VIRGINIA TECH

Contextual Representations and Pre-trained Language Models

Lifu Huang

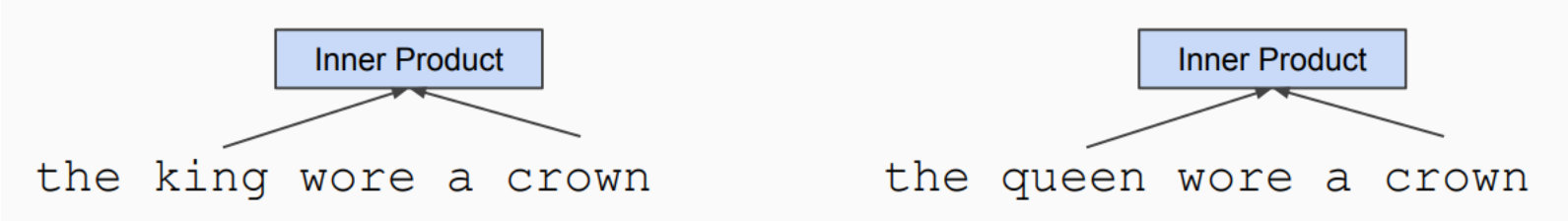
lifuh@vt.edu

Torgersen Hall, Suite 3160E

Pre-trained Word Embeddings

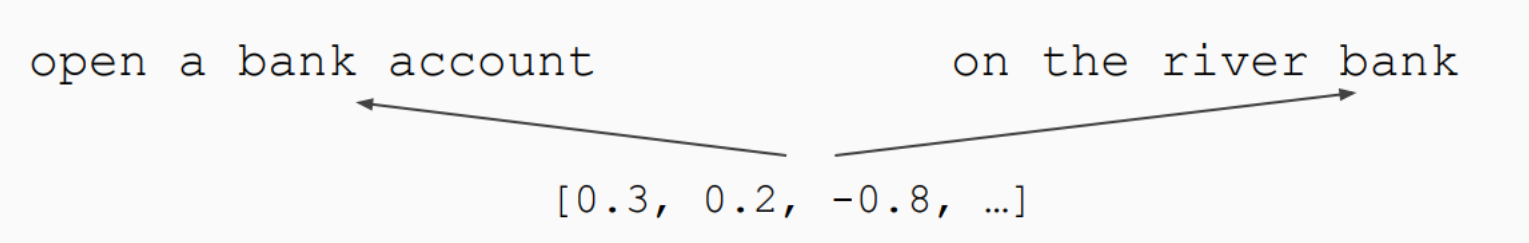
- Word embeddings are the basis of deep learning for NLP

- Word embeddings (e.g., word2vec, GloVe) are often pre-trained on text corpus from co-occurrence statistics



Contextual Representations

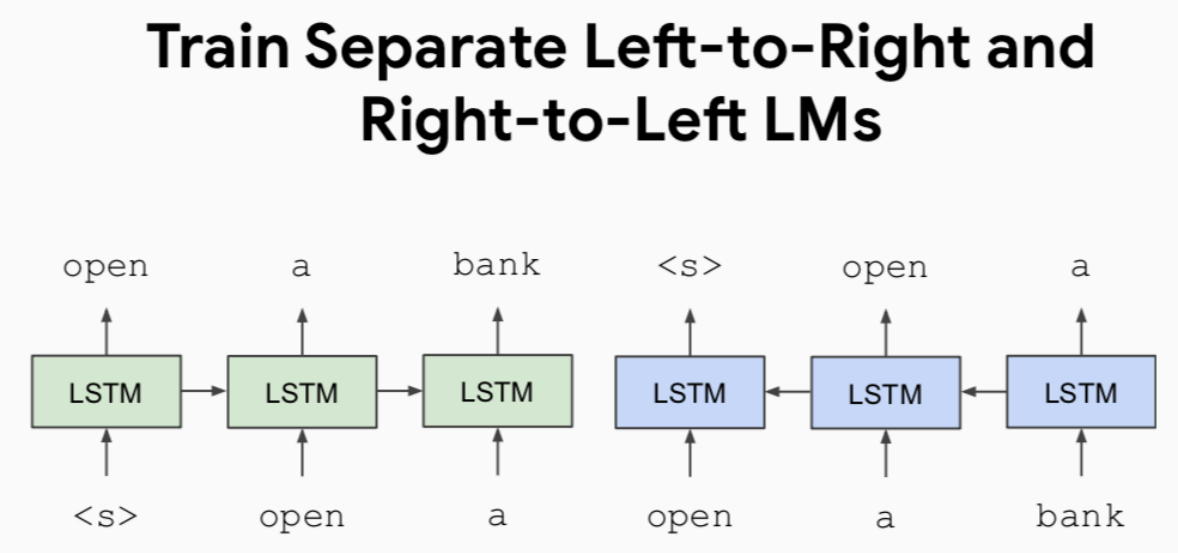
- **Problem**: Word embeddings are static and applied in a context free manner

