N-gram Language Models

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Agenda

In this session, we will discuss:

- Pre-trained language model
- Language Modeling
- N-gram language model

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Pre-Trained Language Models

- Deep neural networks trained on large amounts of data:
  - Especially transformer based;
  - Millions of parameters.

- Language Modeling:
  - Predict the next token given a sequence.
  - Self-supervised task: No need for human annotations.

- Transfer Learning:
  - Reuse knowledge learned from one task in another task.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Language Modeling

- Language models estimate the probability of text sequences.
- Useful to predict the next word:
  - Given a sequence ($W_1, W_2, W_3, \ldots, W_t$) , calculate the most likely word to be W_{t+1}.

I like pizza with

	$P(w_t \mid w_1, w_2, \ldots, w_{t-1})$	
mushrooms	0.1413	P(mushrooms | I, like, pizza, with)
olives	0.1376	P(olives | I, like, pizza, with)
anchovies	0.1320	P(anchovies | I, like, pizza, with)
…	…	
pineapple	0.0003	P(pineapple | I, like, pizza, with)

- A Language Model is a system trained to solve this task.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Language Modeling (Cont.)

- Language models estimate the probability of text sequences.
- Useful to predict the next word:
  - Given a sequence (W_1, W_2, W_3, … , W_t) , calculate the most likely word to be W_{t+1}.

How to learn this? → $P(w_t | w_1, w_2, \dots, w_{t-1})$

I like pizza with

mushrooms	0.1413	P(mushrooms | I, like, pizza, with)
olives	0.1376	P(olives | I, like, pizza, with)
anchovies	0.1320	P(anchovies | I, like, pizza, with)
…	…	
pineapple	0.0003	P(pineapple | I, like, pizza, with)

- A Language Model is a system trained to solve this task.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model

- **Assumption:**
  - The probability of a word only depends on the *n-1* previous words:

with n=3 (trigram):	$P\,(w_t\,|\,w_{t-2},\,w_{t-1})$	E.g., P (mushrooms | pizza, with)
with n=2 (bigram):	$P\,(w_t\,|\,w_{t-1})$	E.g., P(mushrooms | with)
with n=1 (unigram):	$P\,(w_t)$	E.g., P(mushrooms)

  - To calculate an n-gram probability:

$$P(w_t|w_{t-\mathrm{n}+1},\ldots,w_{t-1}) = \frac{count(w_{t-\mathrm{n}+1},\ldots w_{t-1},w_t)}{count(w_{t-\mathrm{n}+1},\ldots w_{t-1})}$$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- **Assumption:**
  - The probability of a word only depends on the *n-1* previous words:

with n=3 (trigram):	$P\,(w_t \mid w_{t-2},\, w_{t-1})$	E.g., P (mushrooms | pizza, with)
with n=2 (bigram):	$P\,(w_t \mid w_{t-1})$	E.g., P(mushrooms | with)
with n=1 (unigram):	$P\,(w_t)$	E.g., P(mushrooms)

  - To calculate an n-gram probability:

n=3 (trigram)

$$P\left(w_t \middle| w_{t-2,}w_{t-1}\right) = \frac{count(w_{t-2}, w_{t-1}, w_t)}{count(w_{t-2}, w_{t-1})}$$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- **Assumption:**
  - The probability of a word only depends on the n-1 previous words:

with n=3 (trigram):	$P\,(w_t\,\vert\,w_{t-2},\,w_{t-1})$	E.g., P (mushrooms | pizza, with)
with n=2 (bigram):	$P\,(w_t\,\vert\,w_{t-1})$	E.g., P(mushrooms | with)
with n=1 (unigram):	$P\,(w_t)$	E.g., P(mushrooms)

  - To calculate an n-gram probability:

n=3 (trigram)

$$P(\text{mushrooms} \mid \text{pizza, with}) = \frac{count(\text{pizza with mushrooms})}{count(\text{pizza with})}$$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- **Assumption:**
  - The probability of a word only depends on the *n-1* previous words:

with n=3 (trigram):	$P(w_t \mid w_{t-2}, w_{t-1})$	E.g., P (mushrooms | pizza, with)
with n=2 (bigram):	$P(w_t \mid w_{t-1})$	E.g., P(mushrooms | with)
with n=1 (unigram):	$P(w_t)$	E.g., P(mushrooms)

  - To calculate an n-gram probability:

n=1 (unigram)

$$P(w_t) = \frac{count(w_t)}{all\ words\ in\ corpus}$$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

THE UNIVERSITY OF ARIZONA

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with ______

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with ______

For the 3-gram model we use only the last 2 previous words.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with ______

$P(w \mid \text{pizza, with})$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with mushrooms

$P(w \mid \text{pizza, with})$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with mushrooms ______

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with mushrooms ______

$$P(w \mid \text{with, mushrooms})$$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with mushrooms and

$$P(w \mid \text{with, mushrooms})$$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with mushrooms and butter

$P(w \mid \text{mushrooms, and})$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

n-gram Language Model (Cont.)

- We can use an n-gram Language Model for Text Generation.
  - E.g., 3-gram Language Model:

I like pizza with mushrooms and butter

$P(w \mid \text{mushrooms, and})$

A Language Model using a short context produces incoherent text.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Autoregressive and Masked Language Model

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Agenda

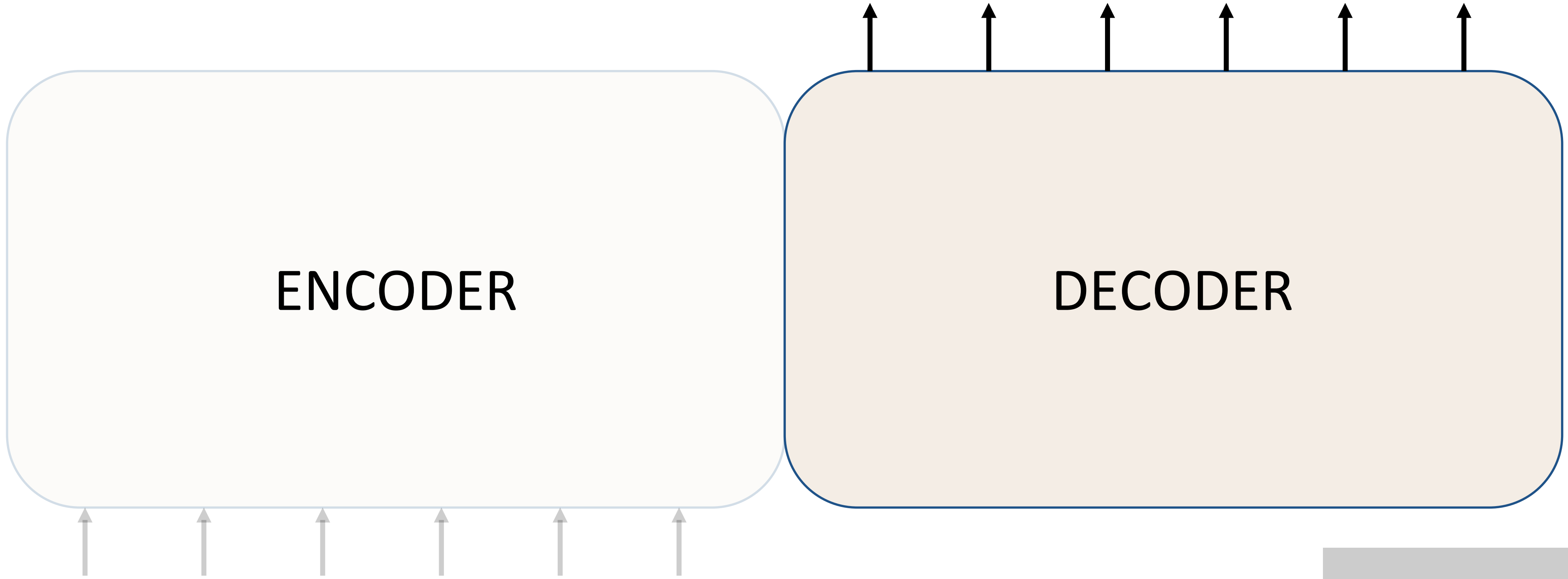
In this session, we will discuss:

- Autoregressive Language Model
- Masked Language Modeling

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

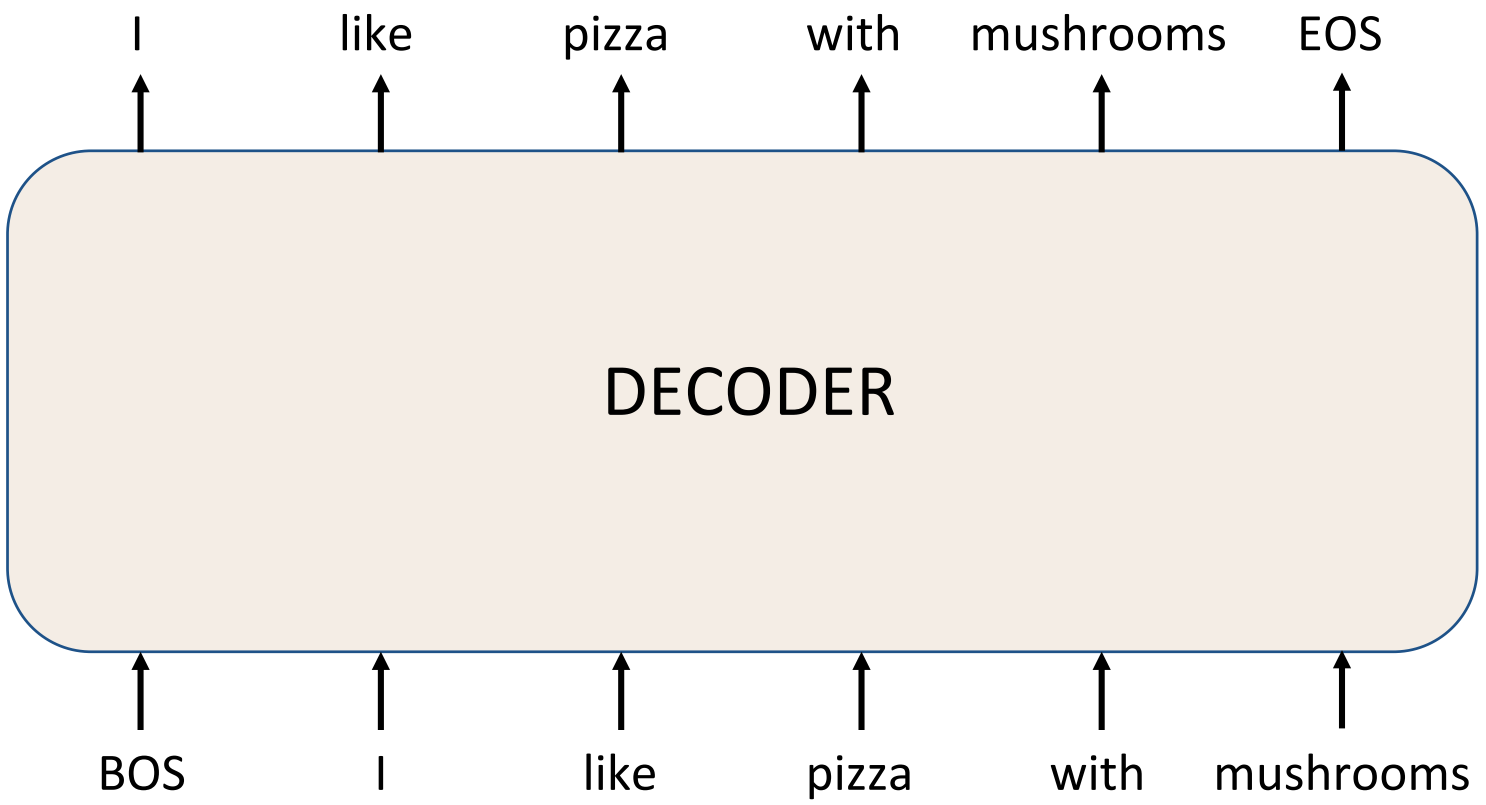
Autoregressive Language Model

- We can use a Decoder to train a Language Model.



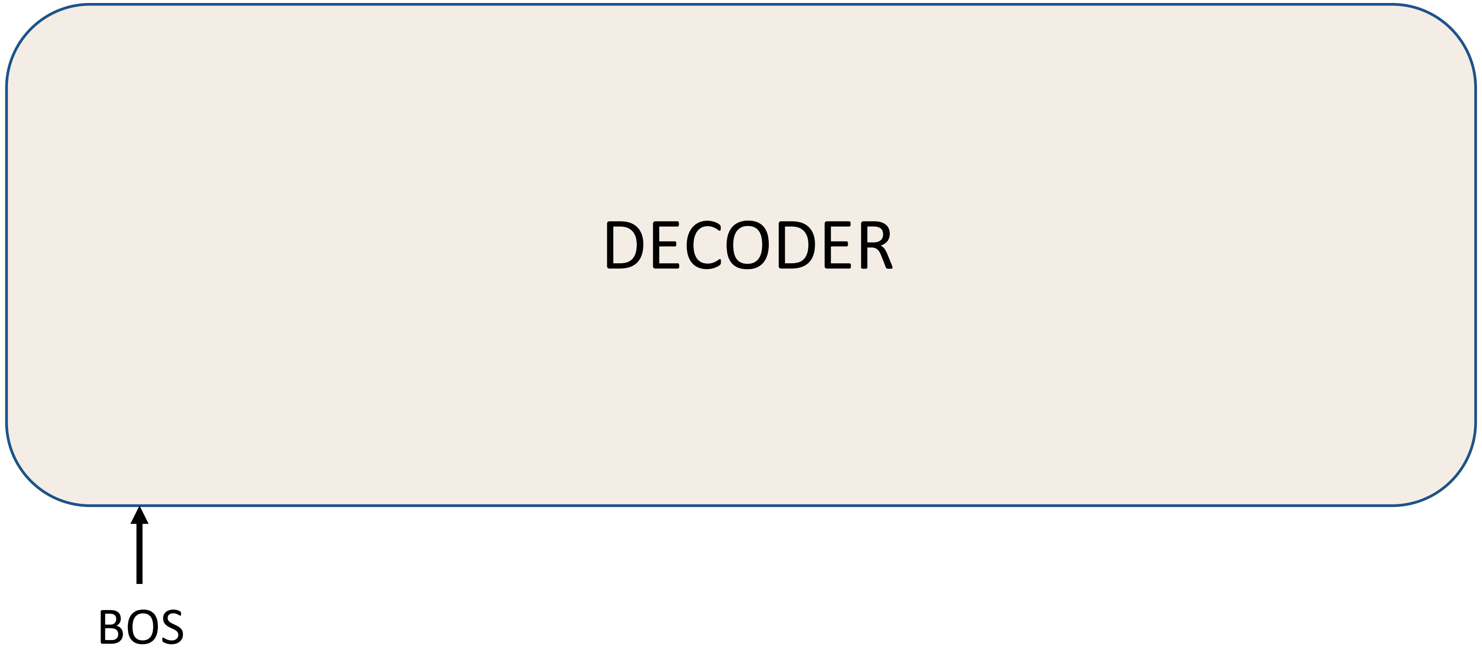
This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Autoregressive Language Model: Training



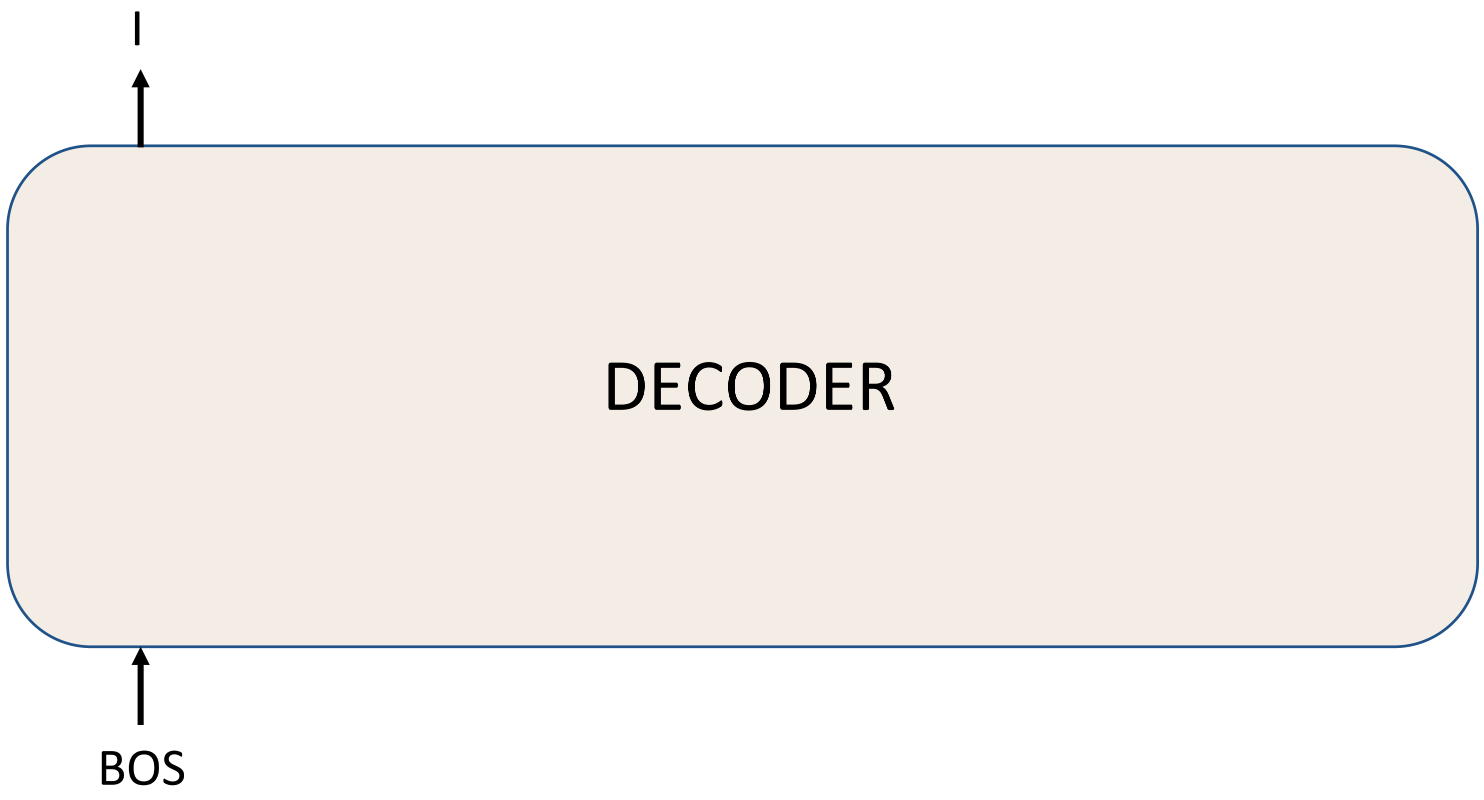
This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Autoregressive Language Model: Prediction



