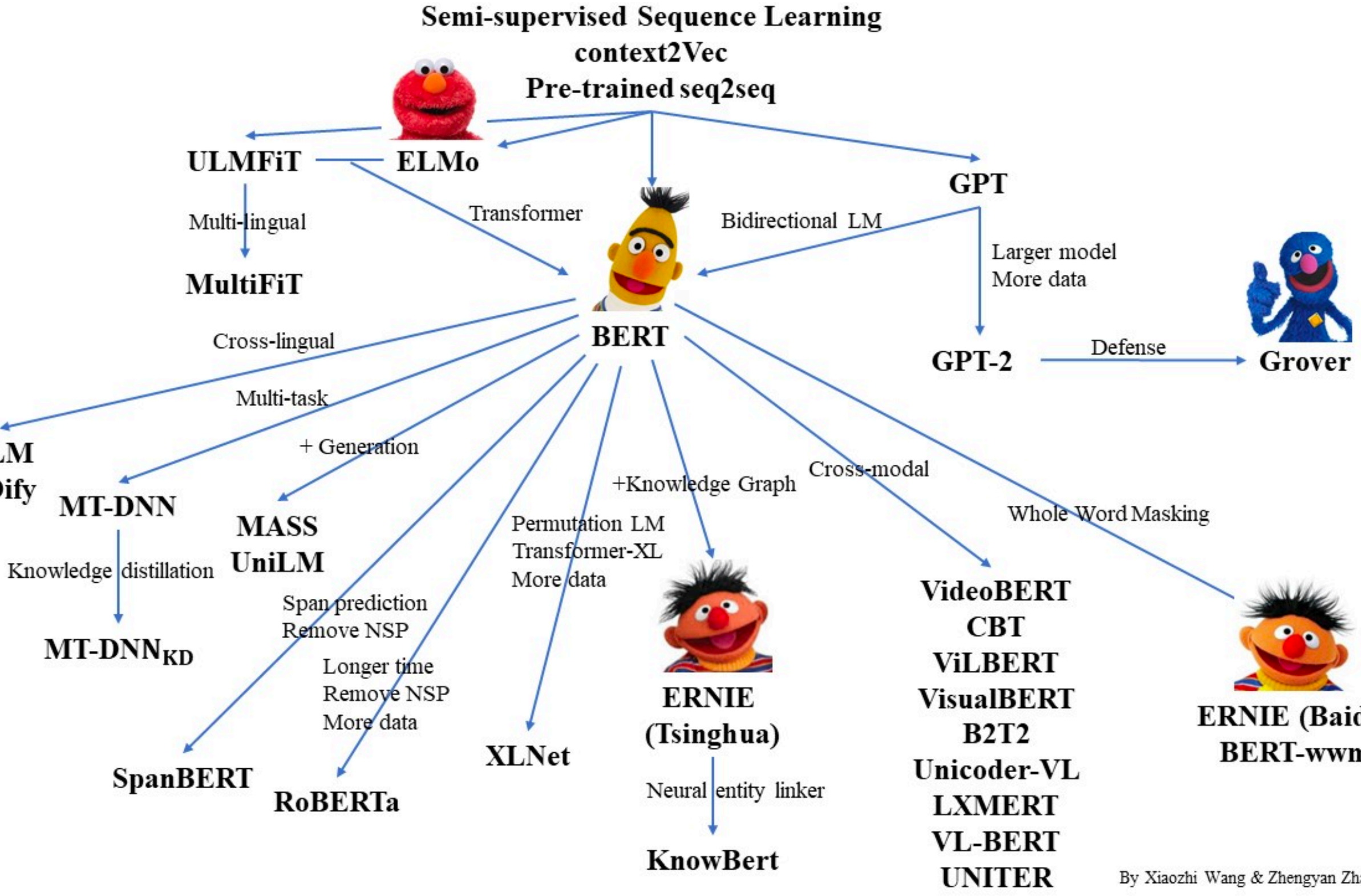
rough details about midterm

- two practice exams on Piazza (solutions soon)
  - many problems are on topics we haven't covered so don't worry about knowing everything

- **topics that can definitely be on the midterm:**
  - naive Bayes
  - ngram LMs + smoothing
  - word embeddings (e.g., word2vec)
  - neural networks (e.g., backprop, training setup)
  - fixed-window / RNN LMs
  - Transformers
  - Transfer learning (e.g., ELMo / BERT)
  - sequence labeling (next week)