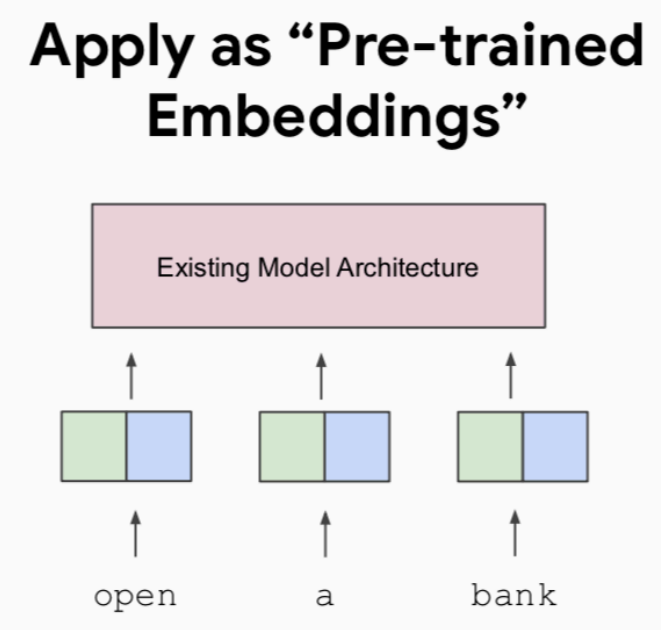


- **Solution**: Train contextual representations on text corpus

Learning of Contextual Representations

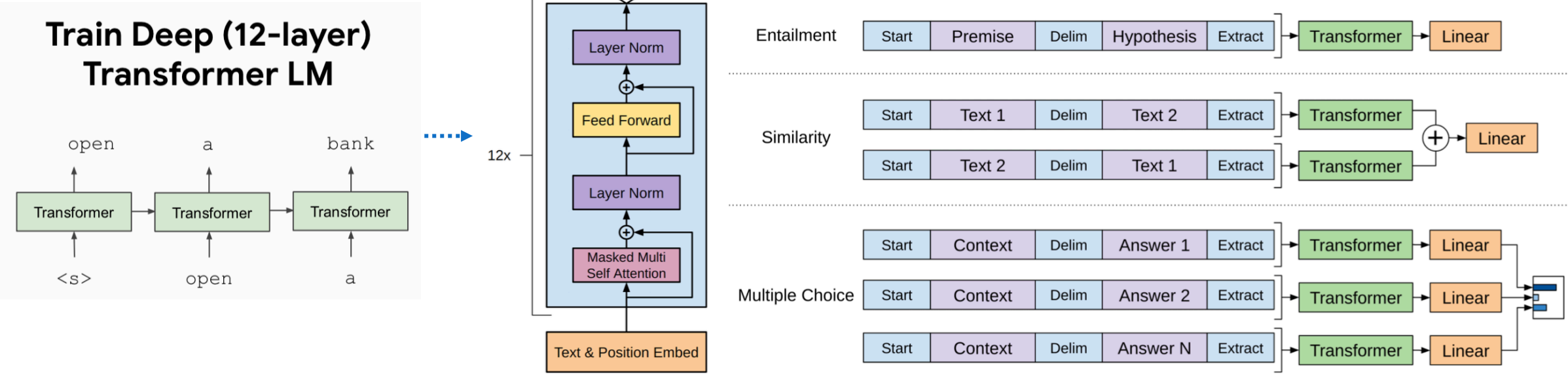
- *ELMo: Deep Contextual Word Embeddings*, AI2 & University of Washington, 2017





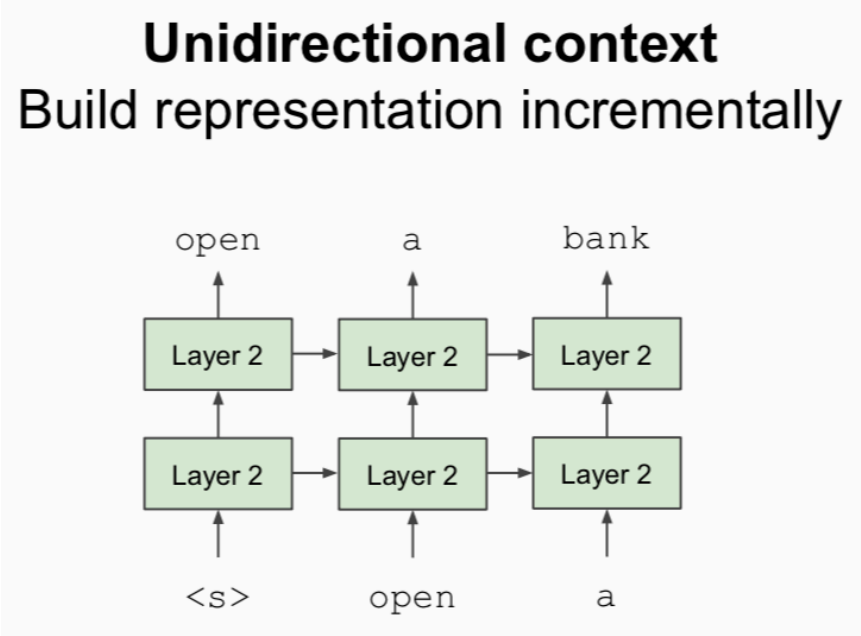
OpenAI-GPT: Improving Language Understanding by Generative Pre-Training

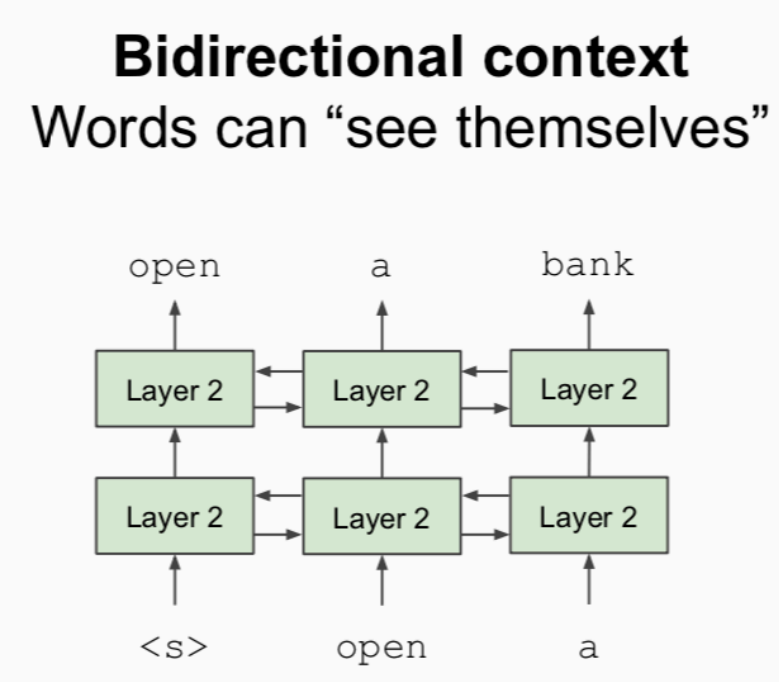
- Generative Pre-training (Unsupervised) and Discriminative Fine-tuning (Supervised)



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

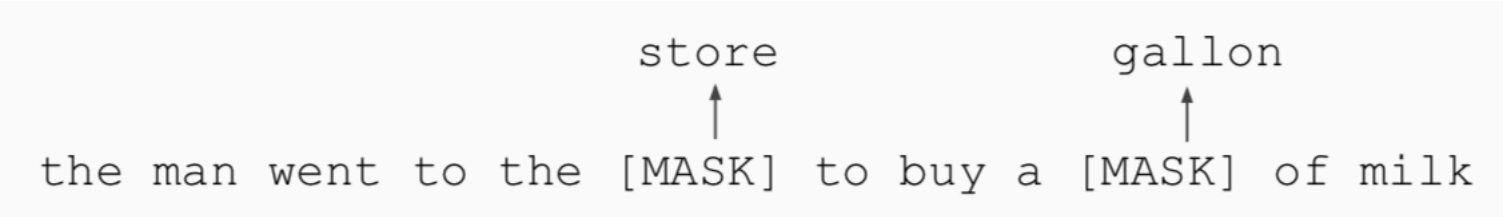
- **Problem**: Language models only use left context or right context, but language understanding is bidirectional
- Why are LMs unidirectional?