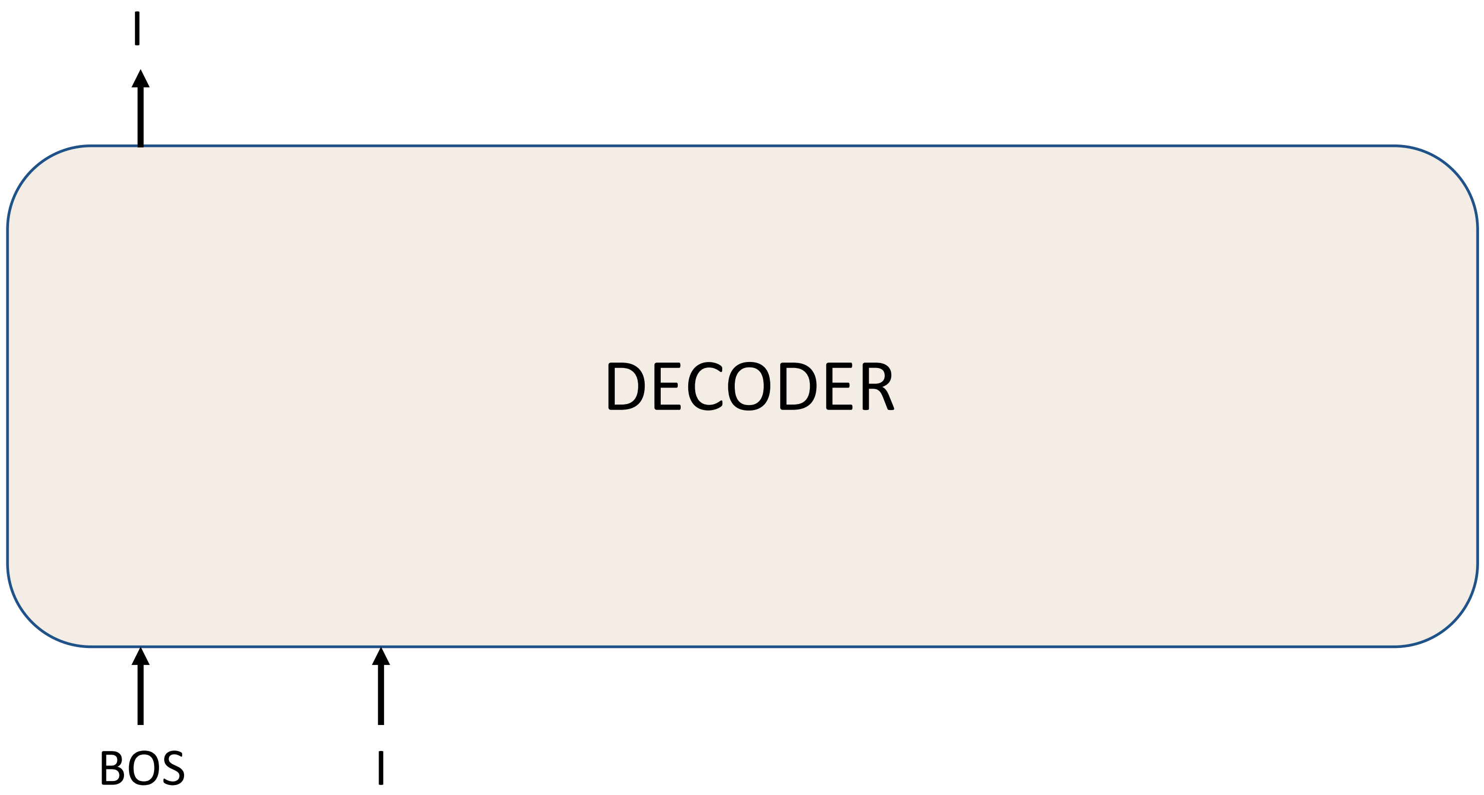
This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Autoregressive Language Model: Prediction (Cont.)



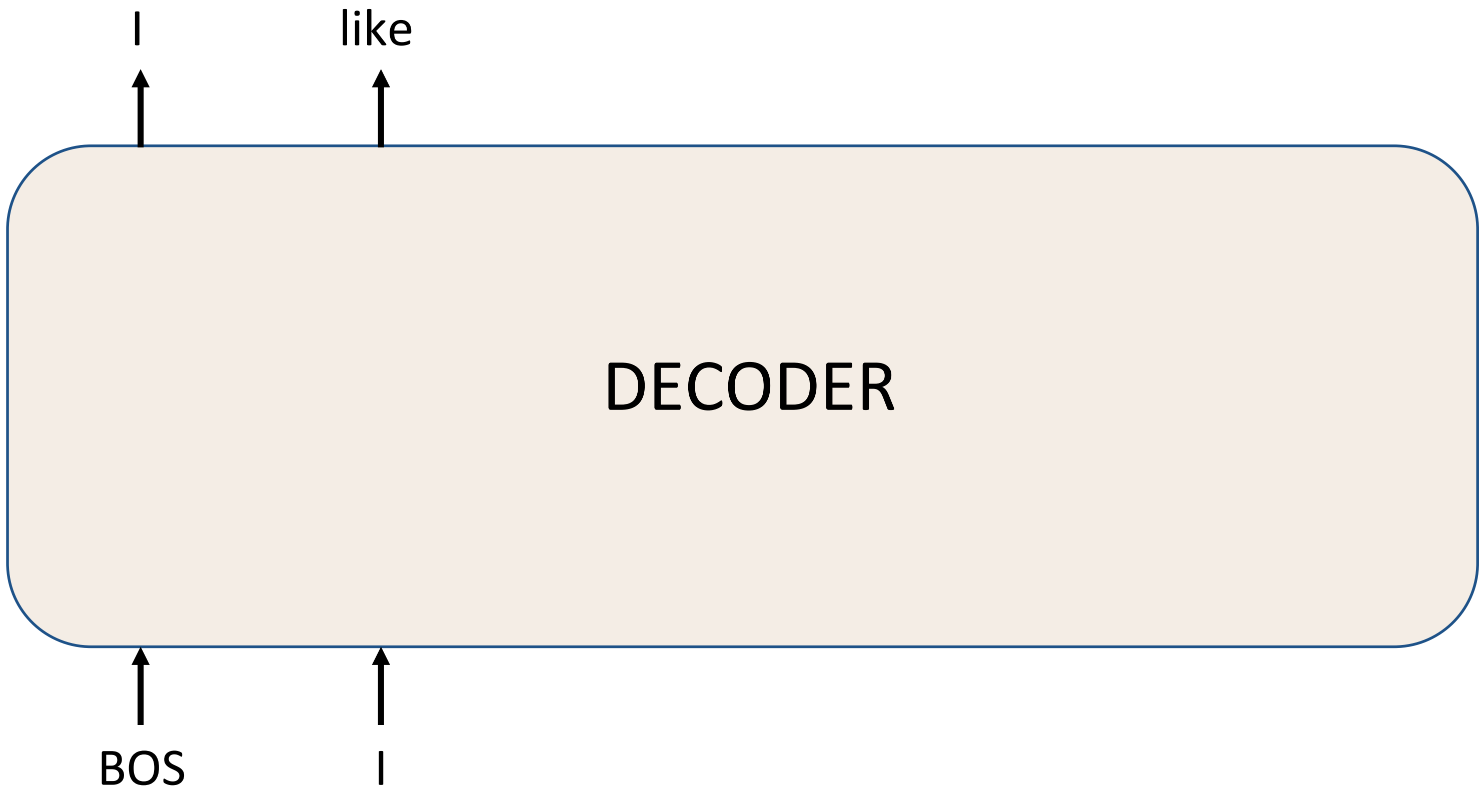
This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Autoregressive Language Model: Prediction (Cont.)



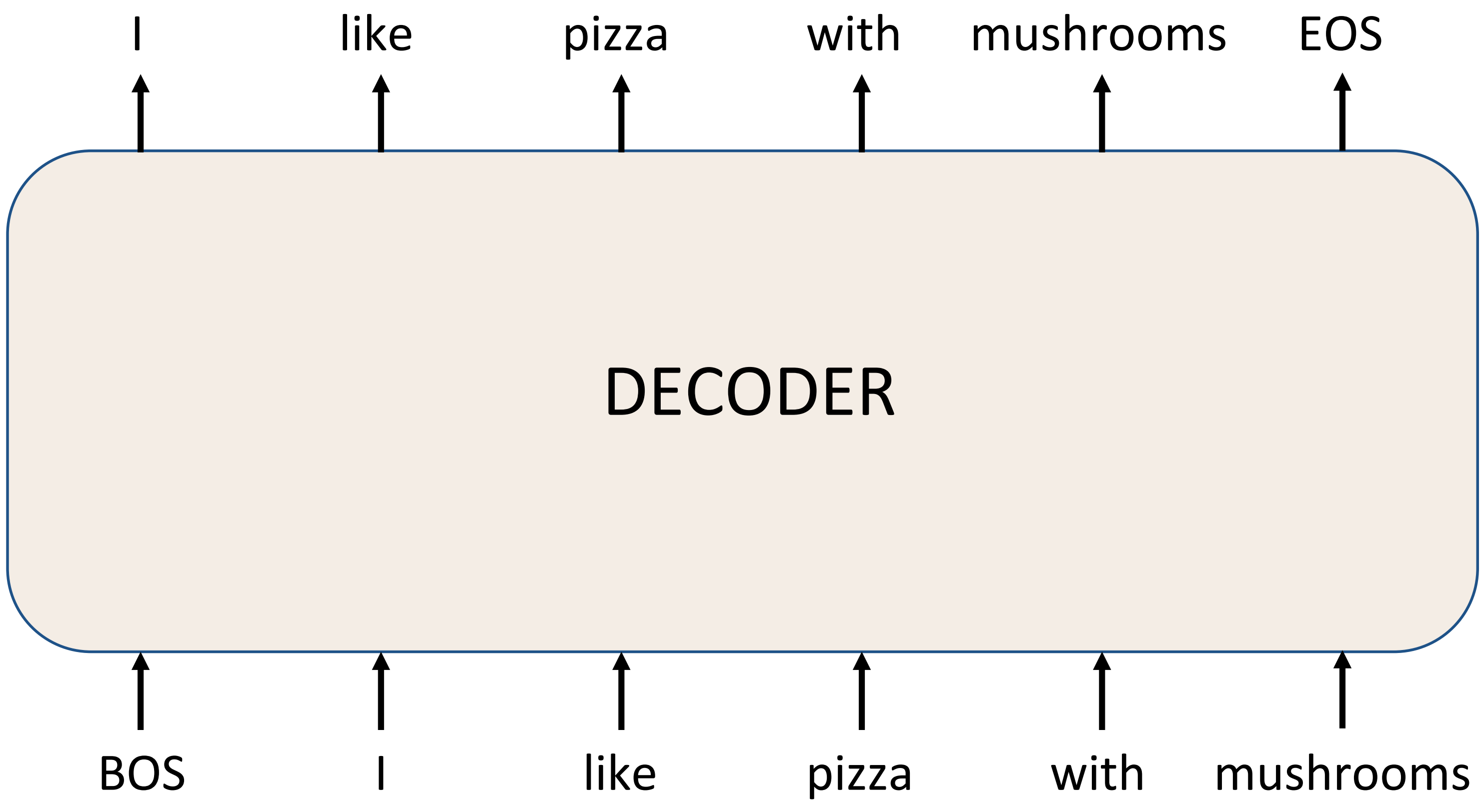
This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Autoregressive Language Model: Prediction (Cont.)



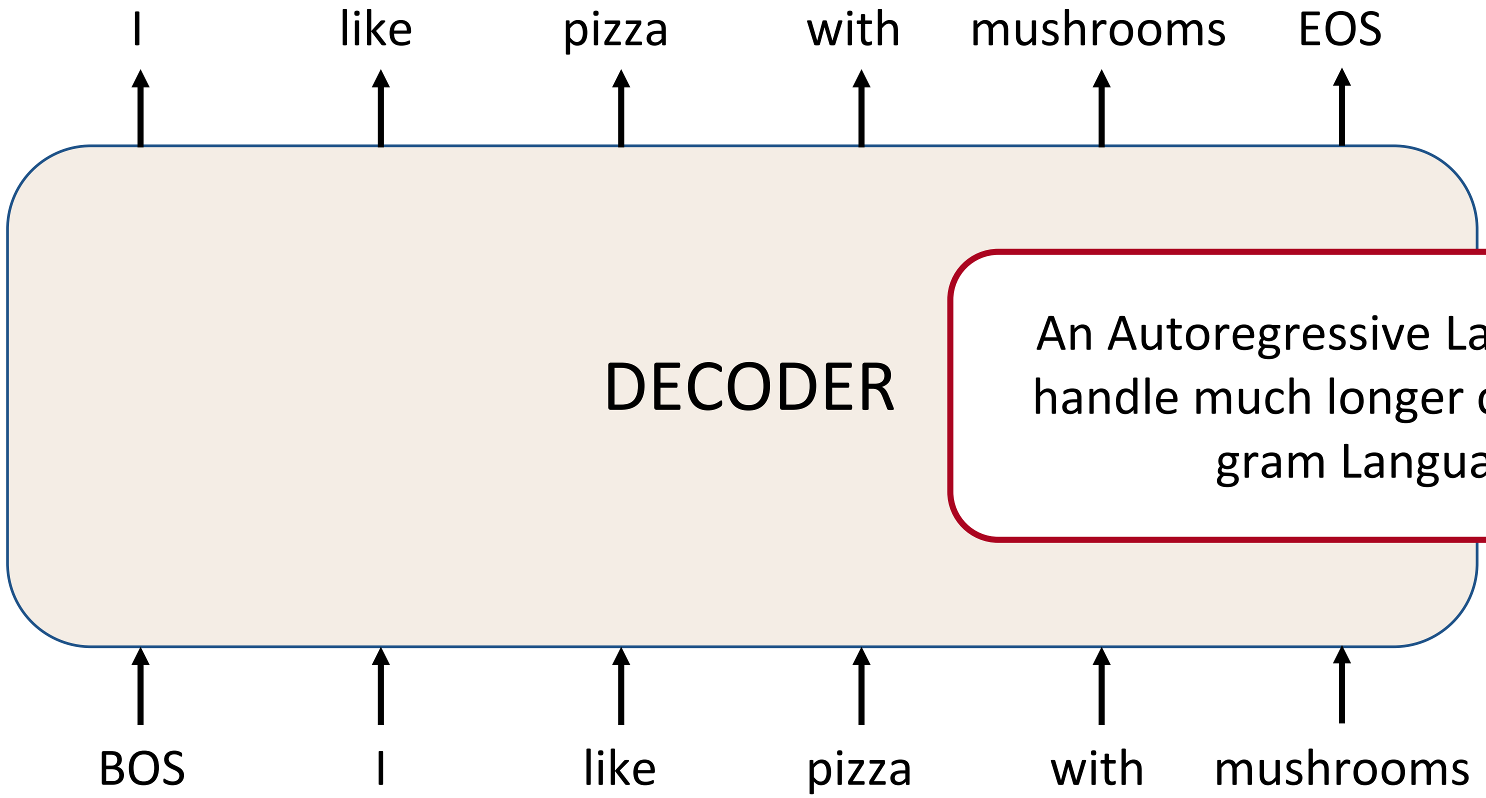
This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Autoregressive Language Model: Prediction (Cont.)