FROM

TO



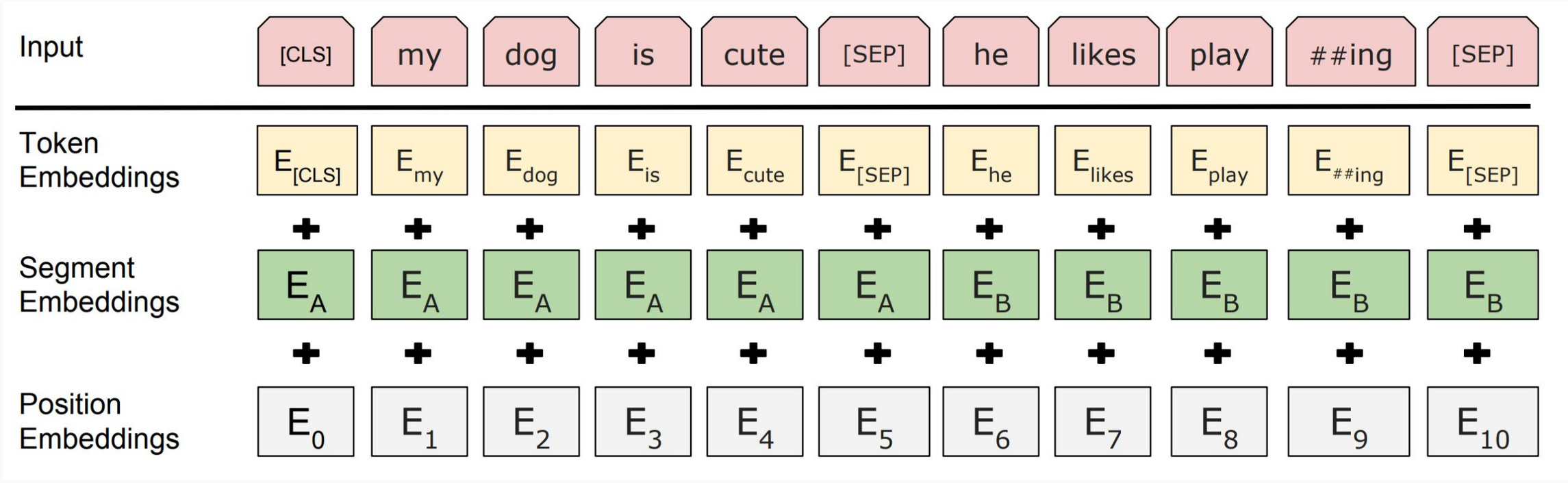
source: https://github.com/thunlp/PLMpapers

Masked LM

- **Solution**: Mask out *k*% of the input words, and then predict the masked words
  - We always use *k* = 15%

What are the pros and cons of increasing *k*?

Input Representation

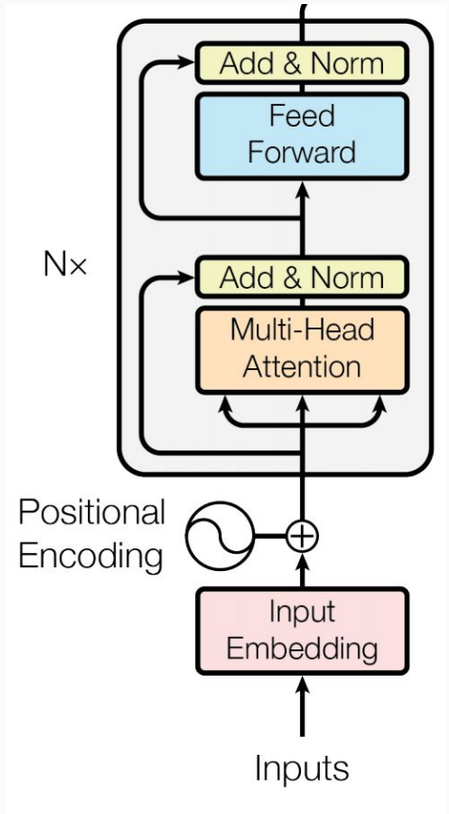


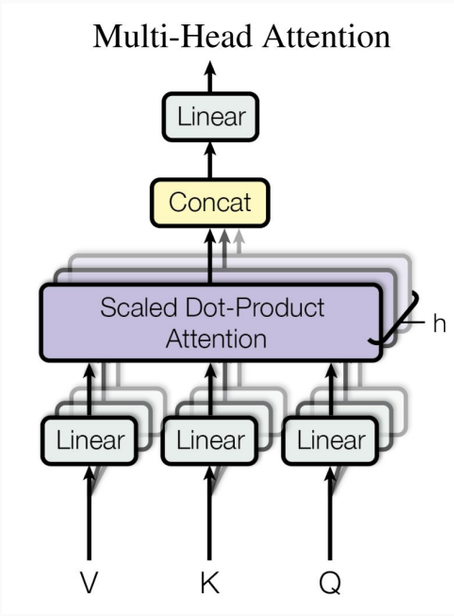
- Use 30,000 WordPiece vocabulary on input.
- Each token is sum of three embeddings
- Single sequence is much more efficient.