Masked Language Model (Masked LM)

- Mask out k% of the input words, and then predict the masked words
  - k =15%



- Too little masking: Too expensive to train
- Too much masking: Not enough context

Masked Language Model (Masked LM)

- **Problem**: Mask token never seen at fine-tuning
- Solution: 15% of the words to predict but don't replace with [MASK] 100% of the time
- 80% of the time, replace with [MASK]
  - went to the store → went to the [MASK]
- 10% of the time, replace random word
  - went to the store → went to the running
- 10% of the time, keep same
  - went to the store → went to the store

Next Sentence Prediction

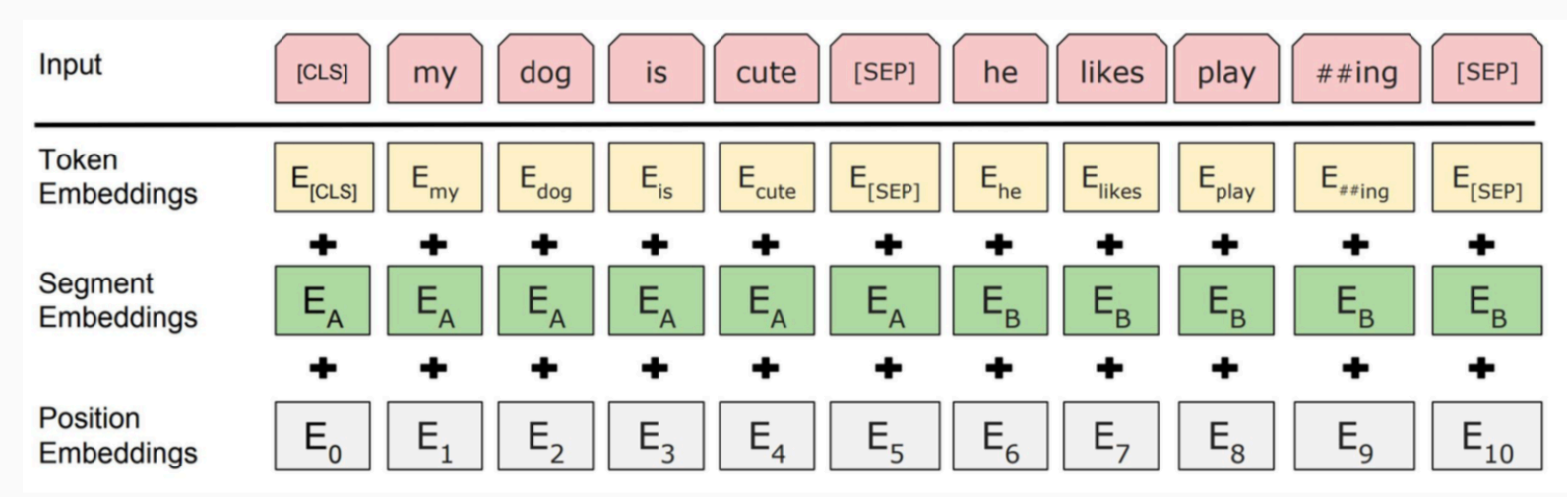
- To learn *relationships* between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence
  - how to segment a story into different chunks
  - whether one event is following or preceding of another event
  - whether one sentence can be implied or entailed by another sentence
  - whether the second sentence answers the questions of the first sentence
  - … …

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`

Model Details

- Input Representation
  - Use 30,000 WordPiece vocabulary on input
  - Each token is sum of three embeddings
    - Token Embeddings, Segment Embeddings, Position Embeddings
  - Add special tokens [CLS] (aggregated sequence representation) and [SEP] (separator)
  - Single sequence is much more efficient