This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Autoregressive Language Model: Prediction (Cont.)



An Autoregressive Language Model can handle much longer contexts than an n-gram Language Model.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Masked Language Modeling

- Traditional Language Modeling only uses the previous context.
- Understanding language usually requires bi-directionality.
- Masked Language Modeling:
  - Predicts a word in any position (fill-in-the-blanks).

I like ______ with mushrooms

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Masked Language Modeling (Cont.)

- Traditional Language Modeling only uses the previous context.
- Understanding language usually requires bi-directionality.
- Masked Language Modeling:
  - Predicts a word in any position (fill-in-the-blanks).

I like [MASK] with mushrooms

We can use a special token to represent the words we want to predict. We call this “masking”.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Masked Language Modeling (Cont.)

- Traditional Language Modeling only uses the previous context.
- Understanding language usually requires bi-directionality.
- Masked Language Modeling:
  - Predicts a word in any position (fill-in-the-blanks).

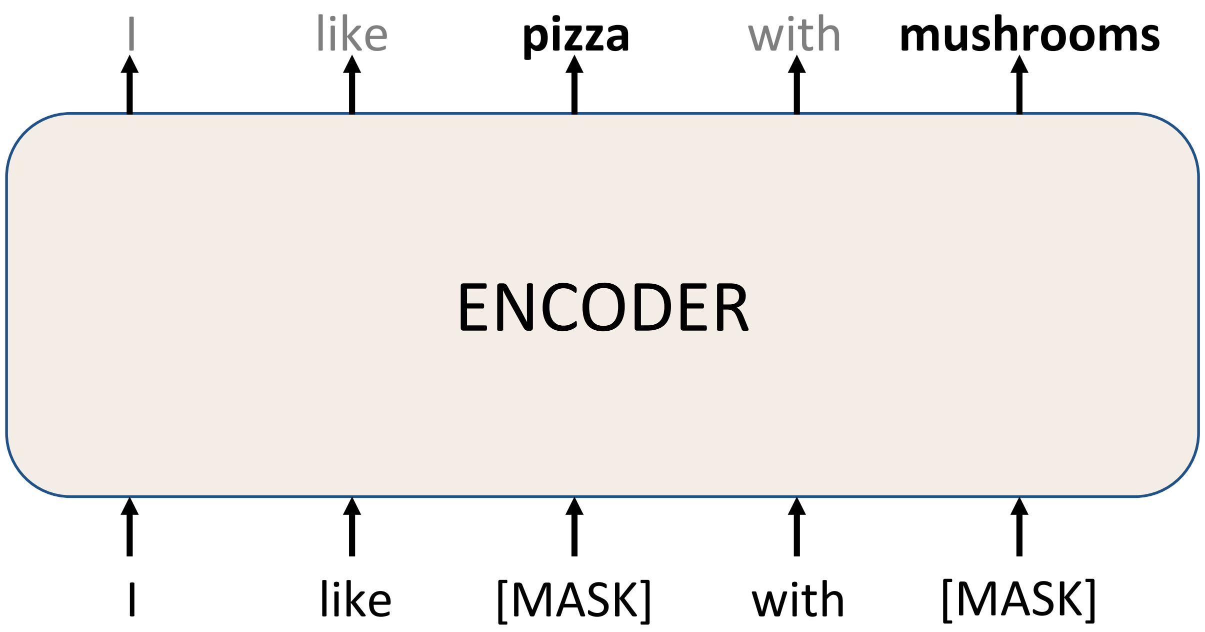
I like [MASK] with mushrooms

$$P(w \mid \text{I, like, [MASK], with, mushrooms})$$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Masked Language Modeling (Cont.)

- Traditional Language Modeling only uses the previous context.
- Understanding language usually requires bi-directionality.
- Masked Language Modeling:
  - Predicts a word in any position (fill-in-the-blanks).



We can use an Encoder with a sequential output for this task.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Transfer Learning

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Agenda

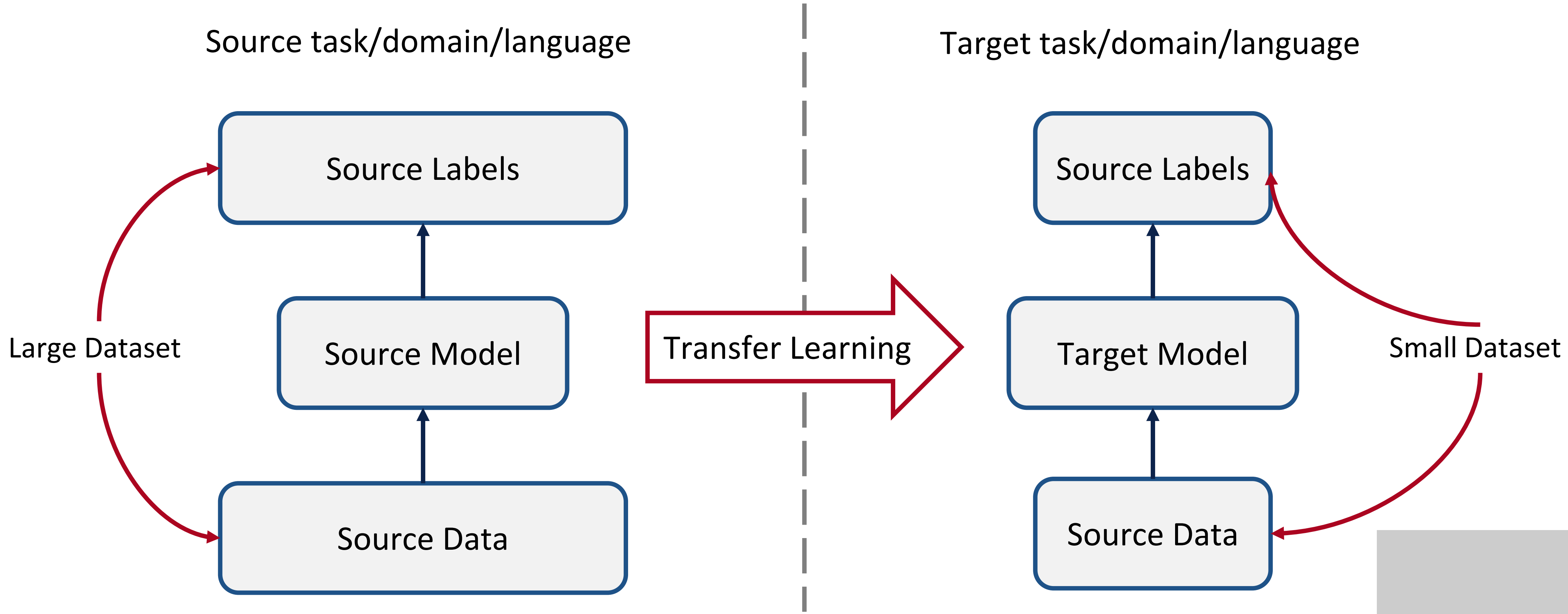
In this session, we will discuss:

- Introduction to Transfer Learning and its types
- Contextual Word Embedding
- Contextualized Word Vectors
- Pre-trained Bidirectional Language Models
- Universal Language Models

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Transfer Learning

Transfer knowledge from one task/domain/language to another.



This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Transfer Learning (Cont.)

- Types of Transfer Learning:
  - Domain Adaptation:
    - Same tasks, different domains
  - Cross-lingual Learning:
    - Same tasks, different languages
  - Multi-task Learning:
    - Different tasks learned jointly
  - Sequential Transfer Learning:
    - Different tasks learned sequentially

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Transfer Learning (Cont.)

- Types of Transfer Learning:
  - Domain Adaptation:
    - Same tasks, different domains
  - Cross-lingual Learning:
    - Same tasks, different languages
  - Multi-task Learning:
    - Different tasks learned jointly
  - Sequential Transfer Learning:
    - Different tasks learned sequentially

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Sequential Transfer Learning

- Approach:
  1. Select a source task with a large training set.
  2. Pre-Training:
     - Train a model in the source task.
  3. Reuse the model or part of the model as starting point for the target task.
     - Reuse the learned weights.
  4. Fine-tuning:
     - Continue training in the target task.

- If the source task relies on unlabeled data, we could use a huge corpus.
  - E.g., Word Embeddings

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Contextual Word Embeddings

- Word Embeddings are static.
  - A word is represented always with the same vector without considering the context.

The **bat** flew into the cave to escape the rain.

The player signed the **bat** for the young fan.

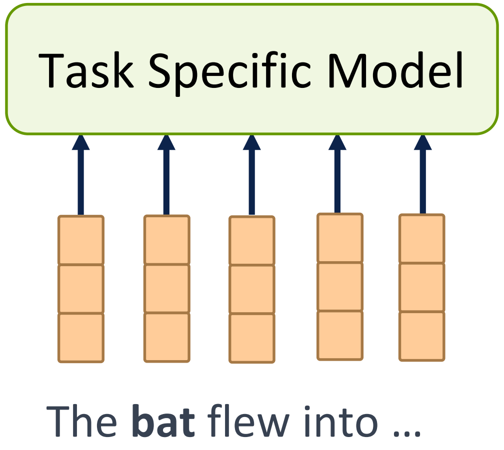
$$[0.27550674, 0.95053627, 0.04023079, \ldots]$$

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

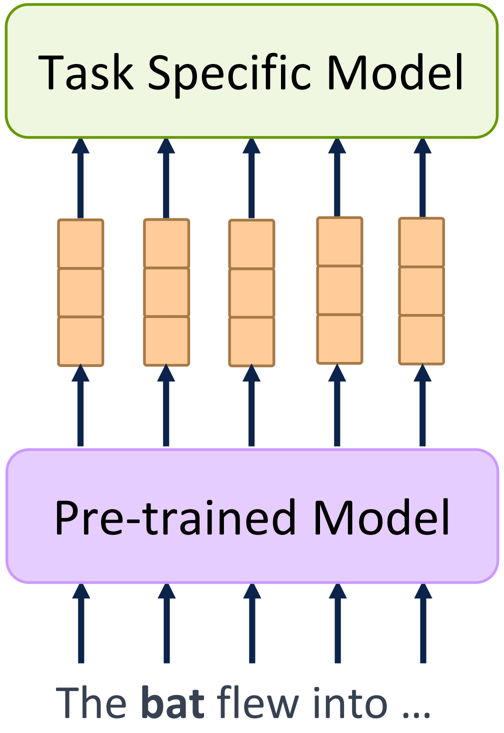
Contextual Word Embeddings (Cont.)

- Pre-trained models that learn to represent words according to context.

Static Word Embeddings



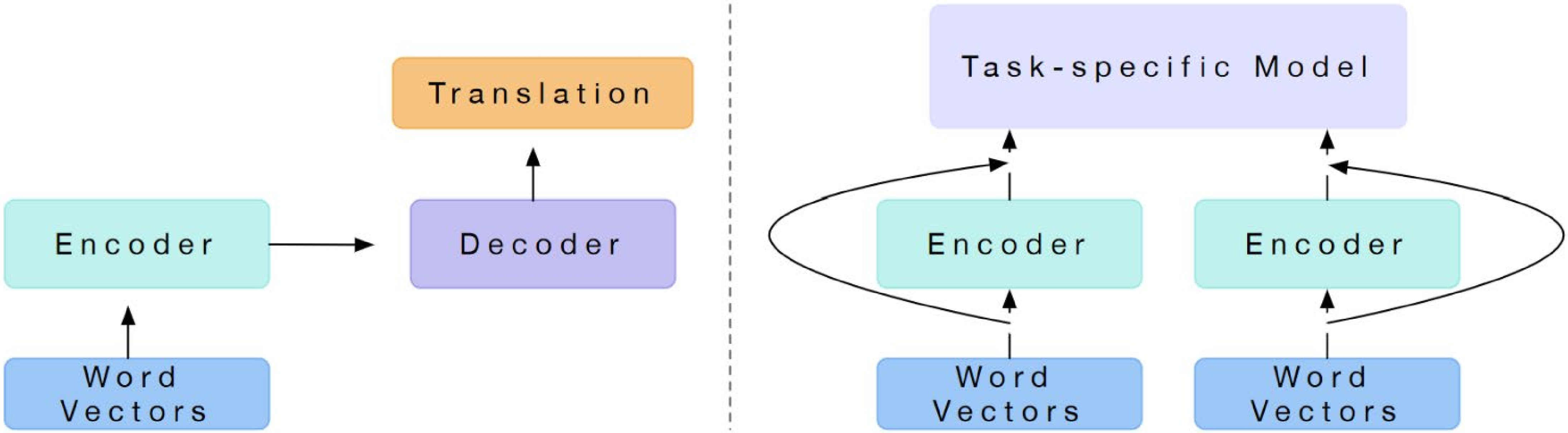
Contextual Word Embeddings



This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

CoVe: Contextualized Word Vectors

McCann et al., 2017

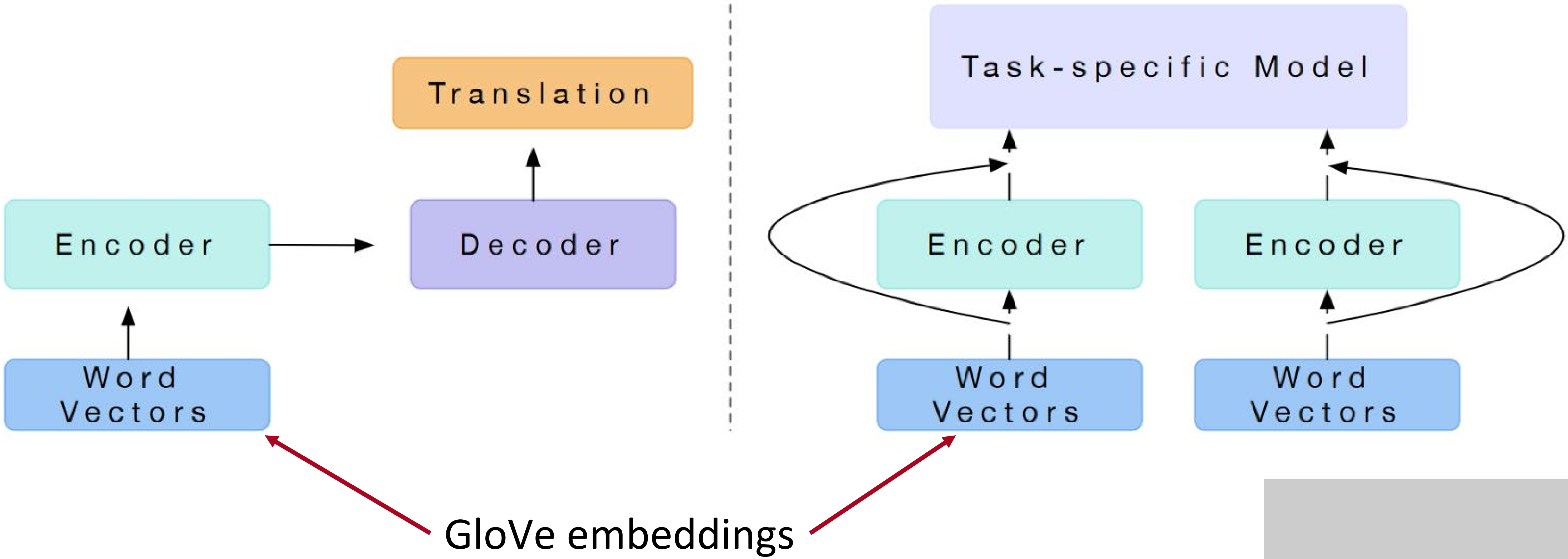


- Pre-train Encoder-Decoder model on machine translation.
- Use the Encoder to get contextualized embeddings on a different task.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

CoVe: Contextualized Word Vectors (Cont.)