Model Architecture

Transformer encoder

- Multi-headed self attention
  - Models context
- Feed-forward layers
  - Computes non-linear hierarchical features
- Layer norm and residuals
  - Makes training deep networks healthy
- Positional embeddings
  - Allows model to learn relative positioning





Model Architecture

- Empirical advantages of Transformer vs. LSTM:

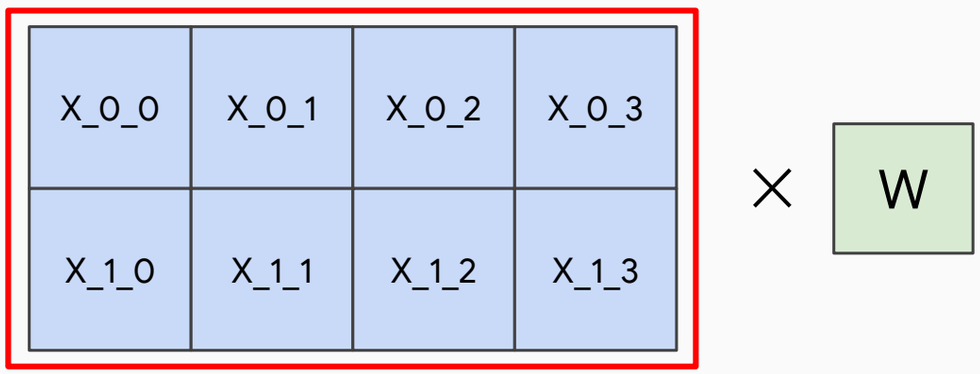
What are they?

Model Architecture

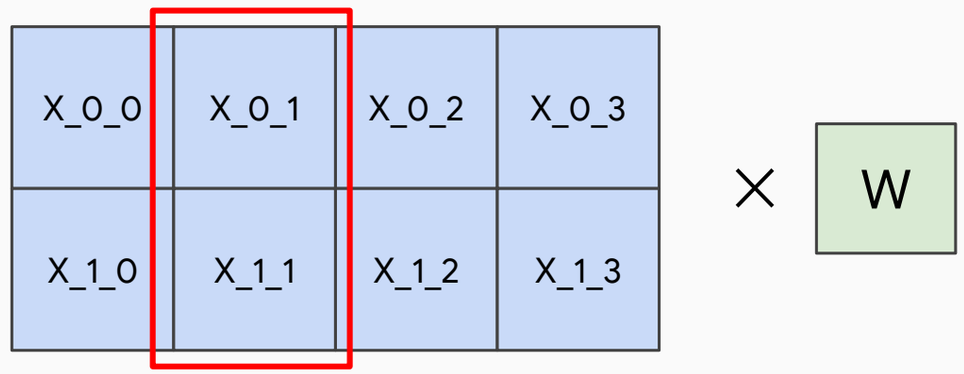
- Empirical advantages of Transformer vs. LSTM:

1. Self-attention == no locality bias
   - Long-distance context has “equal opportunity”
2. Single multiplication per layer == efficiency on TPU
   - Effective batch size is number of *words*, not *sequences*