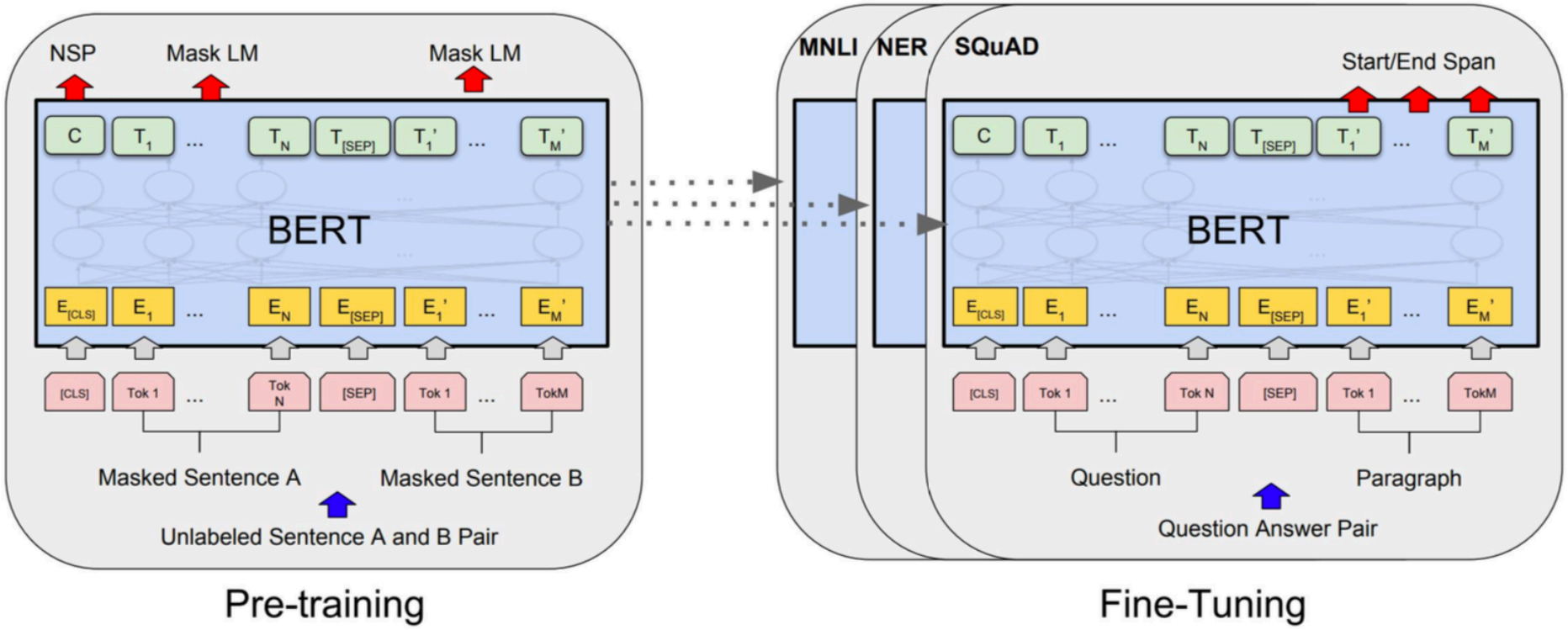


Pre-training

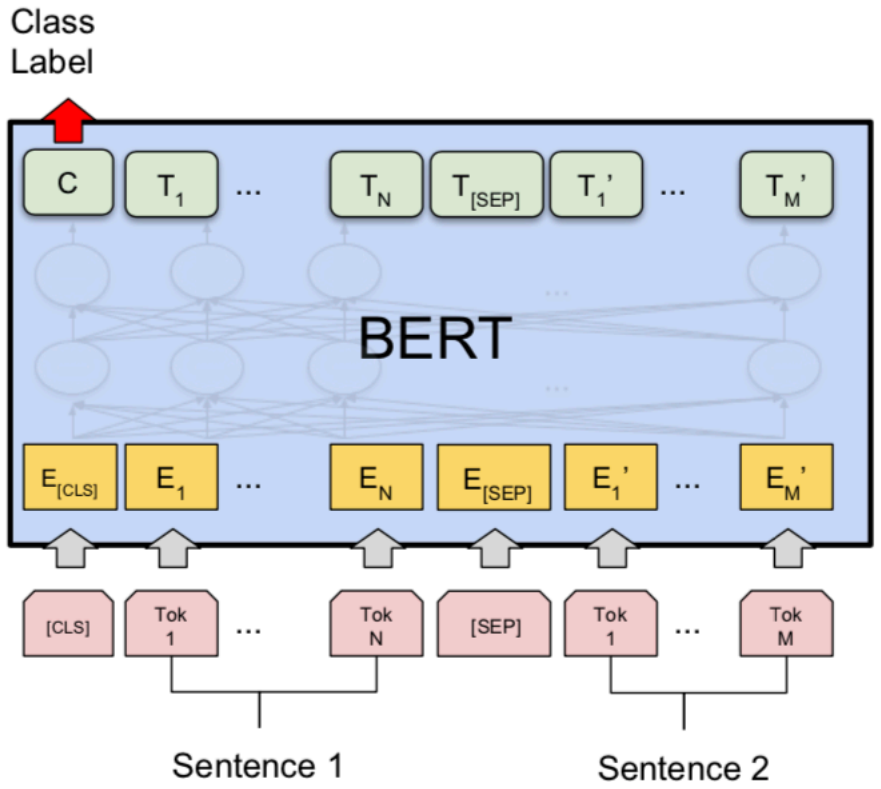
- Pre-training on large-scale unlabeled corpus with **Mask LM** and **Next Sentence prediction**
- **Data**: Wikipedia (2.5B words) + BookCorpus (800M words)
- **Batch Size**: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
- **Training Time**: 1M steps (~40 epochs)
- **Optimizer**: 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 on target tasks – Transfer learning



Fine-tuning Architectures – Sentence Pair Classification



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

MNLI:

S1: At the other end of Pennsylvania Avenue, people began to line up for a White House tour.

S2: People formed a line at the end of Pennsylvania Avenue. [Entailment]

MRPC:

S1: Charles O. Prince, 53, was named as Mr. Weill's successor.

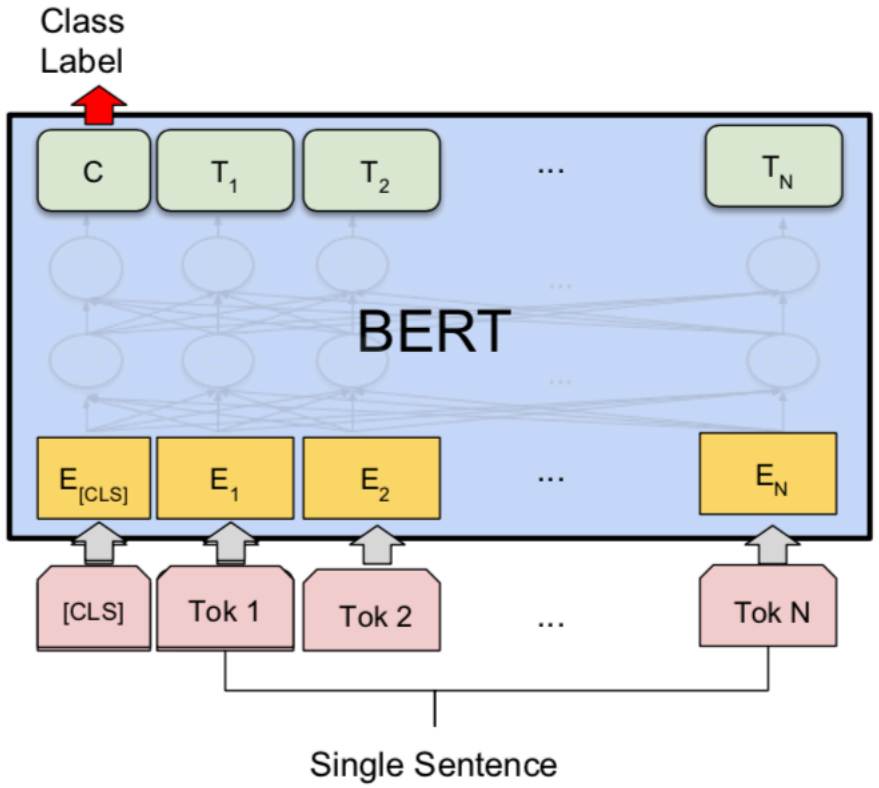
S2: Mr. Weill's longtime confidant, Charles O. Prince, 53, was named as his successor.