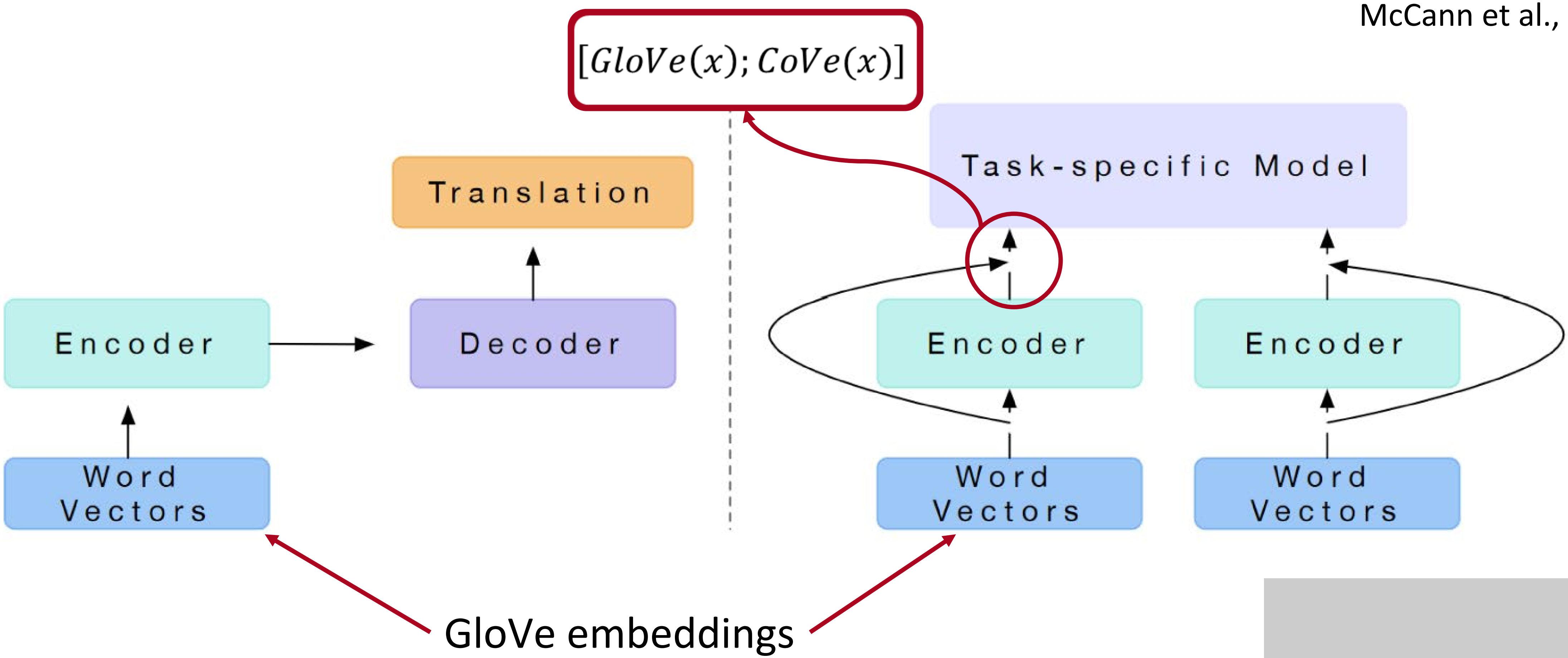
McCann et al., 2017



This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

CoVe: Contextualized Word Vectors (Cont.)

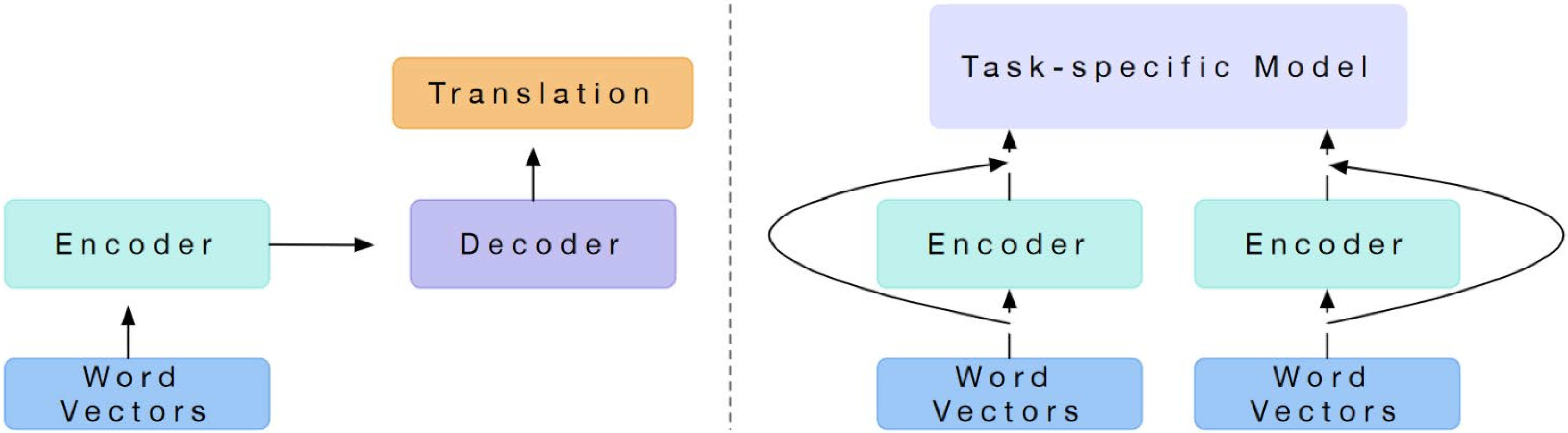
McCann et al., 2017



This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

CoVe: Contextualized Word Vectors (Cont.)

McCann et al., 2017



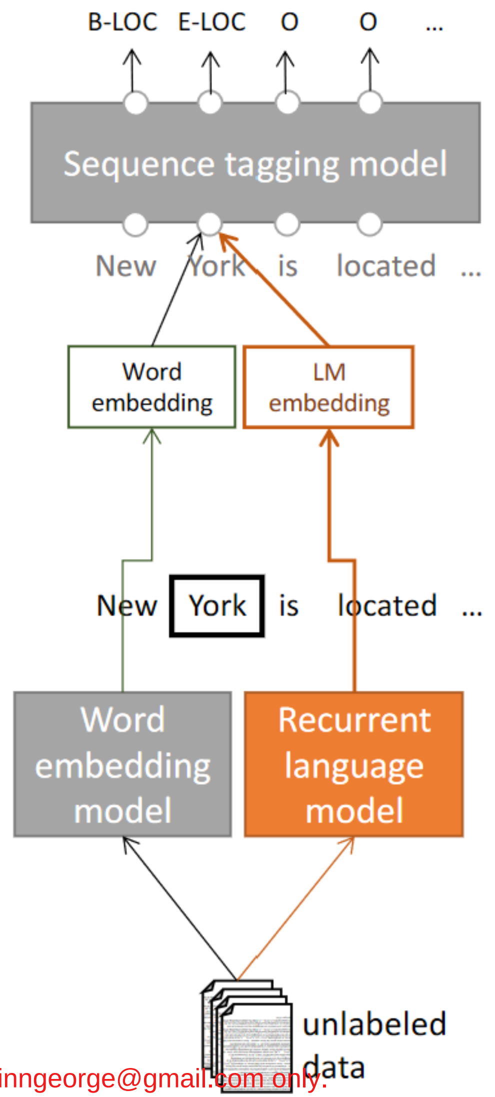
- Pre-training CoVe is limited by available datasets on the translation task.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Pre-trained Bidirectional Language Models

Peters et al., 2017

- Pre-train a Language Model on a large and unlabeled corpus.
- Bidirectional:
  - An LM to predict the following word.
  - An LM to predict the preceding word.
  - Concatenate both LM embeddings.
- Implementations:
  - TagLM: Language-model augmented sequence tagger
  - ELMo: Embeddings from Language Models

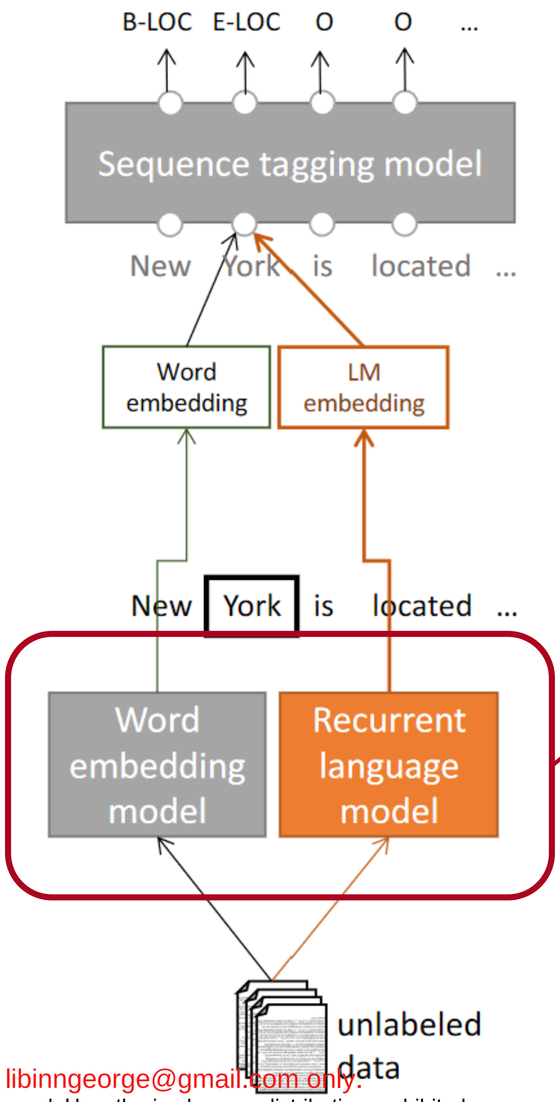


This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Pre-trained Bidirectional Language Models (Cont.)

- Pre-train a Language Model on a large and unlabeled corpus.
- Bidirectional:
  - An LM to predict the following word.
  - An LM to predict the preceding word.
  - Concatenate both LM embeddings.
- Implementations:
  - TagLM: Language-model augmented sequence tagger
  - ELMo: Embeddings from Language Models

Peters et al., 2017

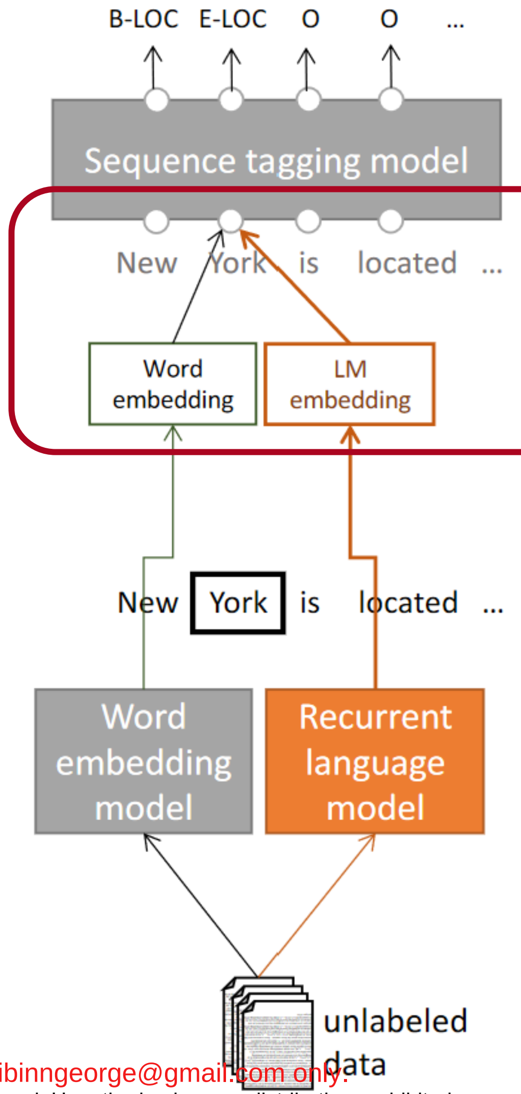


Pre-trained static word embeddings and a language model

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Pre-trained Bidirectional Language Models (Cont.)

- Pre-train a Language Model on a large and unlabeled corpus.
- Bidirectional:
  - An LM to predict the following word.
  - An LM to predict the preceding word.
  - Concatenate both LM embeddings.
- Implementations:
  - TagLM: Language-model augmented sequence tagger
  - ELMo: Embeddings from Language Models



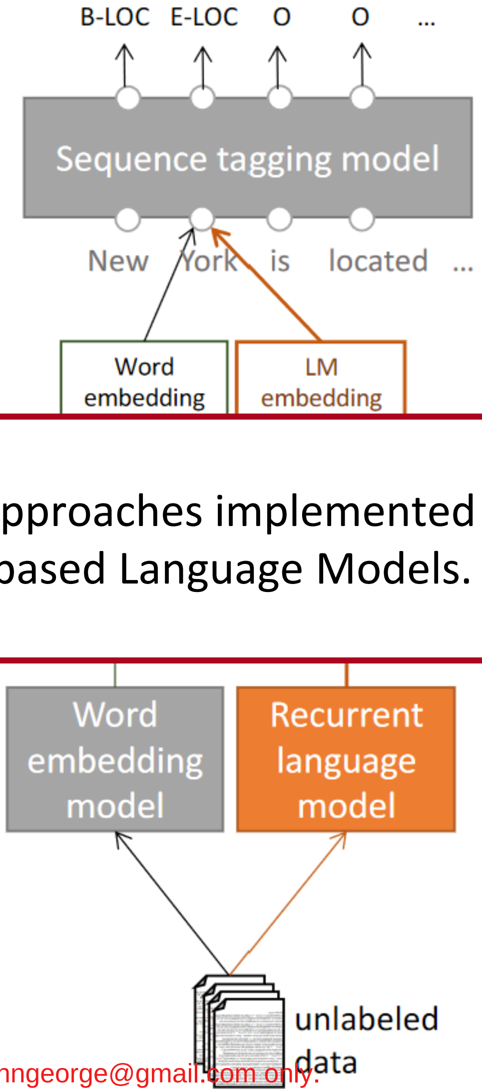
Concatenate static and contextualized embeddings for sequence labelling.

Peters et al., 2017

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Pre-trained Bidirectional Language Models (Cont.)

- Pre-train a Language Model on a large and unlabeled corpus.
- Bidirectional:
  - An LM to predict the following word.
  - An LM to predict the preceding word.
  - Concatenate both LM embeddings.
- Implementations:
  - TagLM: Language-model augmented sequence tagger
  - ELMo: Embeddings from Language Models



Both approaches implemented RNN-based Language Models.

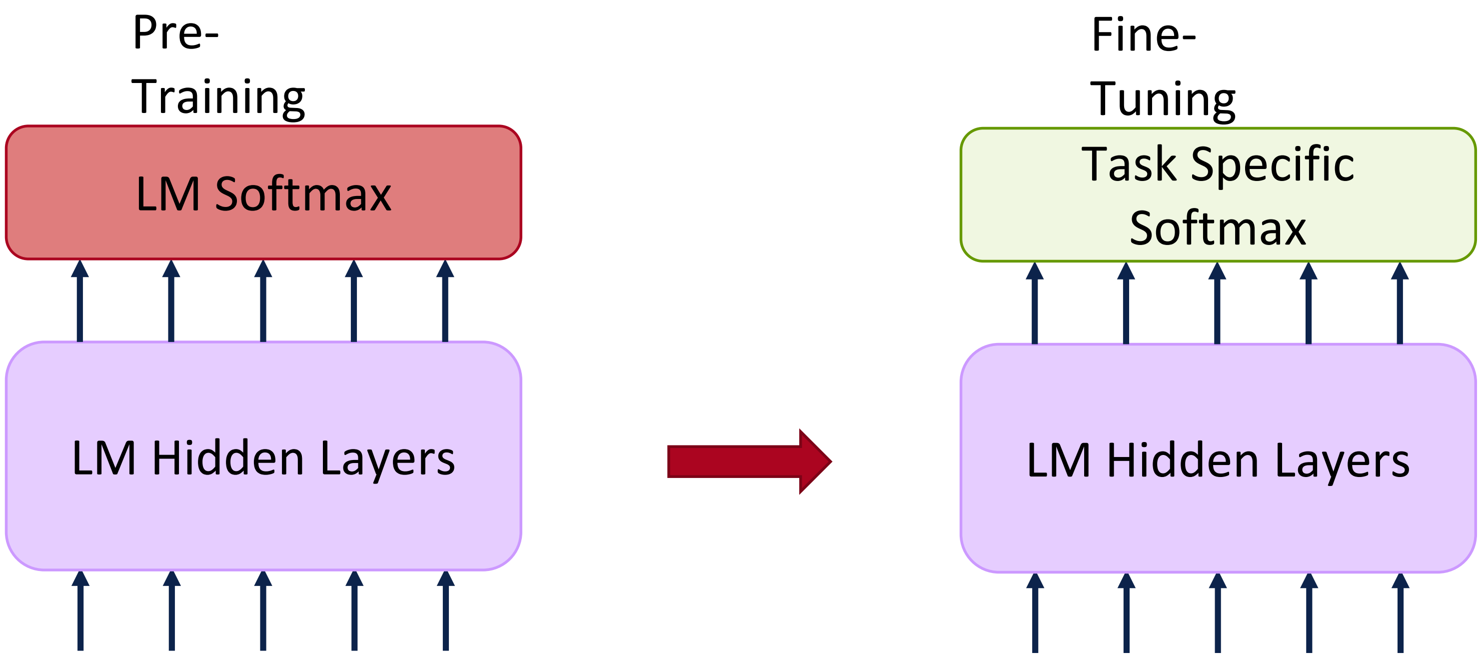
Peters et al., 2017

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

THE UNIVERSITY OF ARIZONA

Universal Language Models

- CoVe/TagLM/ELMO are used as an input layer.
  - For each specific task, we still have to train a whole specific model.
- Universal Language Model:

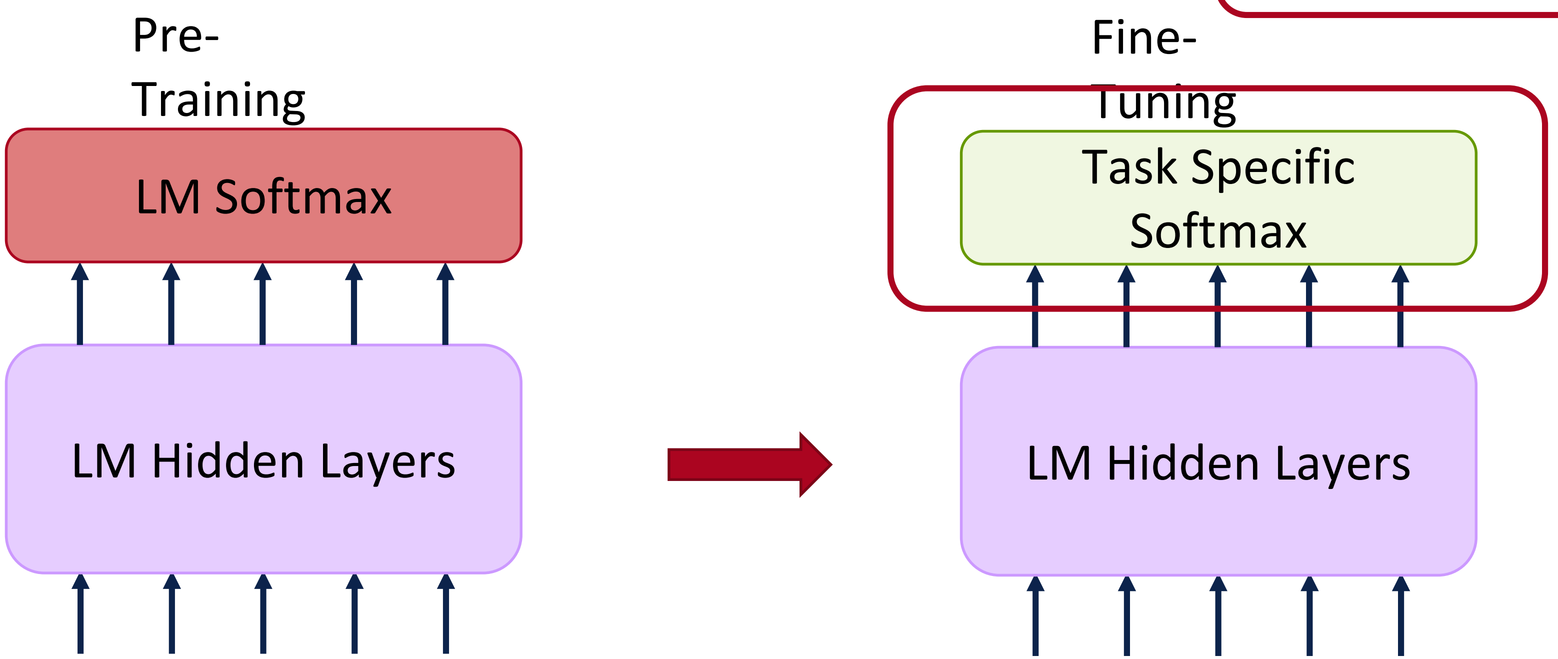


This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Universal Language Models (Cont.)

- CoVe/TagLM/ELMO are used as an input layer.
  - For each specific task, we still have to train a whole specific model.
- Universal Language Model:

We only need a new output layer for each task.



This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Agenda

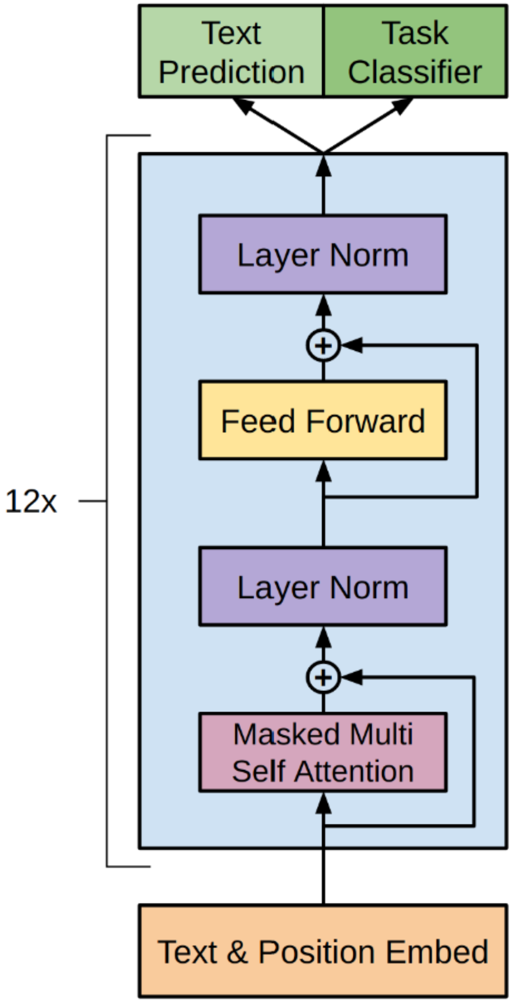
In this session, we will discuss:

- Introduction to GPT
- Fine-tuning a GPT model

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Generative Pre-Training for Language Understanding

- Autoregressive Language Model based on Transformer Decoder.

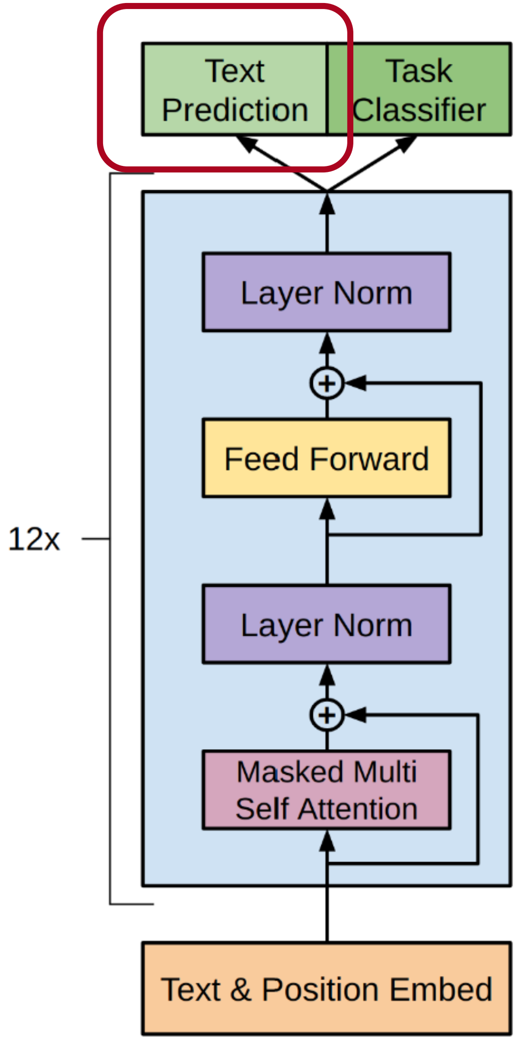


Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Generative Pre-Training for Language Understanding (Cont.)

- Autoregressive Language Model based on Transformer Decoder.
- During pre-training, a Linear+Softmax layer is trained to predict the next token.



Radford et al., 2018

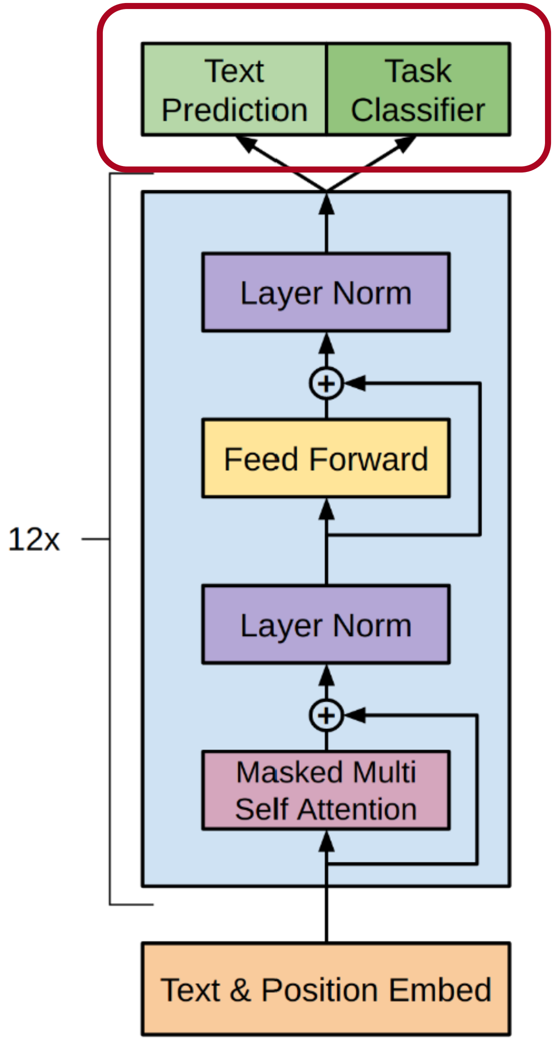
This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

GPT: Generative Pre-Training for Language Understanding (Cont.)

- Autoregressive Language Model based on Transformer Decoder.

- During pre-training, a Linear+Softmax layer is trained to predict the next token.

- During fine-tuning, a task-specific Linear+Softmax layer is added and trained.
  - GPT continues training the Text Prediction layer during fine-tuning.

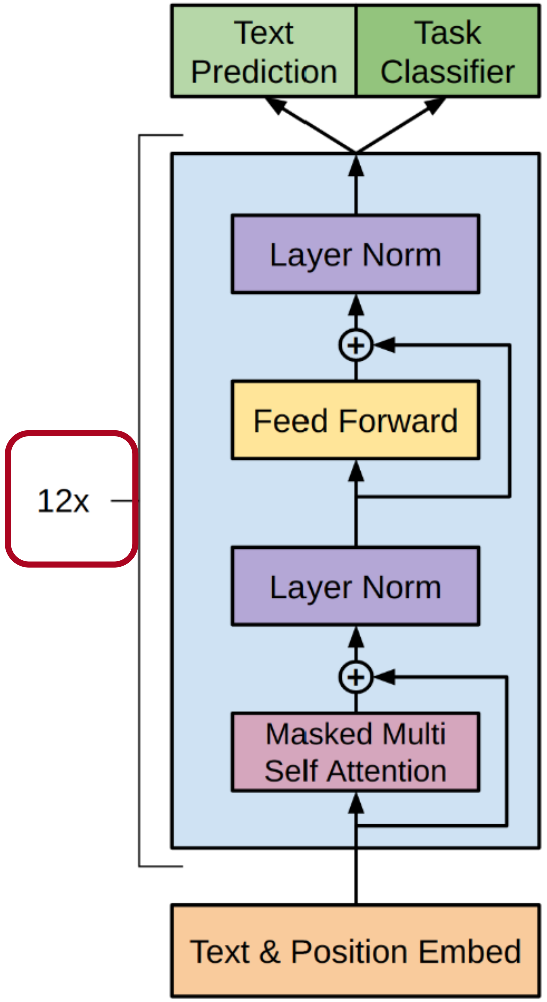


Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Generative Pre-Training for Language Understanding (Cont.)

- Autoregressive Language Model based on Transformer Decoder.
- During pre-training, a Linear+Softmax layer is trained to predict the next token.
- During fine-tuning, a task-specific Linear+Softmax layer is added and trained.
  - GPT continues training the Text Prediction layer during fine-tuning.
- 12 stacked Decoder blocks

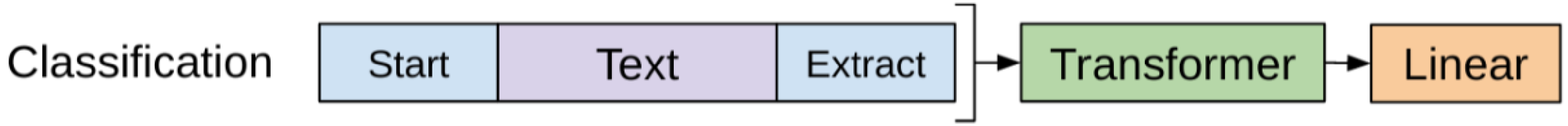


Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

GPT: Fine-tuning

- Input formats for fine-tuning on different tasks:

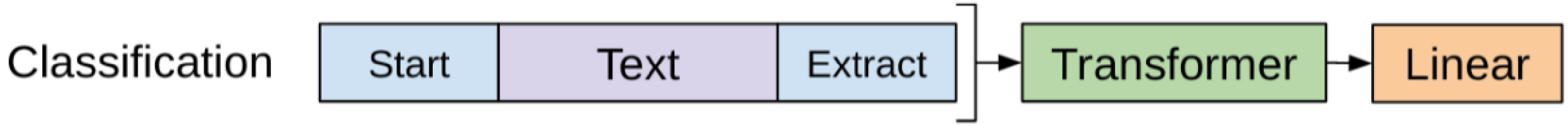


Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



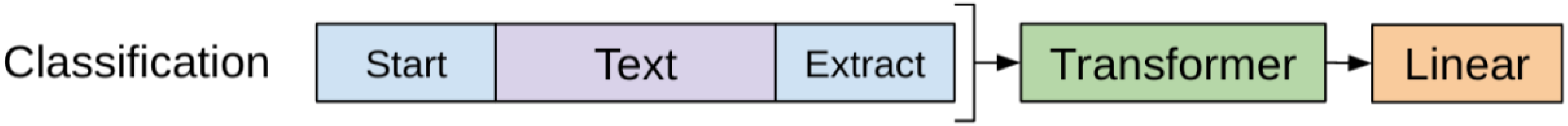
This input is a sequence of subwords.

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



This input is a sequence of sub ##words ##.

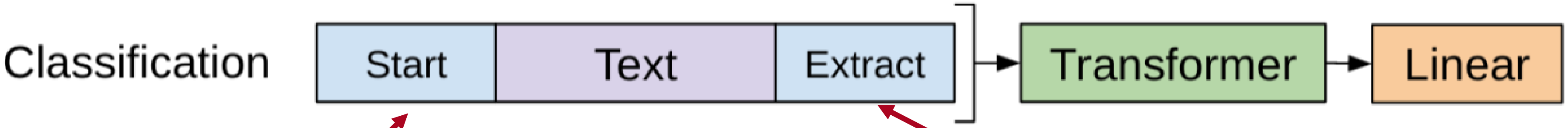
GPT uses subword tokenization.

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



<s> This input is a sequence of sub ##words ##. <e>

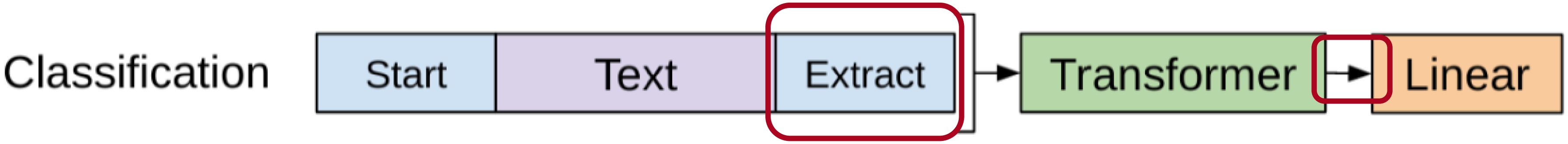
Randomly initialized start
and end/extract special tokens.

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



<s> This input is a sequence of sub ##words ##. <e>

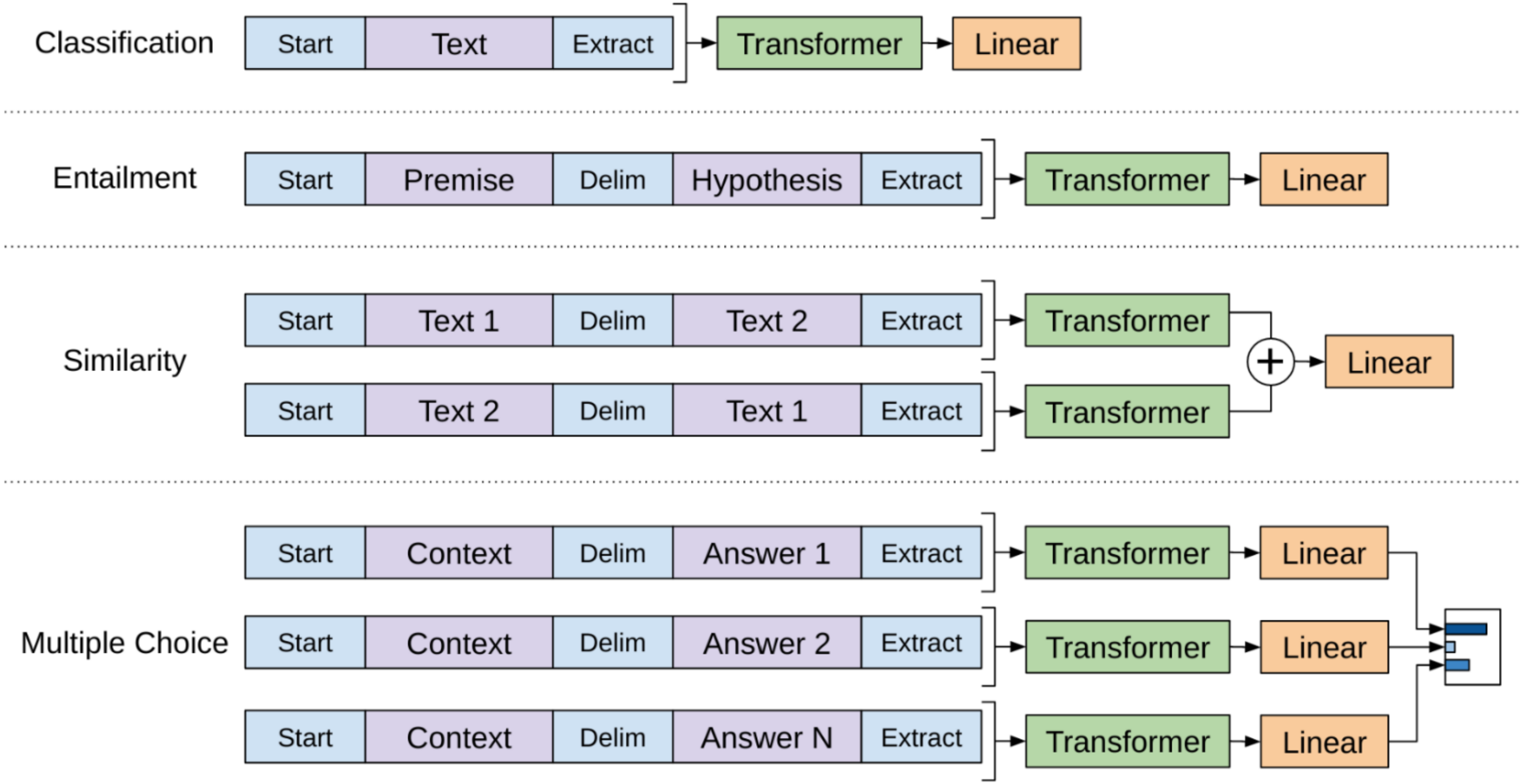
The input of the task classifier is the hidden state of the end/extract token from the last block.