Transformer



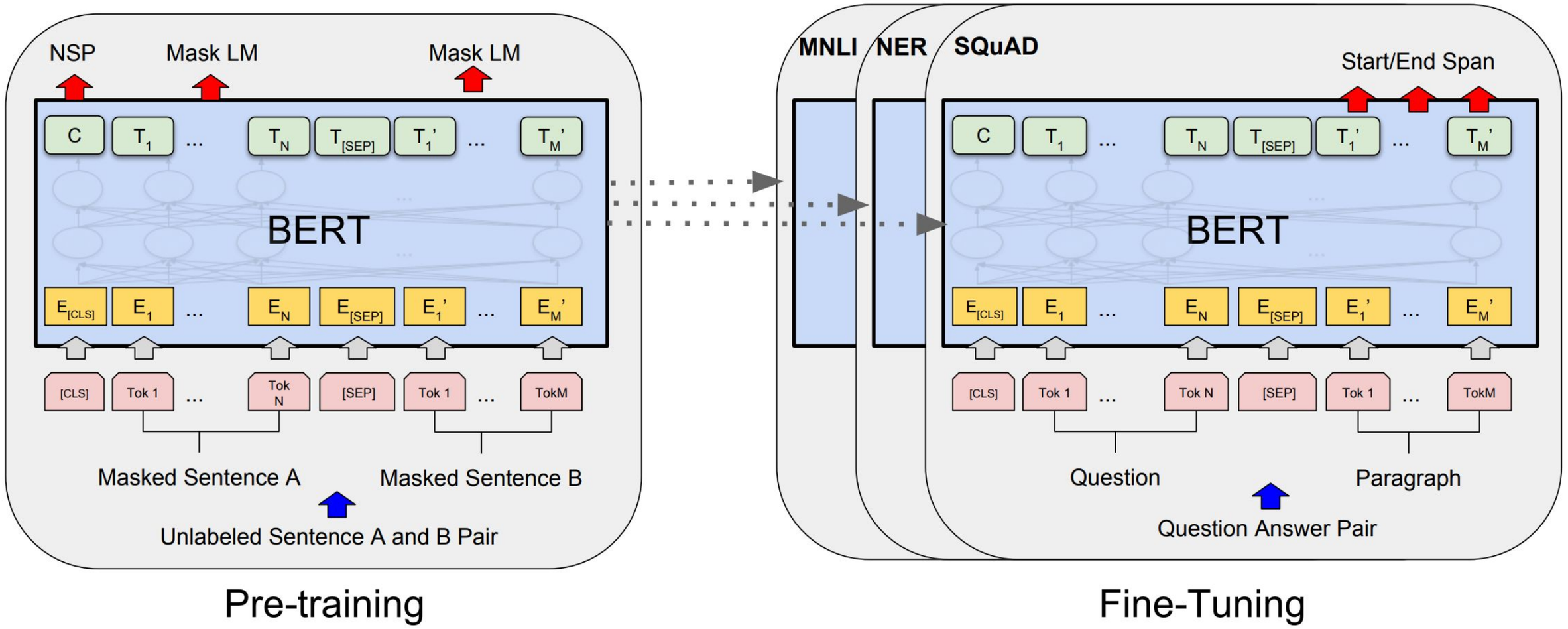
LSTM



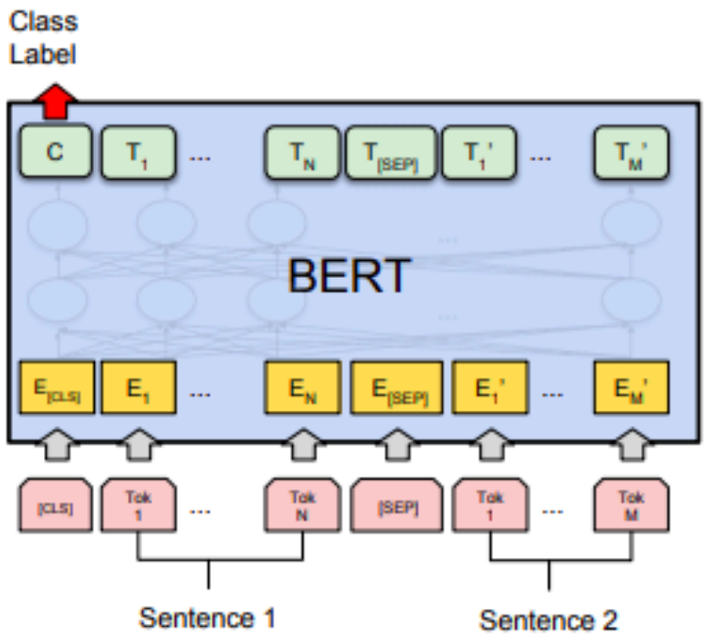
Model Details

- Data: Wikipedia (2.5B words) + BookCorpus (800M words)
- Batch Size: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
- Training Time: 1M steps (~40 epochs)
- Optimizer: AdamW, 1e-4 learning rate, linear decay
- `BERT-Base`: 12-layer, 768-hidden, 12-head
- `BERT-Large`: 24-layer, 1024-hidden, 16-head
- Trained on 4x4 or 8x8 TPU slice for 4 days