SWAG:

On stage, a woman takes a seat at the piano. She

a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
d) nervously sets her fingers on the keys

Fine-tuning Architectures – Single Sentence Classification



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

SST-2:

S: Roger Dodger is one of the most compelling variations on this theme. [Positive]

CoLA:

a. My friend has/*have to go.
b. My friends *has/have to go

Fine-tuning Architectures – Results

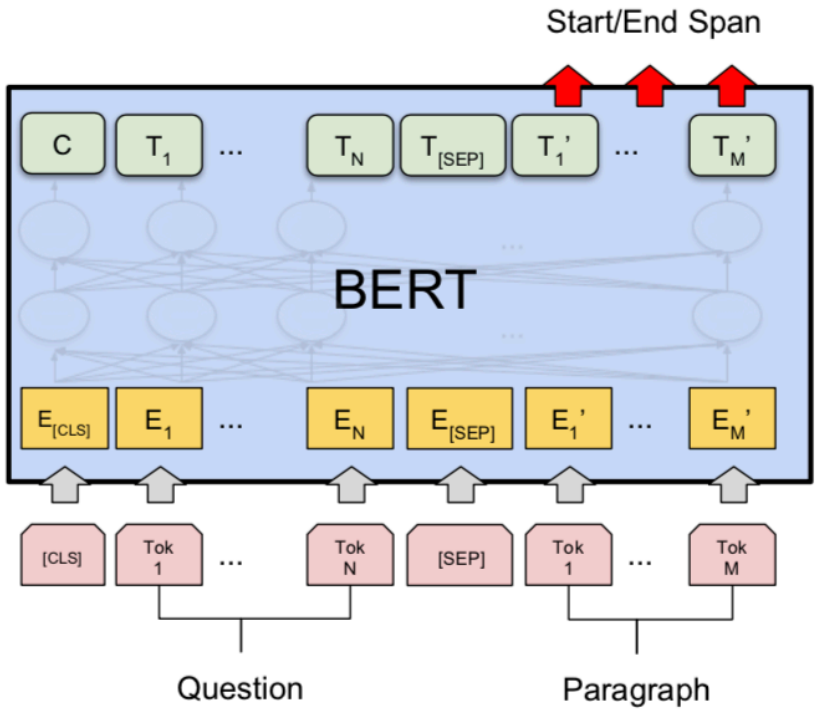
System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Table 1: GLUE Test results, scored by the evaluation server (https://gluebenchmark.com/leaderboard).

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Table 4: SWAG Dev and Test accuracies. †Human performance is measured with 100 samples, as reported in the SWAG paper.

Fine-tuning Architectures – Span Prediction



(c) Question Answering Tasks: SQuAD v1.1

SQuAD 1.1:

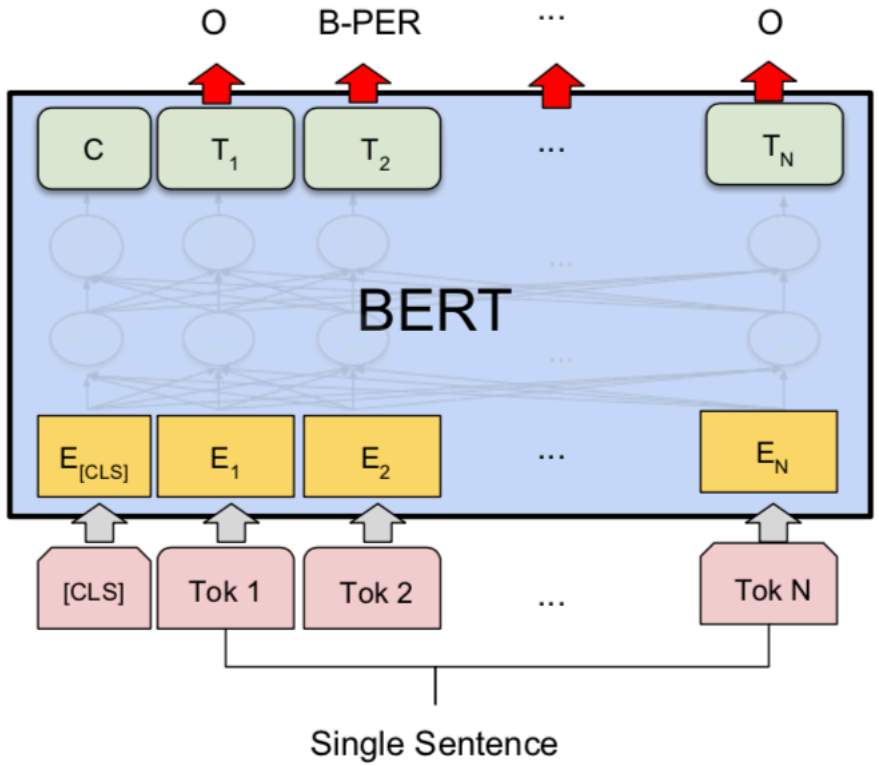
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Fine-tuning Architectures – Sequence Labeling



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

CoNLL-2003:

```
U.N.          NNP   I-NP   I-ORG
official      NN    I-NP   O
Ekeus         NNP   I-NP   I-PER
heads         VBZ   I-VP   O
for           IN    I-PP   O
Baghdad       NNP   I-NP   I-LOC
.             .     O      O
```

Fine-tuning Architectures – Results

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

Table 3: SQuAD 2.0 results. We exclude entries that use BERT as one of their components.

System	Dev F1	Test F1
ELMo (Peters et al., 2018a)	95.7	92.2
CVT (Clark et al., 2018)	-	92.6
CSE (Akbik et al., 2018)	-	**93.1**
Fine-tuning approach		
$\text{BERT}_{\text{LARGE}}$	96.6	92.8
$\text{BERT}_{\text{BASE}}$	96.4	92.4
Feature-based approach ($\text{BERT}_{\text{BASE}}$)		
Embeddings	91.0	-
Second-to-Last Hidden	95.6	-
Last Hidden	94.9	-
Weighted Sum Last Four Hidden	95.9	-
Concat Last Four Hidden	96.1	-
Weighted Sum All 12 Layers	95.5	-

Table 7: CoNLL-2003 Named Entity Recognition results. Hyperparameters were selected using the Dev set. The reported Dev and Test scores are averaged over 5 random restarts using those hyperparameters.