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:

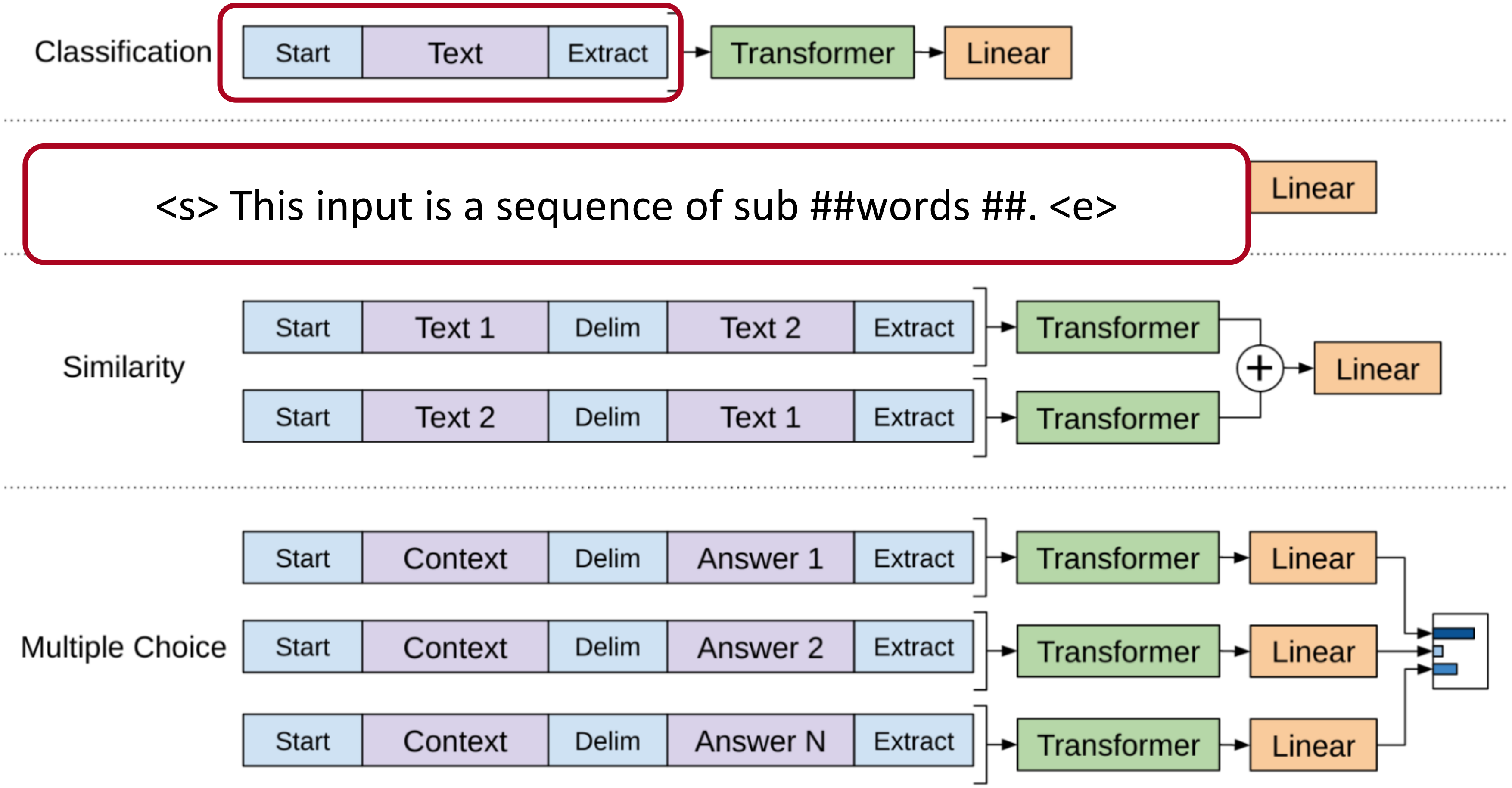


Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



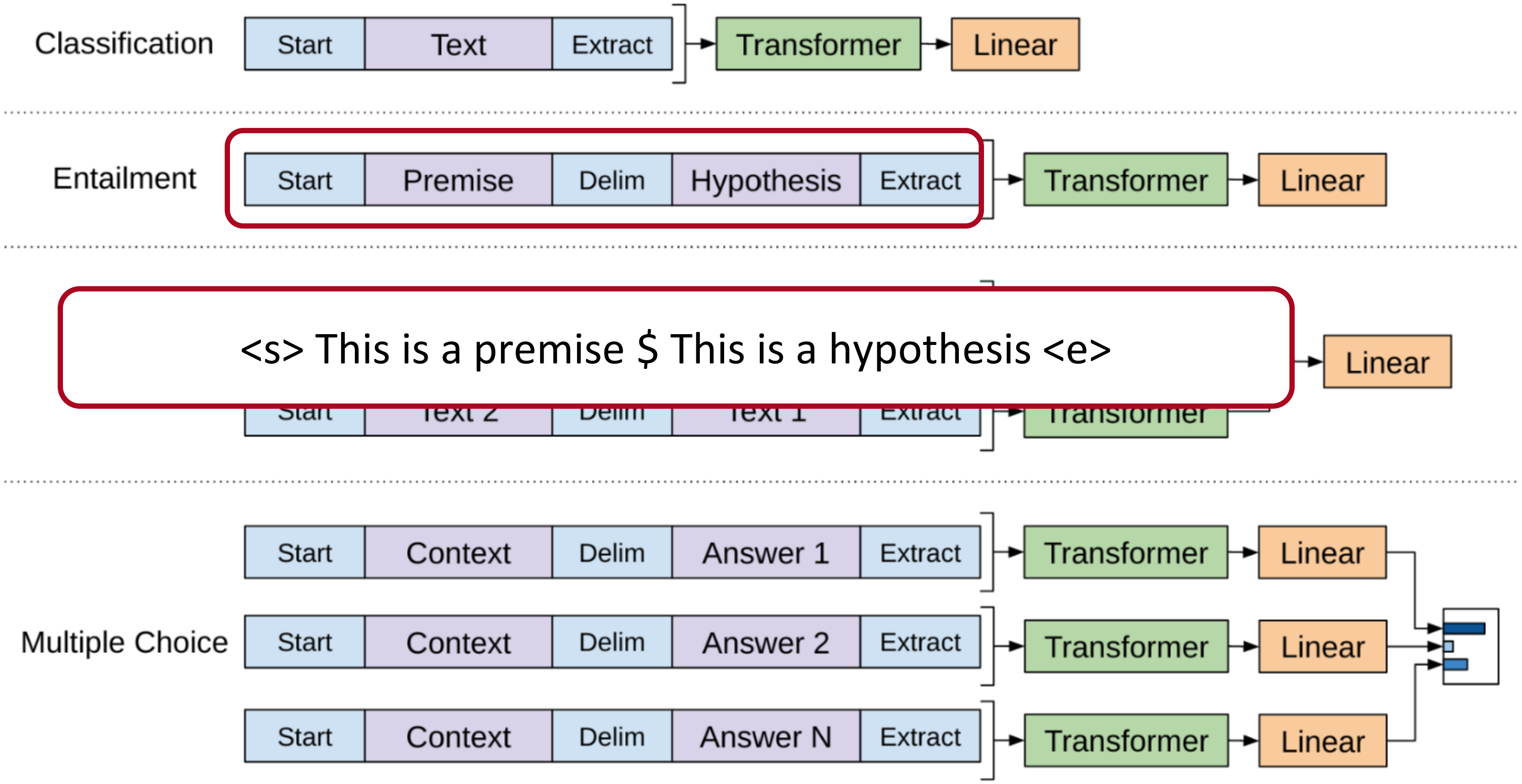
<s> This input is a sequence of sub ##words ##. <e>

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



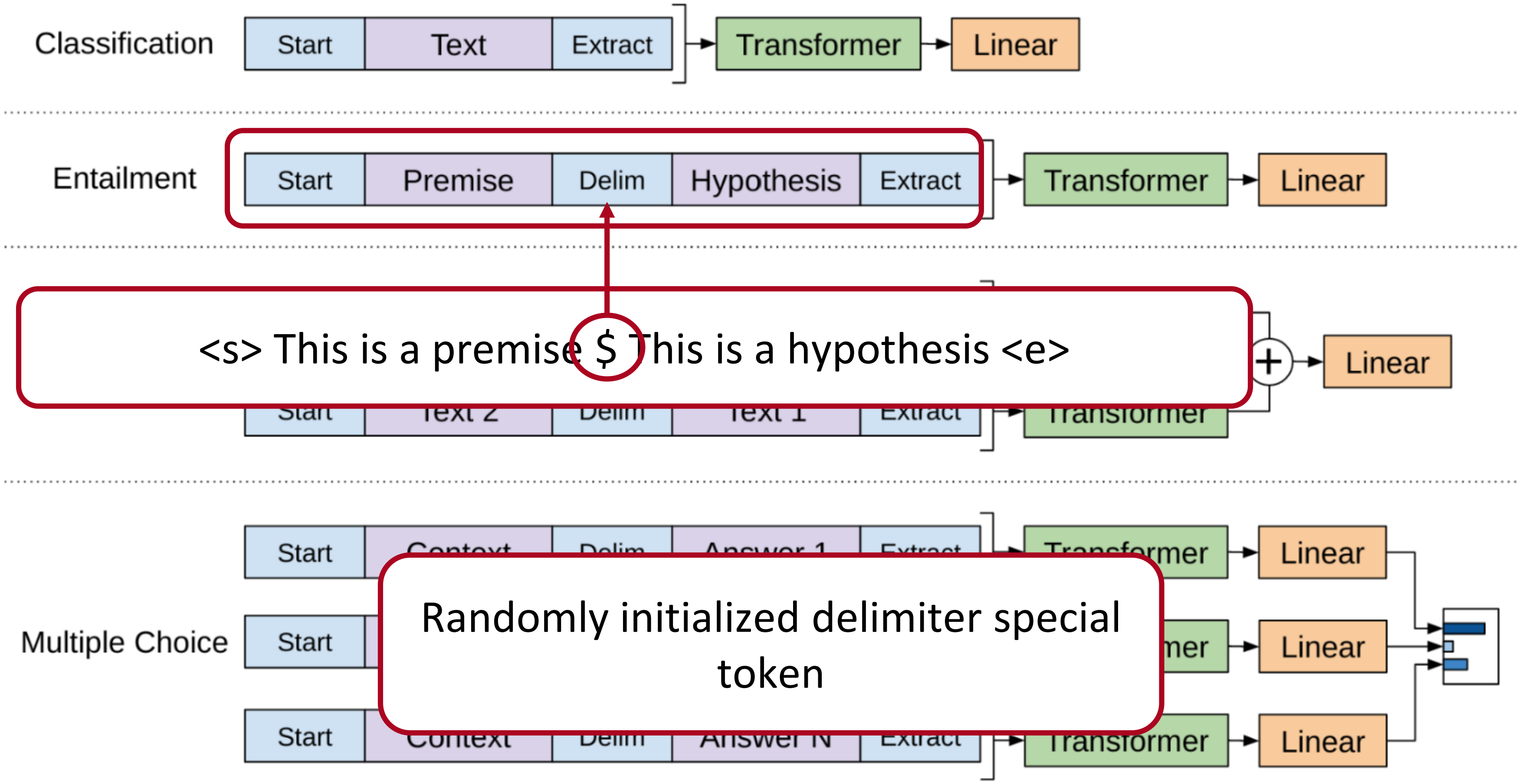
<s> This is a premise $ This is a hypothesis <e>

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



<s> This is a premise $ This is a hypothesis <e>

Randomly initialized delimiter special token

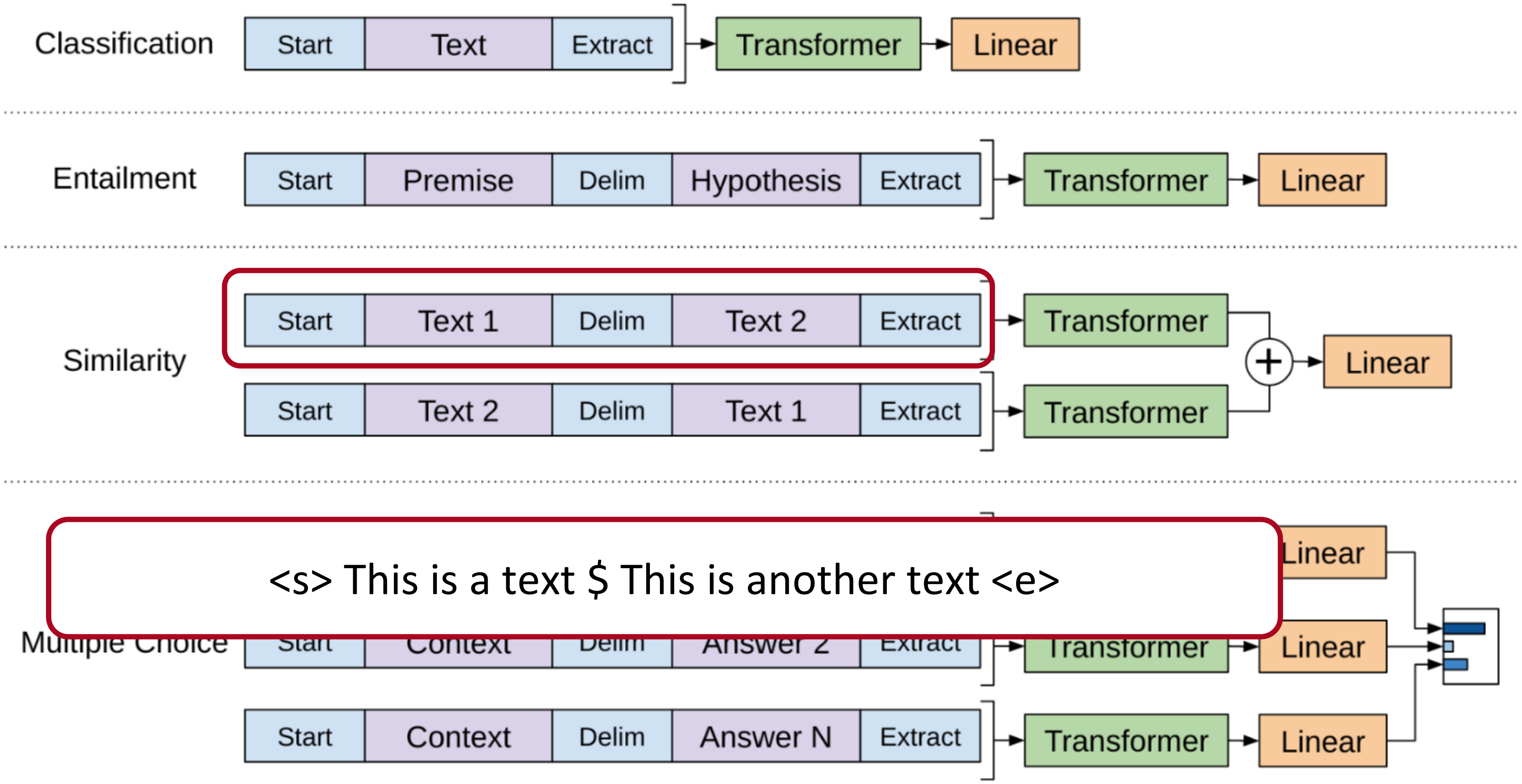
Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