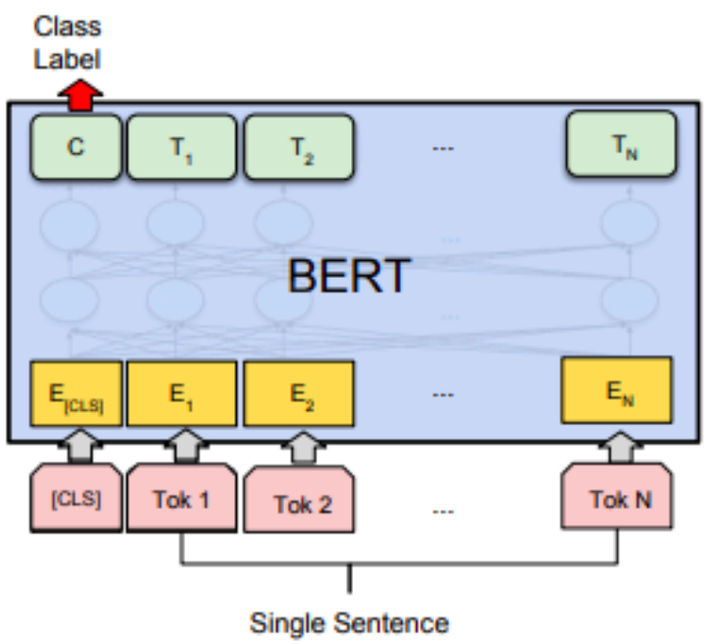
Fine-Tuning Procedure



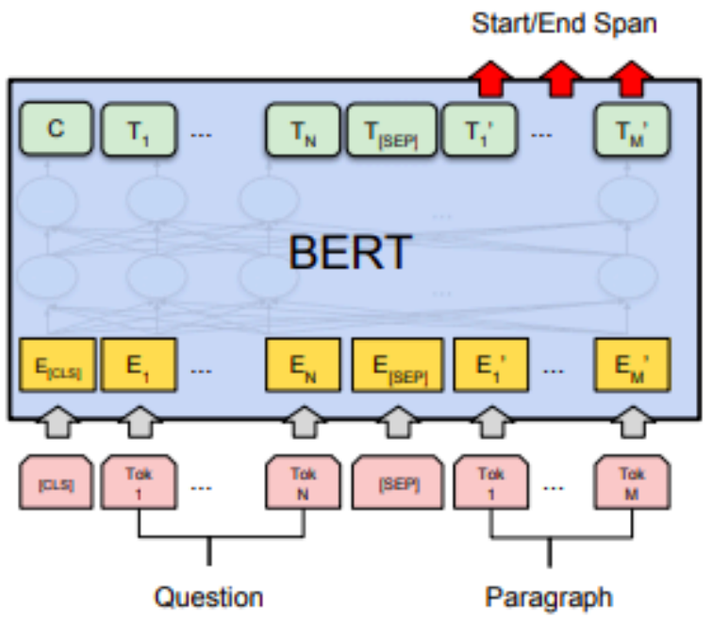
Fine-Tuning Procedure



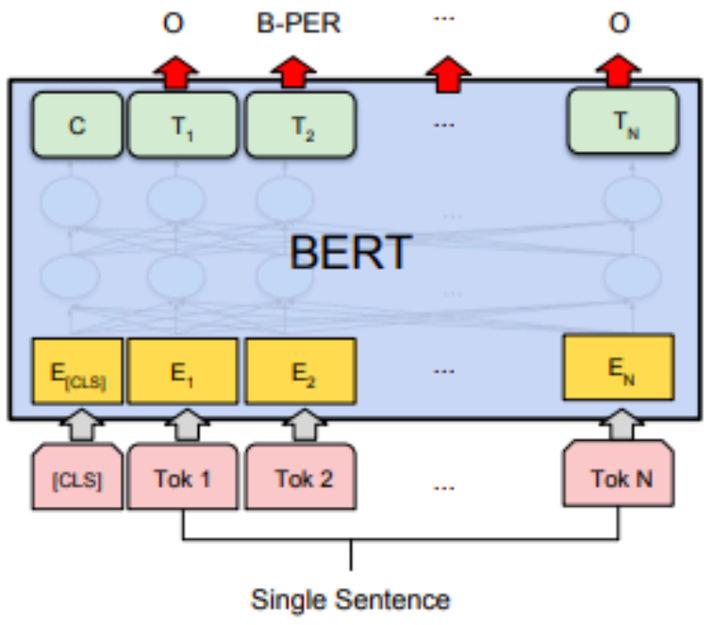
(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG



(b) Single Sentence Classification Tasks:
SST-2, CoLA



(c) Question Answering Tasks:
SQuAD v1.1



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

GLUE Results

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

MultiNLI

Premise: Hills and mountains are especially sanctified in Jainism.
Hypothesis: Jainism hates nature.
Label: Contradiction

CoLa

Sentence: The wagon rumbled down the road.
Label: Acceptable

Sentence: The car honked down the road.
Label: Unacceptable

SWAG

A girl is going across a set of monkey bars. She

(i) jumps up across the monkey bars.

(ii) struggles onto the bars to grab her head.

(iii) gets to the end and stands on a wooden plank.

(iv) jumps up and does a back flip.

- Run each Premise + Ending through BERT.
- Produce logit for each pair on token 0 (`[CLS]`)