Post-BERT Pre-training Advancements - RoBERTa

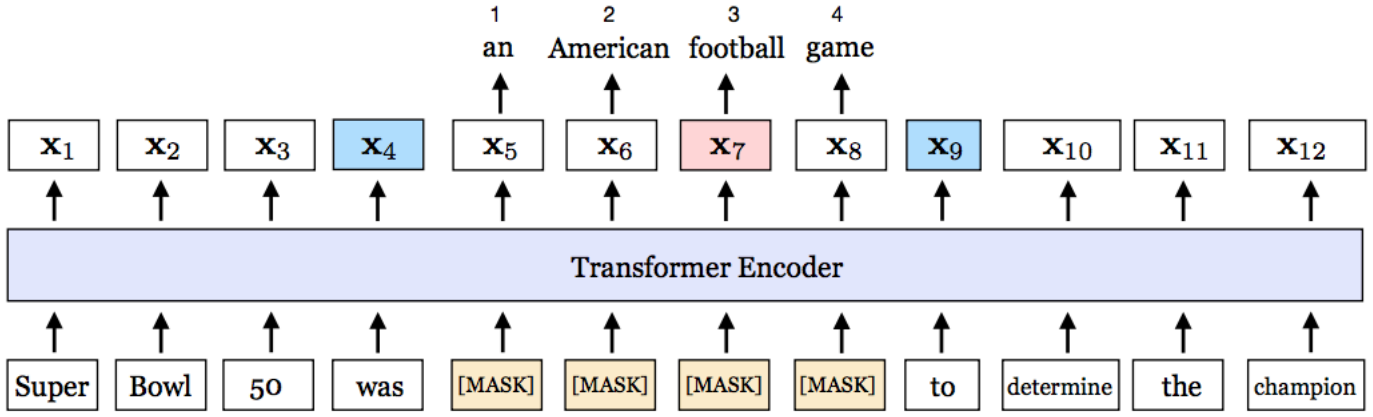
- *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (Liu et al, University of Washington and Facebook, 2019)
- Trained BERT for more epochs and/or on more data
  - Showed that more epochs alone helps, even on same data
  - More data also helps
- Improved masking and pre-training data slightly

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-

Post-BERT Pre-training Advancements - SpanBERT

- SpanBERT: Improving Pre-training by Representing and Predicting Spans. Joshi et al., 2019
- Masking continuous random spans and then predict the masked span with boundary representations
- Single-sequence training instead of bi-sequence training

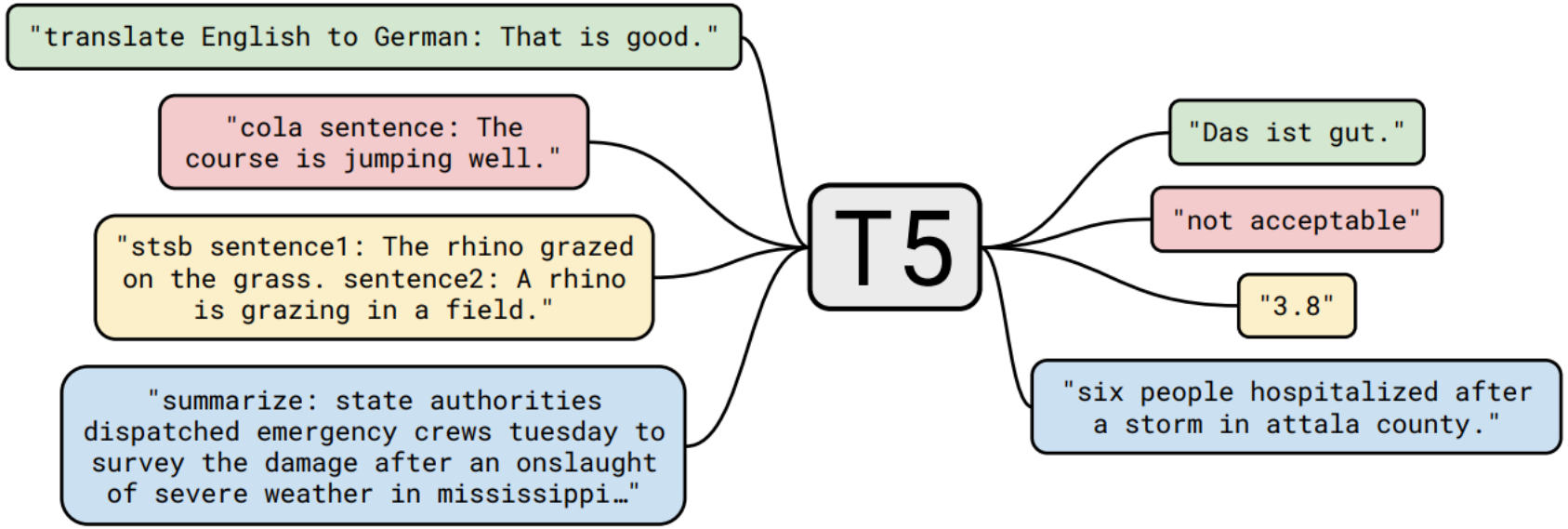
$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$



Perform better on Span based tasks, e.g., SQuAD, Coreference

Post-BERT Pre-training Advancements – T5

- Exploring the Limits of Transfer Learning with a Unified **Text-to-Text Transformer**. Raffel et al., 2020
- Pretraining Objectives
  - Language Modeling
  - Deshuffling
  - Corrupting Spans
- Conclusion:
  - Scaling up model size and amount of training data is very helpful
  - Training Strategies



Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**

Post-BERT Pre-training Advancements - More

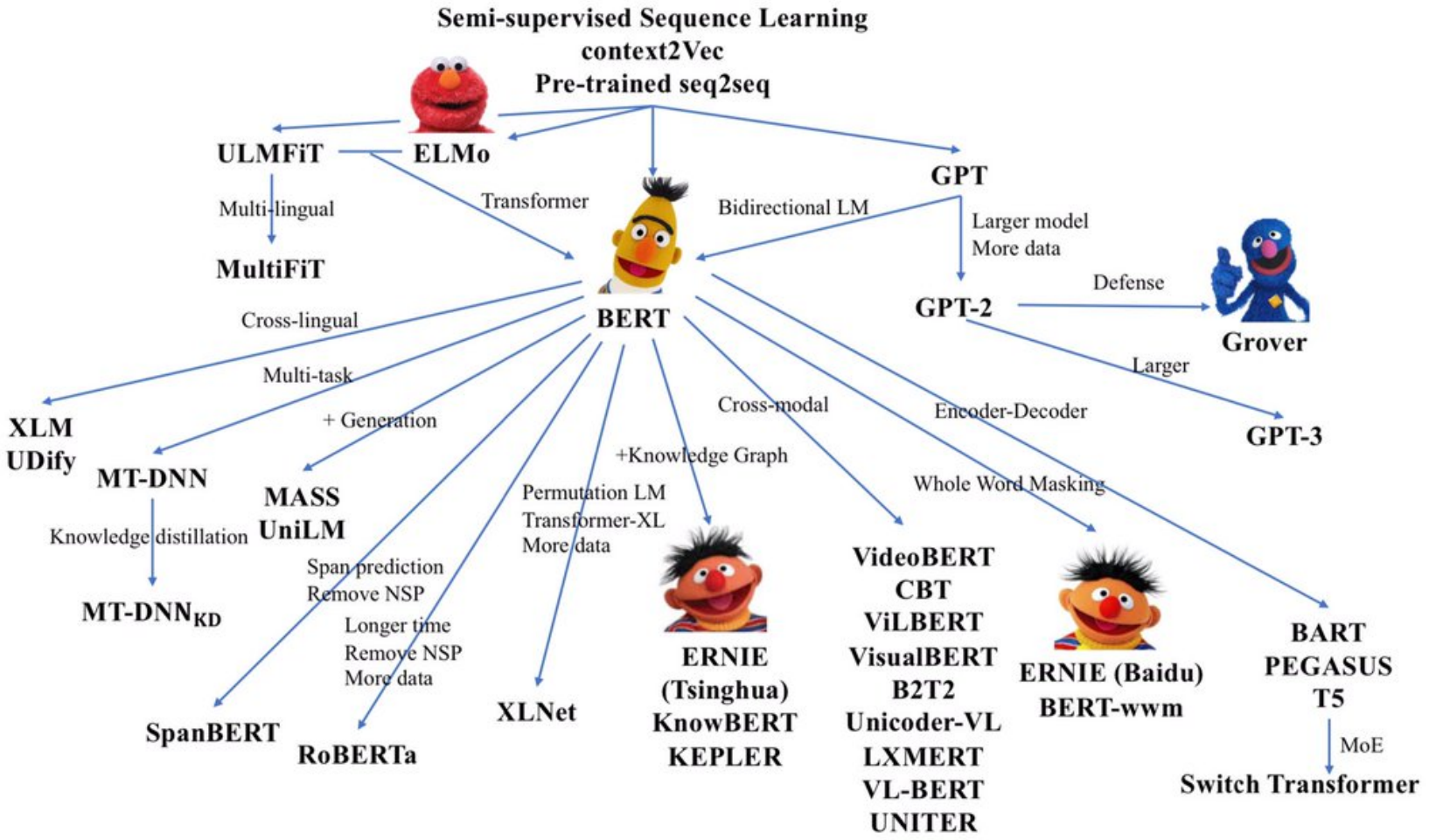
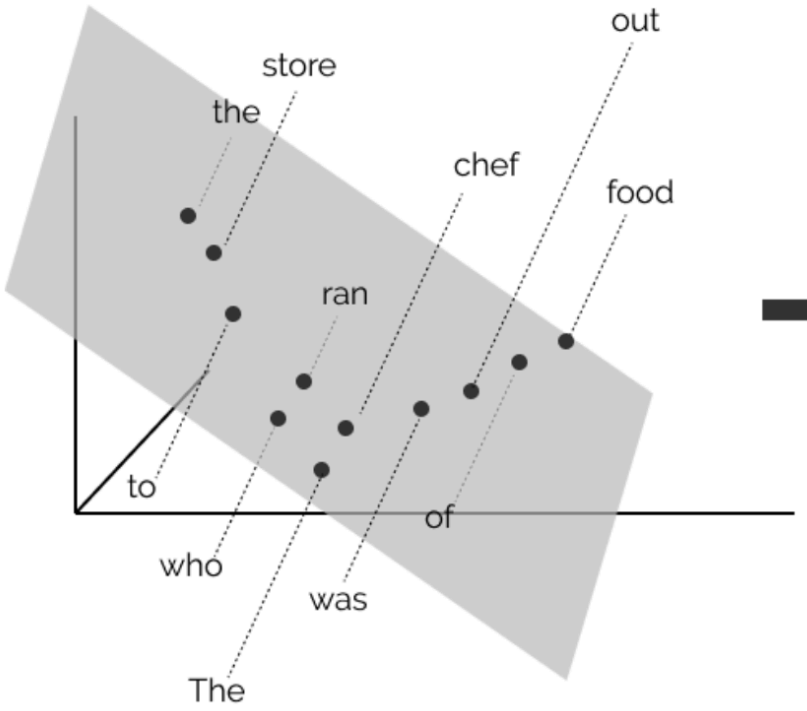


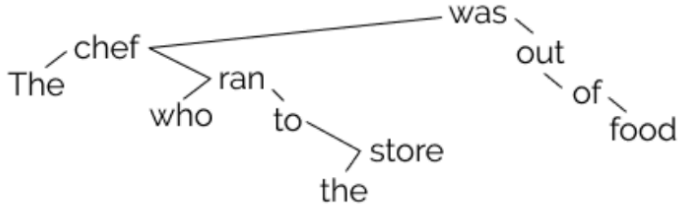
Figure 8: The family of recent typical PTMs, including both pre-trained language models and multimodal models.

Han et al., 2021

Language Model Probing

- Whether the contextual representations have encoded the token-level linguistic phenomena?
  - Part-of-speech tags (Blevins et al., 2018), Morphology (Belinkov et al., 2017), Word-sense disambiguation (Peters et al., 2018), Syntactic Tree (Peters et al., 2018b)





$$\min_B \sum_\ell \frac{1}{|s^\ell|^2} \sum_{i,j} \left| d_{T^\ell}(w_i^\ell, w_j^\ell) - d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell)^2 \right|$$