THE UNIVERSITY OF ARIZONA

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



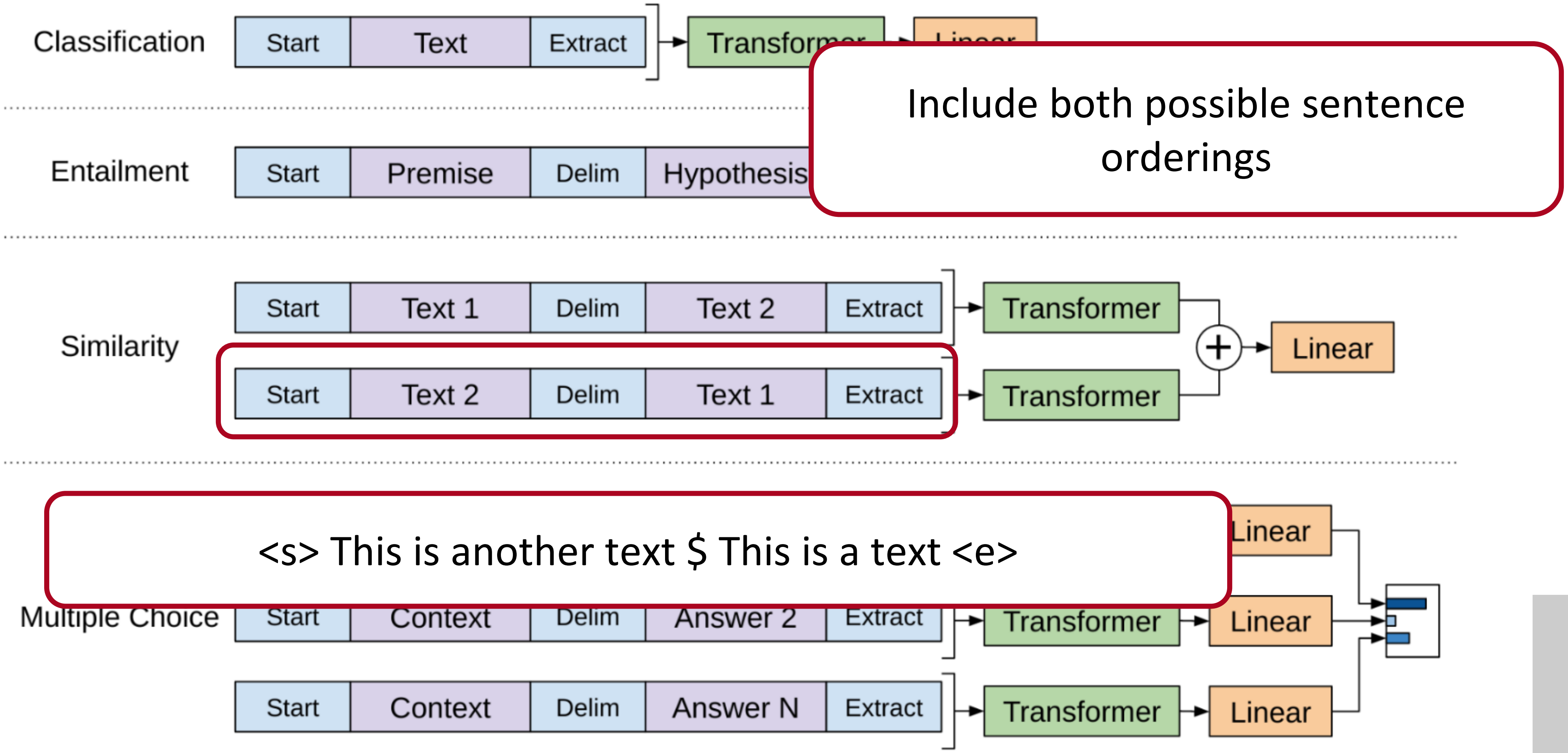
<s> This is a text $ This is another text <e>

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:

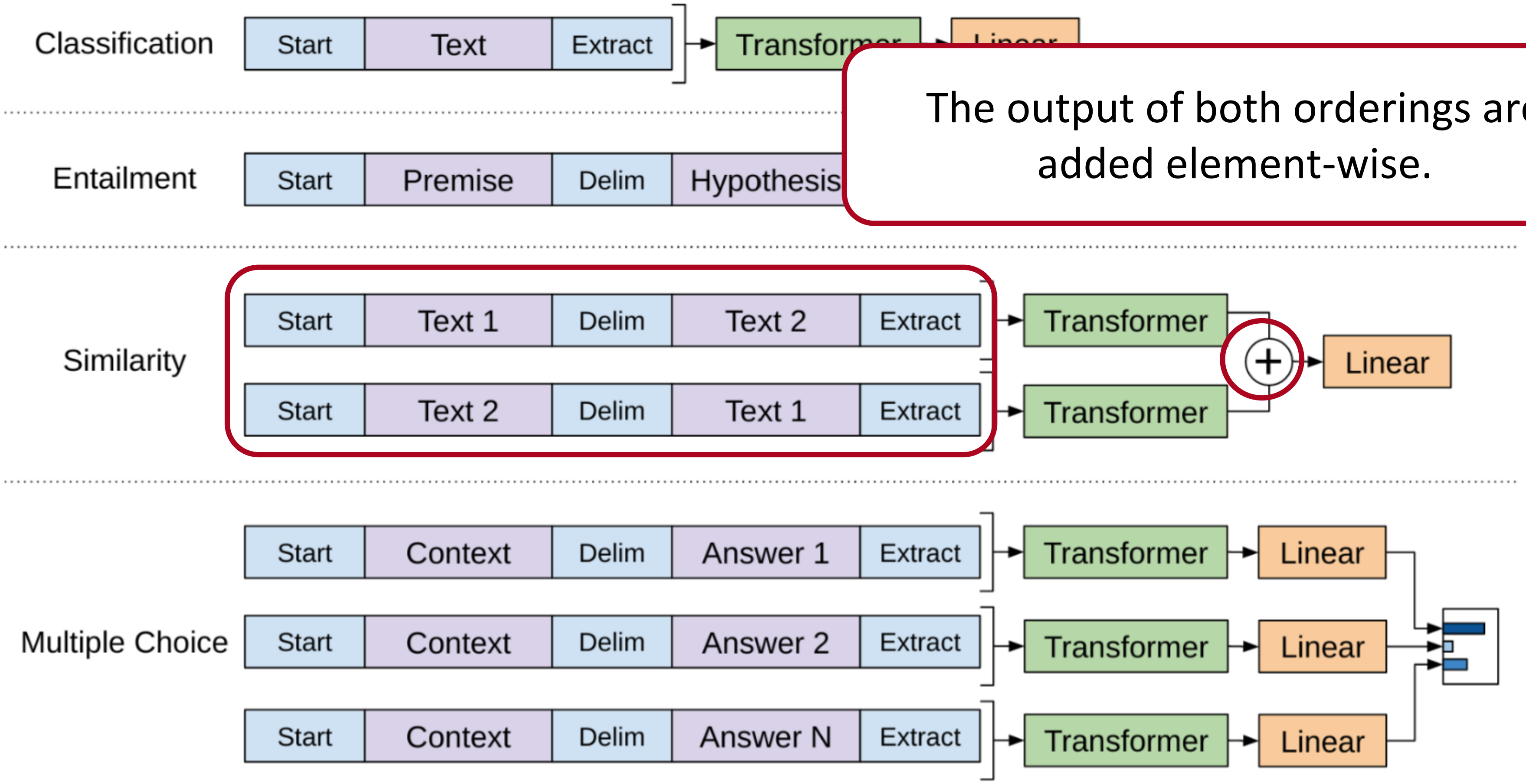


Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



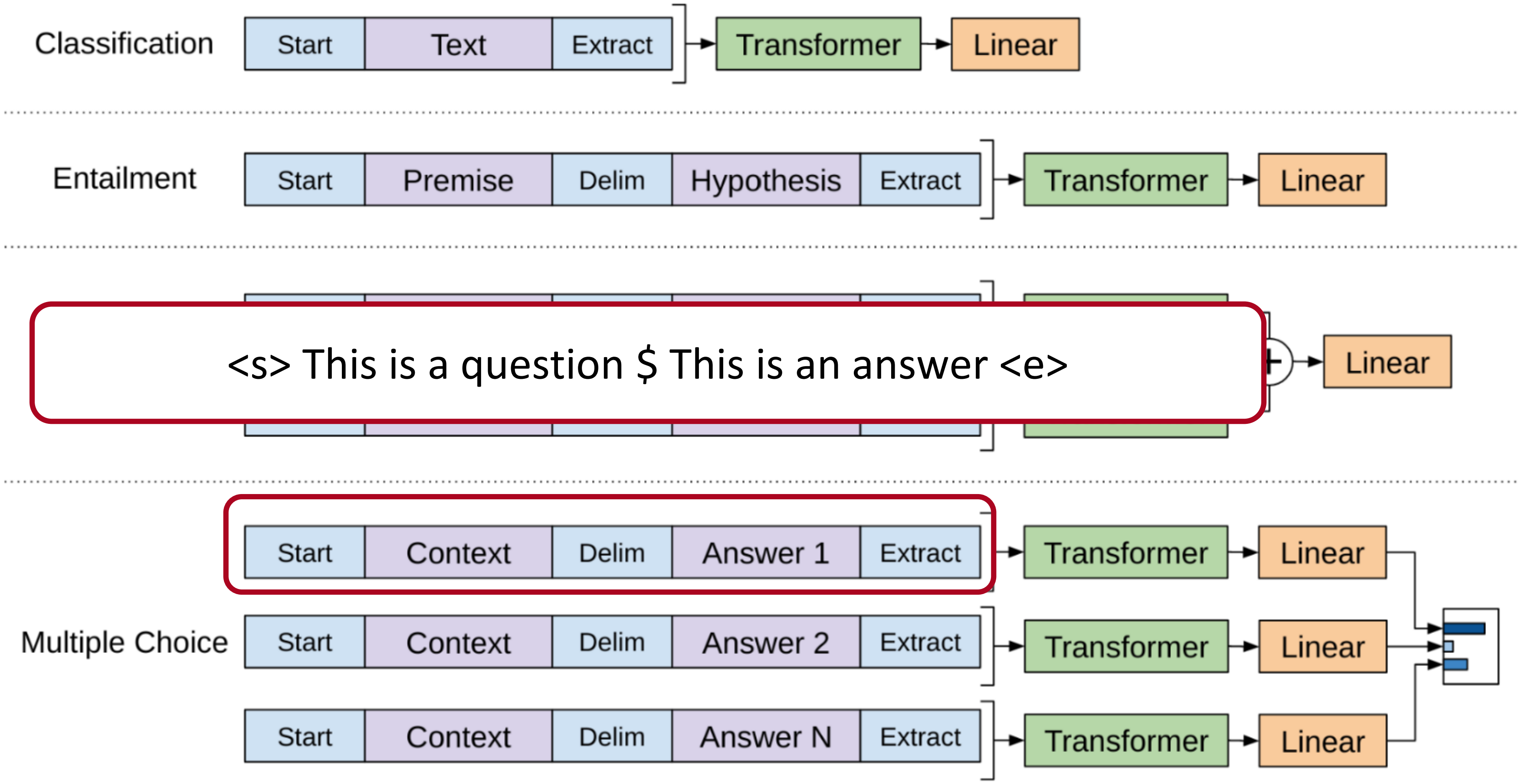
The output of both orderings are added element-wise.

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:

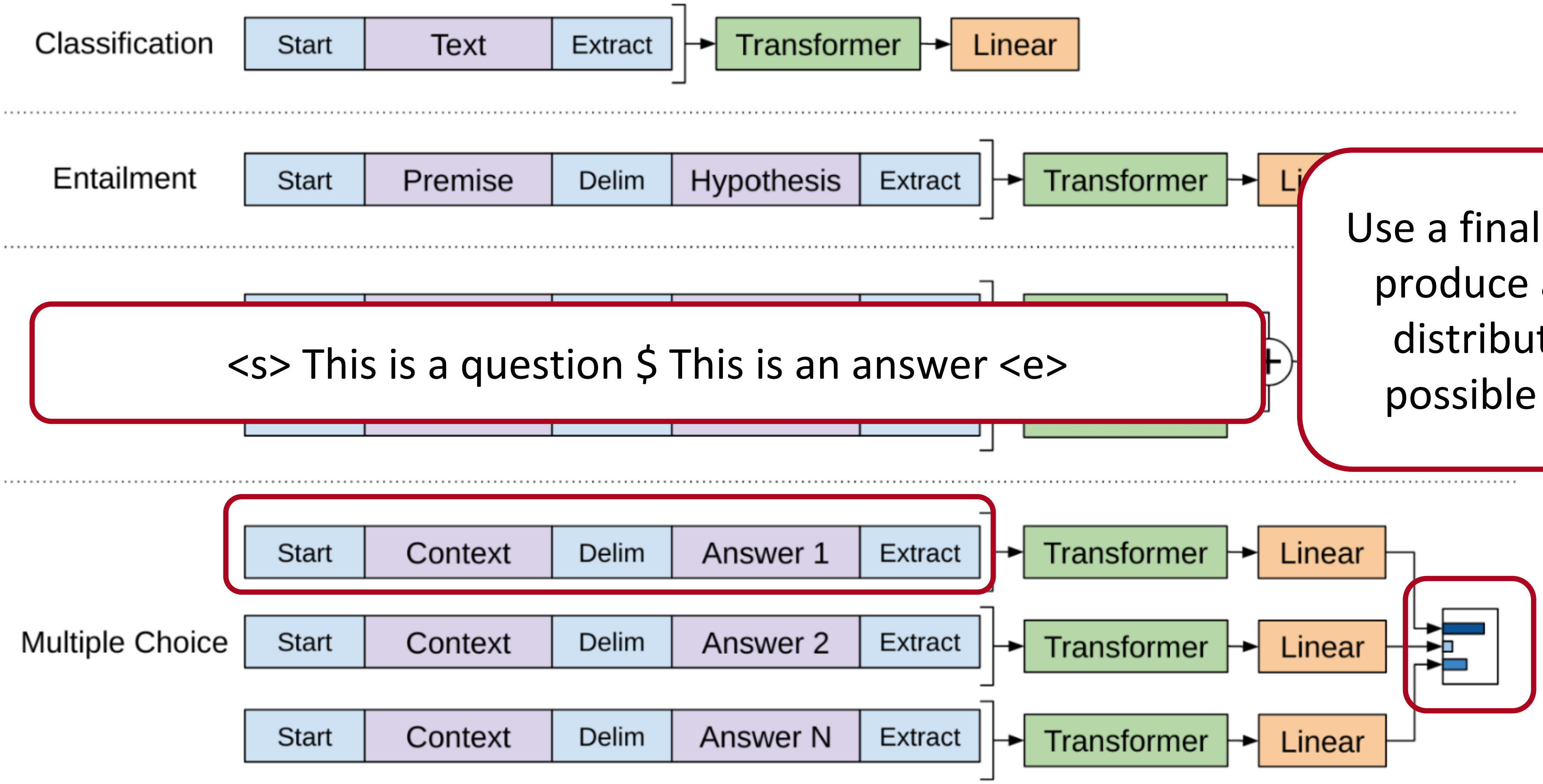


Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

GPT: Fine-tuning (Cont.)

- Input formats for fine-tuning on different tasks:



<s> This is a question $ This is an answer <e>

Use a final softmax to produce an output distribution over possible answers.

Radford et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

BERT

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Agenda

In this session, we will discuss:

- An Introduction to BERT

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Bidirectional Encoder Representations from Transformers

- An Autoregressive Language Model is unidirectional:
  - Pre-trained for the next token prediction.

- BERT is bidirectional:
  - Pre-train Transformer Encoder jointly on:
    - Masked Language Modeling;
    - Next Sentence Prediction.

Is Next: [CLS] I [MASK] to the restaurant [SEP] I ordered a [MASK] of noodles [SEP]

Is Not Next: [CLS] I [MASK] to the restaurant [SEP] A motor ##bike is a type of [MASK] [SEP]

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Bidirectional Encoder Representations from Transformers

- An Autoregressive Language Model is unidirectional:
  - Pre-trained for the next token prediction.

- BERT is bidirectional:
  - Pre-train Transformer Encoder jointly on:
    - Masked Language Modeling;
    - Next Sentence Prediction.

Subword tokenization

Is Next: [CLS] I [MASK] to the restaurant [SEP] I ordered a [MASK] of noodles [SEP]

Is Not Next: [CLS] I [MASK] to the restaurant [SEP] A motor ##bike is a type of [MASK] [SEP]

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Bidirectional Encoder Representations from Transformers

- An Autoregressive Language Model is unidirectional:
  - Pre-trained for the next token prediction.

- BERT is bidirectional:
  - Pre-train Transformer Encoder jointly on:
    - Masked Language Modeling;
    - Next Sentence Prediction.

The input for pre-training are 2 concatenated sentences.

Is Next: [CLS] [MASK] to the restaurant [SEP] I ordered a [MASK] of noodles [SEP]

Is Not Next: [CLS] I [MASK] to the restaurant [SEP] A motor ##bike is a type of [MASK] [SEP]

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

THE UNIVERSITY OF ARIZONA

BERT: Bidirectional Encoder Representations from Transformers

- An Autoregressive Language Model is unidirectional:
  - Pre-trained for the next token prediction.