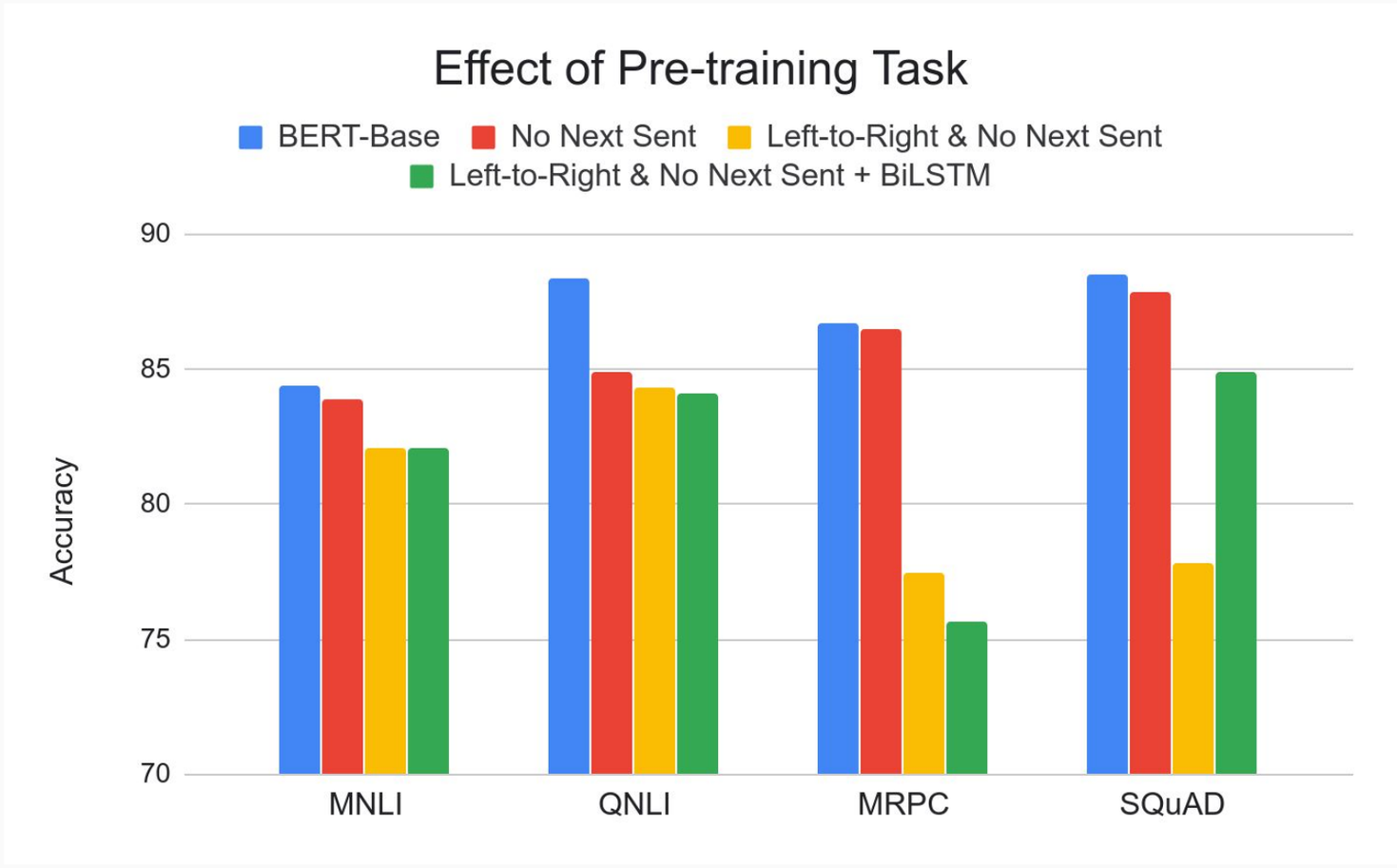
$$P_i = \frac{e^{V \cdot C_i}}{\sum_{j=1}^{4} e^{V \cdot C_j}}$$