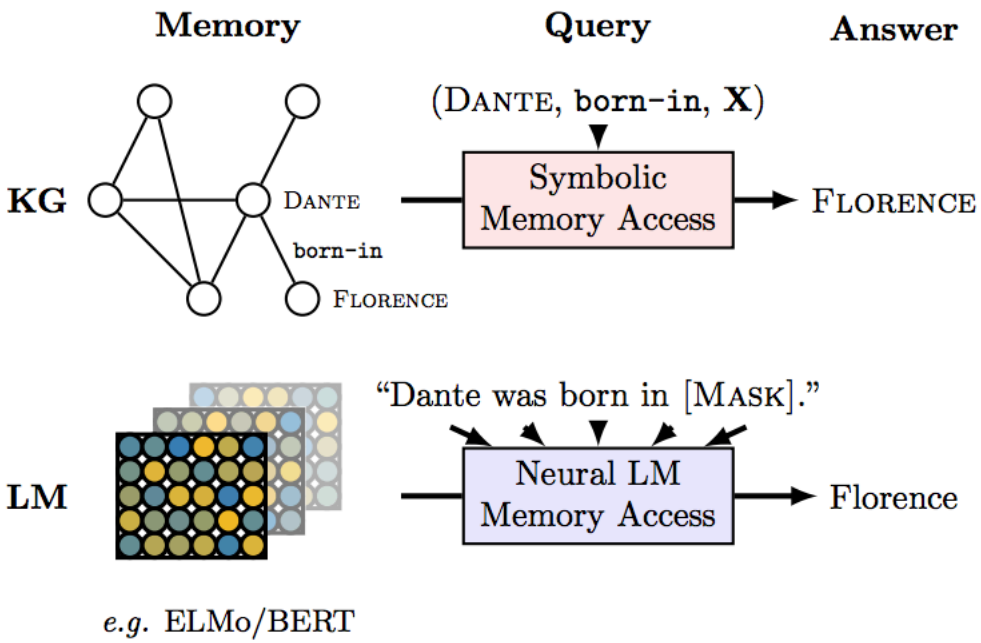
$$\min_B \sum_\ell \frac{1}{|s_\ell|} \sum_i \left(\|w_i\| - \|Bh_i\|^2 \right)$$

Peters et al., 2018b

Language Model Probing

- Whether the contextual representations have encoded various sentence structures, e.g., syntactic or semantic relation?
  - Dependency relation, Semantic role, Coreference, etc. (Tenney et al., 2019)

- Whether the pre-trained language models have encoded the factual knowledge (Petroni et al., 2019) or commonsense knowledge (Zhou et al., 2020, Lin et al., 2020) embedded in the large-scale pre-training corpus?



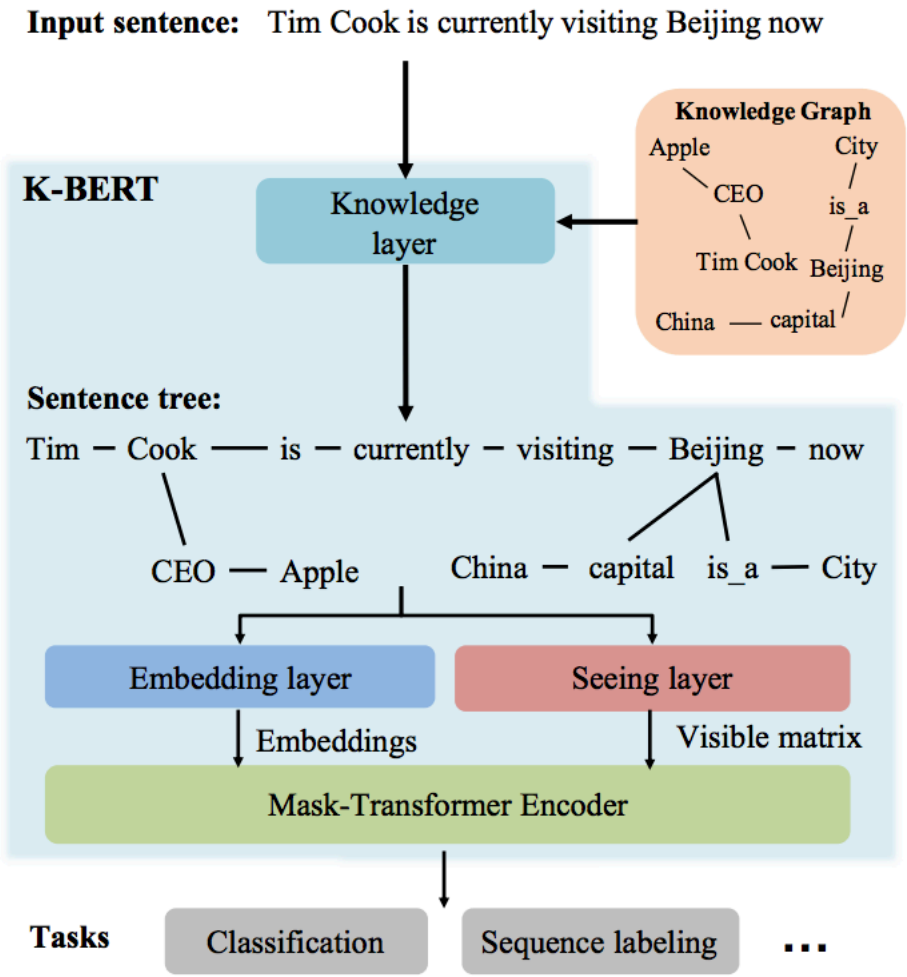
Petroni et al., 2019

Distillation

- BERT and other pre-trained language models are extremely large and expensive
- How can we apply them without much computational resources? – Distillation
- Distillation: allows the model to focus only on a subset of features that are useful for the given tasks
- Simple Technique
  - Train “Teacher”: Use SOTA pre-training + fine-tuning technique to train model with maximum accuracy
  - Label a large amount of unlabeled input examples with Teacher
  - Train “Student”: Much smaller model (e.g., 50x smaller) which is trained to mimic Teacher output
  - Student objective is typically Mean Square Error or Cross Entropy

Knowledge-aware Language Model Pre-training

- Improve the knowledge-awareness of language model pre-training by incorporating external knowledge resources
  - Lexical Constraints, e.g., whether two words have a particular semantic relation or not (Lauscher et al., 2019)
  - Incorporating external knowledge graphs into pre-training (Yao et al., 2019, Liu et al., 2019, He et al., 2020, Wang et al., 2020)



Liu et al., 2019

Prompt Tuning

- Select one or many appropriate prompts to manipulate the model behavior so that the pre-trained LM itself can predict the desired output, sometimes even without any additional task-specific training (Radford et al., 2019, Petroni et al., 2019, Brown et al., 2020, Raffel et al., 2020)
  - Hard Prompts - templates
  - Soft Prompts – continuous parameters
  - future directions?

Type	Task	Input ([X])	Template	Answer ([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1] [X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman ... ...
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...

Liu et al., Pre-train, Prompt and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing 2021