Leaderboard

Human Performance (88.00%)
Running Best
Submissions

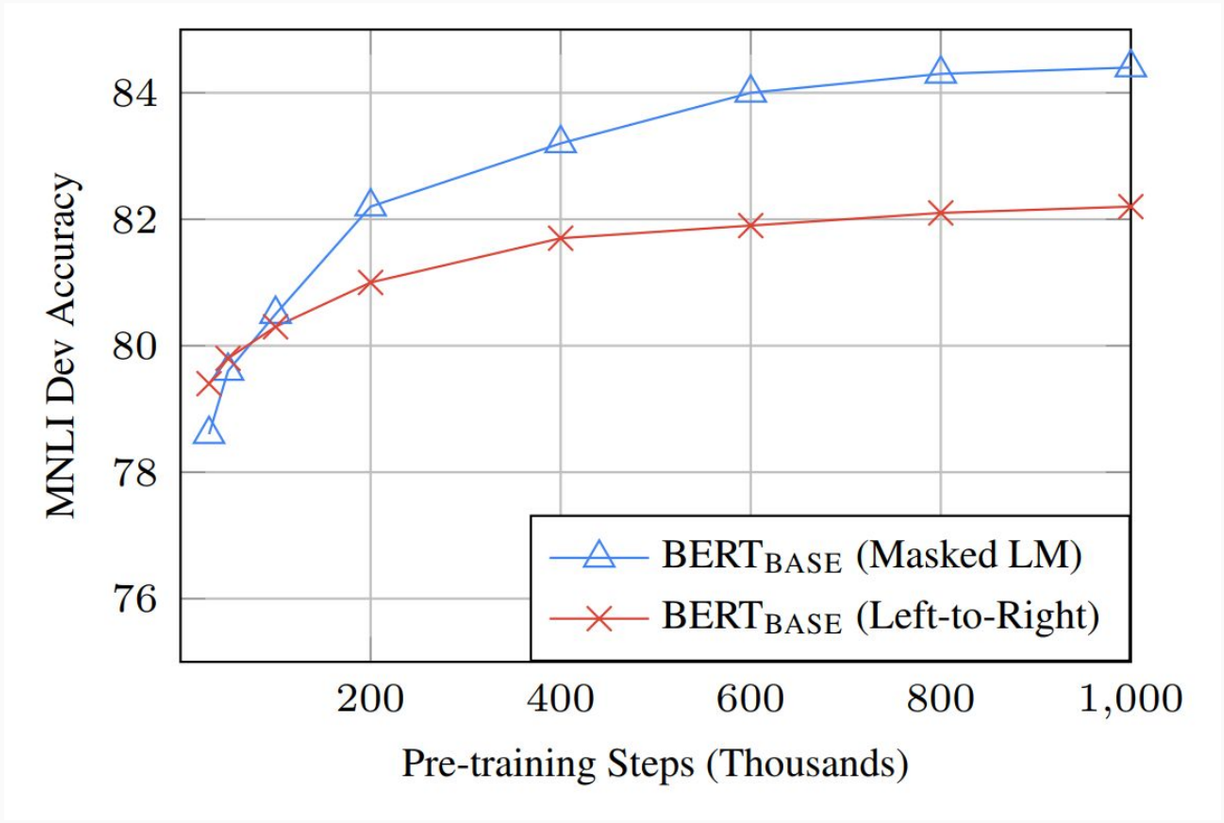
Rank	Model	Test Score
1	**BERT (Bidirectional Encoder Representations from Transfo...** *Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova* 10/11/2018	**86.28%**
2	**OpenAI Transformer Language Model** *Original work by Alec Radford, Karthik Narasimhan, Tim Salimans, ...* 10/11/2018	**77.97%**
3	**ESIM with ELMo** *Zellers, Rowan and Bisk, Yonatan and Schwartz, Roy and Choi, Yejin* 08/30/2018	**59.06%**
4	**ESIM with Glove** *Zellers, Rowan and Bisk, Yonatan and Schwartz, Roy and Choi, Yejin* 08/29/2018	**52.45%**

Effect of Pre-training Task



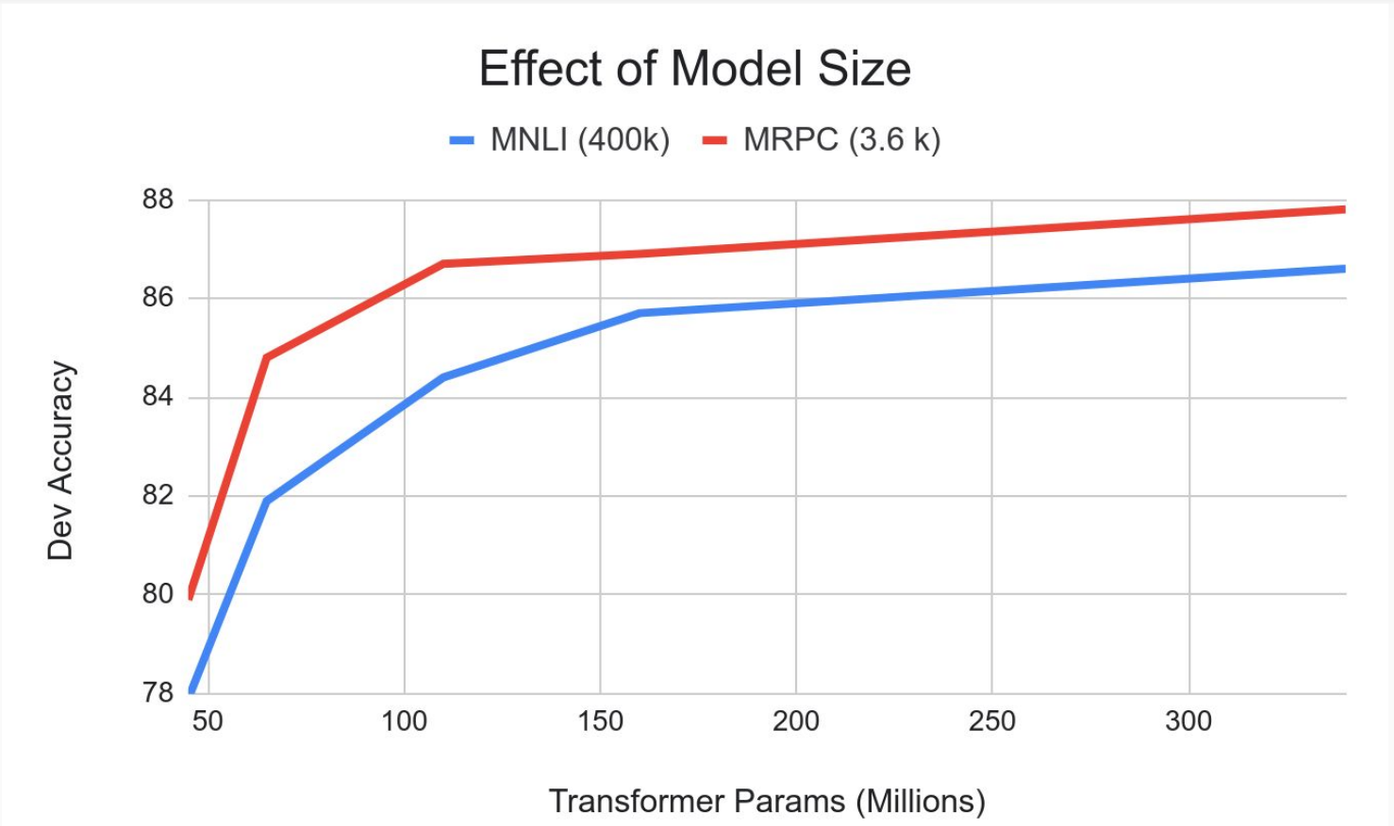
- Masked LM (compared to left-to-right LM) is very important on some tasks, Next Sentence Prediction is important on other tasks.
- Left-to-right model does very poorly on word-level task (SQuAD), although this is mitigated by BiLSTM

Effect of Directionality and Training Time



- Masked LM takes slightly longer to converge because we only predict 15% instead of 100%
- But absolute results are much better almost immediately

Effect of Model Size



- Big models help *a lot*
- Going from 110M -> 340M params helps even on datasets with 3,600 labeled examples
- Improvements have *not* asymptoted

Multilingual BERT

- Trained single model on 104 languages from Wikipedia. Shared 110k WordPiece vocabulary.

System	English	Chinese	Spanish
XNLI Baseline - Translate Train	73.7	67.0	68.8
XNLI Baseline - Translate Test	73.7	68.4	70.7
BERT - Translate Train	81.9	76.6	77.8
BERT - Translate Test	81.9	70.1	74.9
BERT - Zero Shot	81.9	63.8	74.3

- XNLI is MultiNLI translated into multiple languages.
- Always evaluate on human-translated Test.
- Translate Train: MT English Train into Foreign, then fine-tune.
- Translate Test: MT Foreign Test into English, use English model.
- Zero Shot: Use Foreign test on English model.

Common Questions

- Why did no one think of this before?
- Better question: Why wasn't contextual pre-training popular before 2018 with ELMo?
- Good results on pre-training is >1,000x to 100,000 more expensive than supervised training.
  - E.g., 10x-100x bigger model trained for 100x-1,000x as many steps.
  - Imagine it's 2013: Well-tuned 2-layer, 512-dim LSTM sentiment analysis gets 80% accuracy, training for 8 hours.
  - Pre-train LM on same architecture for a week, get 80.5%.
  - Conference reviewers: "Who would do something so expensive for such a small gain?"

Common Questions

- The model must be learning more than “contextual embeddings”
- Alternate interpretation: Predicting missing words (or next words) requires learning many types of language understanding features.
  - syntax, semantics, pragmatics, coreference, etc.
- Implication: Pre-trained model is much bigger than it needs to be to solve specific task
- Task-specific model distillation words very well

Common Questions

- Is modeling “solved” in NLP? I.e., is there a reason to come up with novel model architectures?
  - But that’s the most fun part of NLP research :(
- Maybe yes, for now, on some tasks, like SQuAD-style QA.
  - At least using the same deep learning “lego blocks”
- Examples of NLP models that are not “solved”:
  - Models that minimize total training cost vs. accuracy on modern hardware
  - Models that are very parameter efficient (e.g., for mobile deployment)
  - Models that represent knowledge/context in latent space
  - Models that represent structured data (e.g., knowledge graph)
  - Models that jointly represent vision and language

Common Questions

- Personal belief: Near-term improvements in NLP will be mostly about making clever use of “free” data.
  - Unsupervised vs. semi-supervised vs. synthetic supervised is somewhat arbitrary.
  - “Data I can get a lot of without paying anyone” vs. “Data I have to pay people to create” is more pragmatic distinction.
- No less “prestigious” than modeling papers:
  - *Phrase-Based & Neural Unsupervised Machine Translation*, Facebook AI Research, EMNLP 2018 Best Paper

Conclusions

- Empirical results from BERT are great, but biggest impact on the field is:
- With pre-training, bigger == better, without clear limits (so far).
- Unclear if adding things on top of BERT really helps by very much.
  - Good for people and companies building NLP systems.
  - Not necessary a “good thing” for researchers, but important.

draw stuff

masked transformer exercise