- BERT is bidirectional:
  - Pre-train Transformer Encoder jointly on:
    - Masked Language Modeling;
    - Next Sentence Prediction.

Sentence separator special token.

Is Next: [CLS] I [MASK] to the restaurant [SEP] I ordered a [MASK] of noodles [SEP]

Is Not Next: [CLS] I [MASK] to the restaurant [SEP] A motor ##bike is a type of [MASK] [SEP]

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Bidirectional Encoder Representations from Transformers

- An Autoregressive Language Model is unidirectional:
  - Pre-trained for the next token prediction.

- BERT is bidirectional:
  - Pre-train Transformer Encoder jointly on:
    - Masked Language Modeling;
    - Next Sentence Prediction.

Class special token

Is Next: [CLS] I [MASK] to the restaurant [SEP] I ordered a [MASK] of noodles [SEP]

Is Not Next: [CLS] I [MASK] to the restaurant [SEP] A motor ##bike is a type of [MASK] [SEP]

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Bidirectional Encoder Representations from Transformers

- Masked Language Modeling:
  - Learn to predict the masked token.
  - Mask 15% of the tokens.
    - 80% of the time, use a special token [MASK].
    - 10% of the time, use a random token.
    - 10% of the time, do not replace the token.

Is Next: [CLS] I [MASK] to the restaurant [SEP] I ordered a [MASK] of noodles [SEP]

Is Not Next: [CLS] I [MASK] to the restaurant [SEP] A motor ##bike is a type of [MASK] [SEP]

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Bidirectional Encoder Representations from Transformers

- Masked Language Modeling:
  - Learn to predict the masked token.
  - Mask 15% of the tokens.
    - 80% of the time, use a special token [MASK].
    - 10% of the time, use a random token.
    - 10% of the time, do not replace the token.

[MASK] token does not appear during fine-tuning.

These masking alternatives help to mitigate this mismatch between pre-training and fine-tuning.

Is Next: [CLS] I [MASK] to the restaurant [SEP] I ordered a [MASK] of noodles [SEP]

Is Not Next: [CLS] I [MASK] to the restaurant [SEP] A motor ##bike is a type of [MASK] [SEP]

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Bidirectional Encoder Representations from Transformers

- Next Sentence Prediction:
  - Binary classification task:
    - 50% of sentence pairs contain consecutive sentences extracted from training.
    - 50% of sentence pairs contain sentences randomly picked.
  - Help the model to understand relationships between sentences.

Is Next:	[CLS] I [MASK] to the restaurant [SEP] I ordered a [MASK] of noodles [SEP]
Is Not Next:	[CLS] I [MASK] to the restaurant [SEP] A motor ##bike is a type of [MASK] [SEP]

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT- Pre-training and Fine Tuning

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Agenda

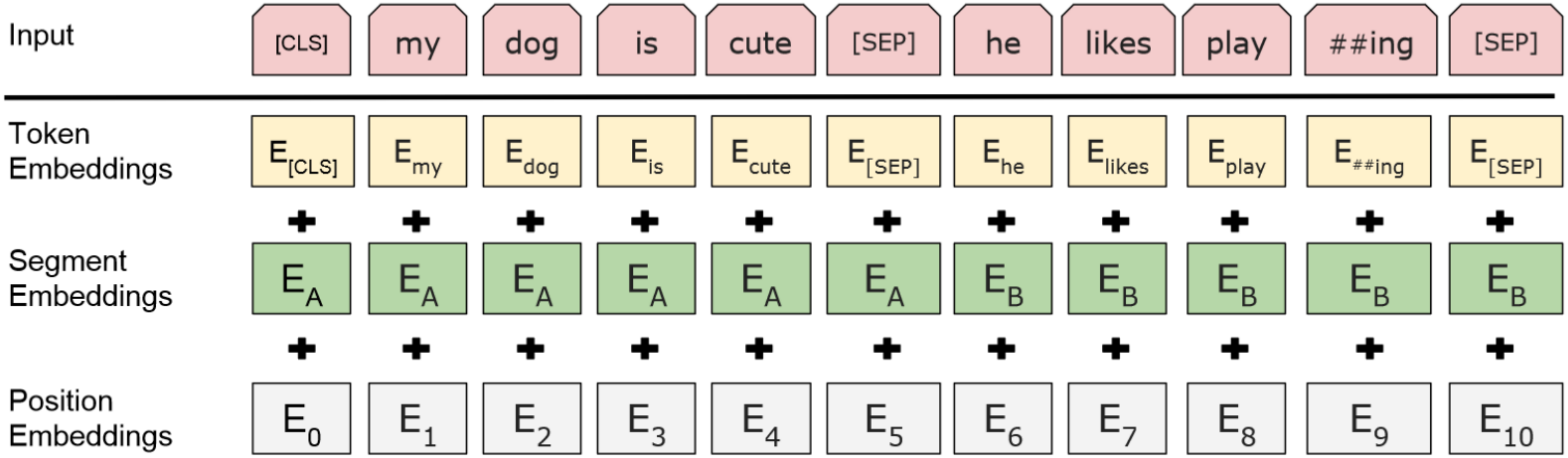
In this session, we will discuss:

- Pre-training of BERT
- Fine-tuning of BERT

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Input Representation

Sum of the token embeddings, the segmentation embeddings, and the position embeddings.

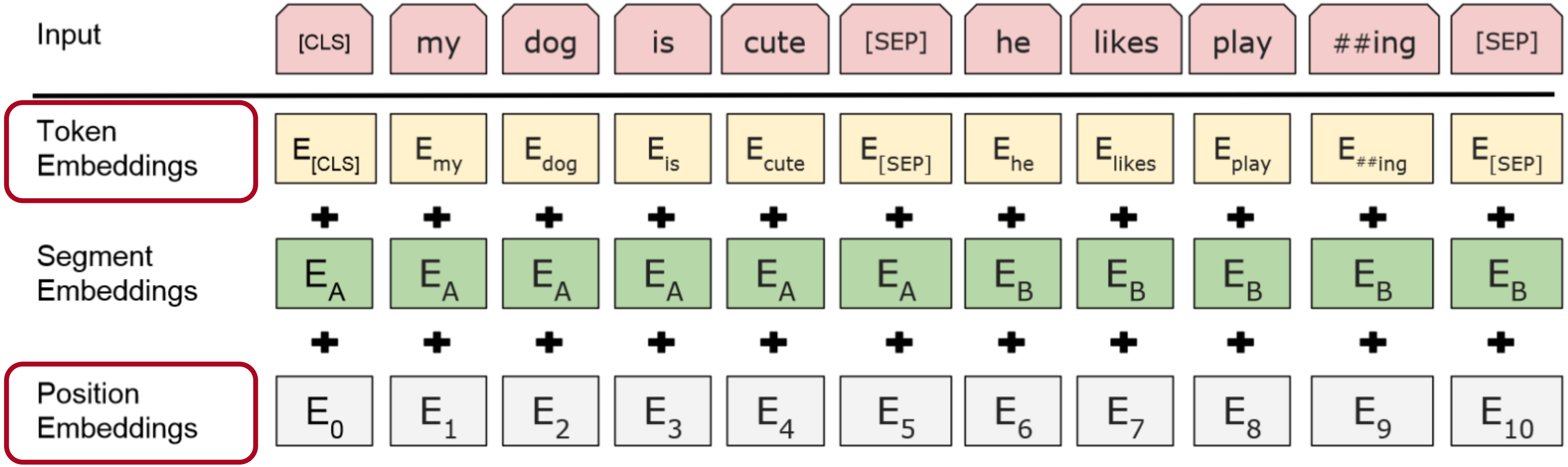


Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

BERT: Input Representation (Cont.)

Sum of the token embeddings, the segmentation embeddings, and the position embeddings.



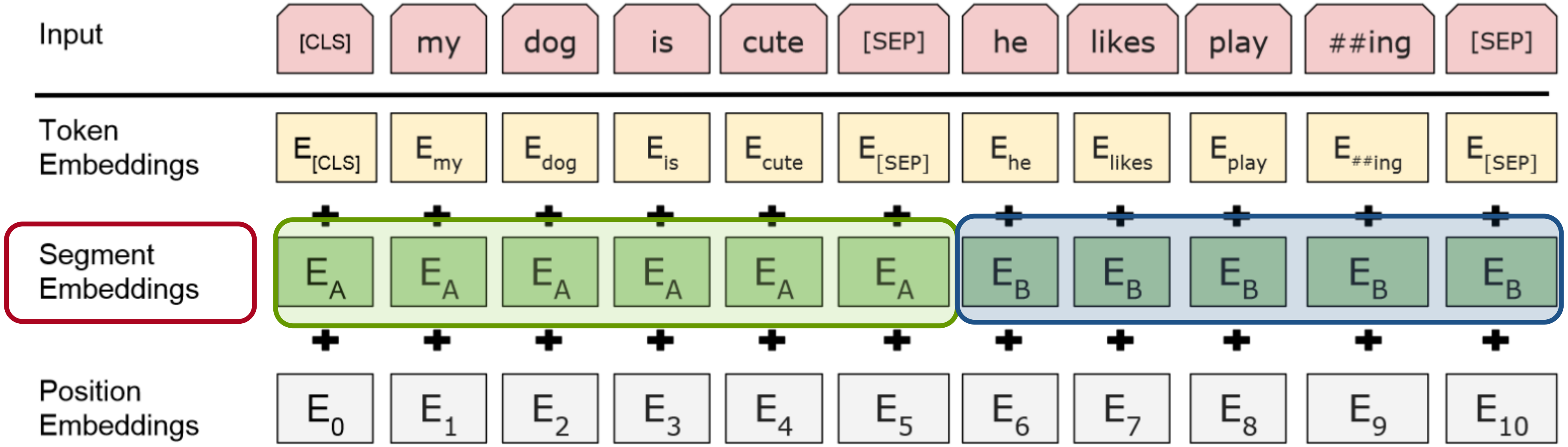
These are like standard Transformers.

Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

BERT: Input Representation (Cont.)

Sum of the token embeddings, the segmentation embeddings, and the position embeddings.

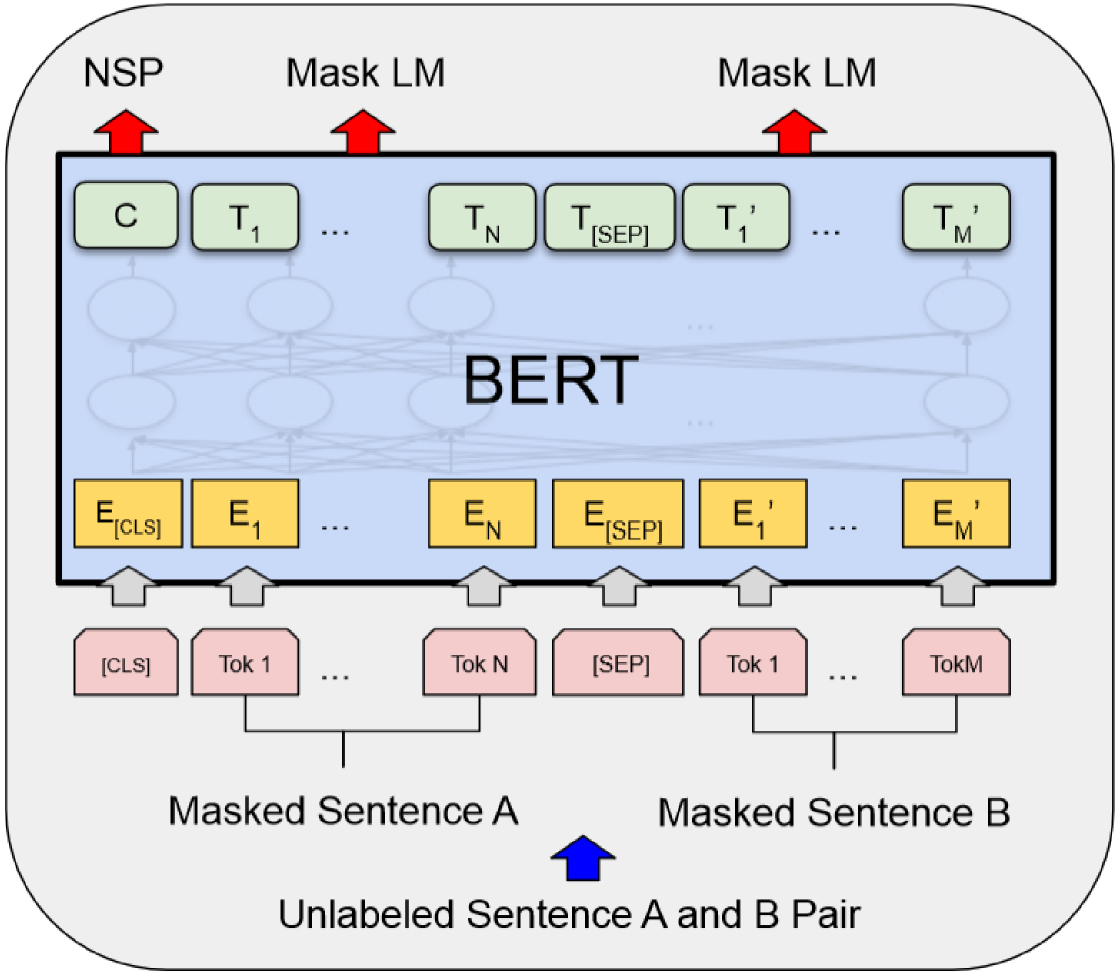


These help to differentiate between sentences.

Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

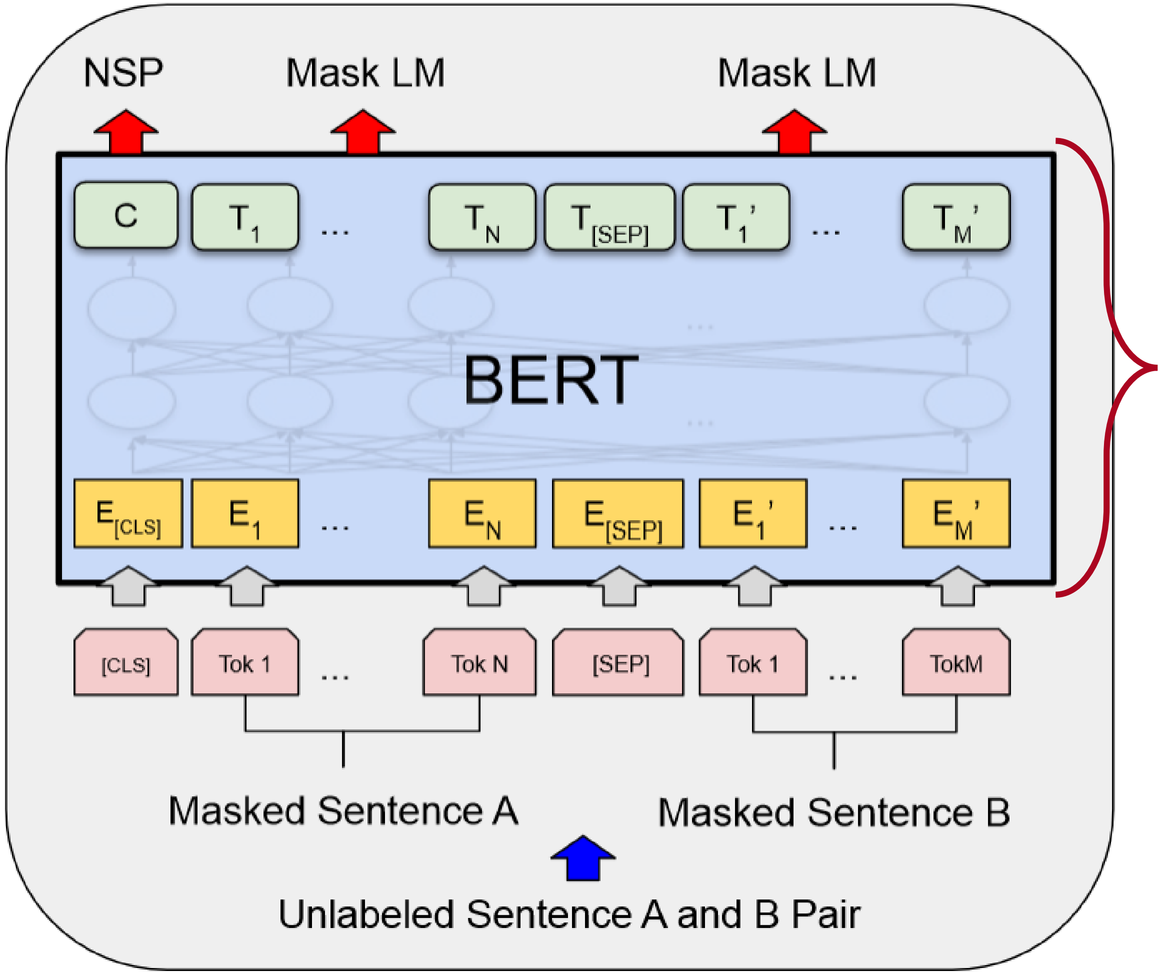
BERT: Pre-training



Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Pre-training (Cont.)

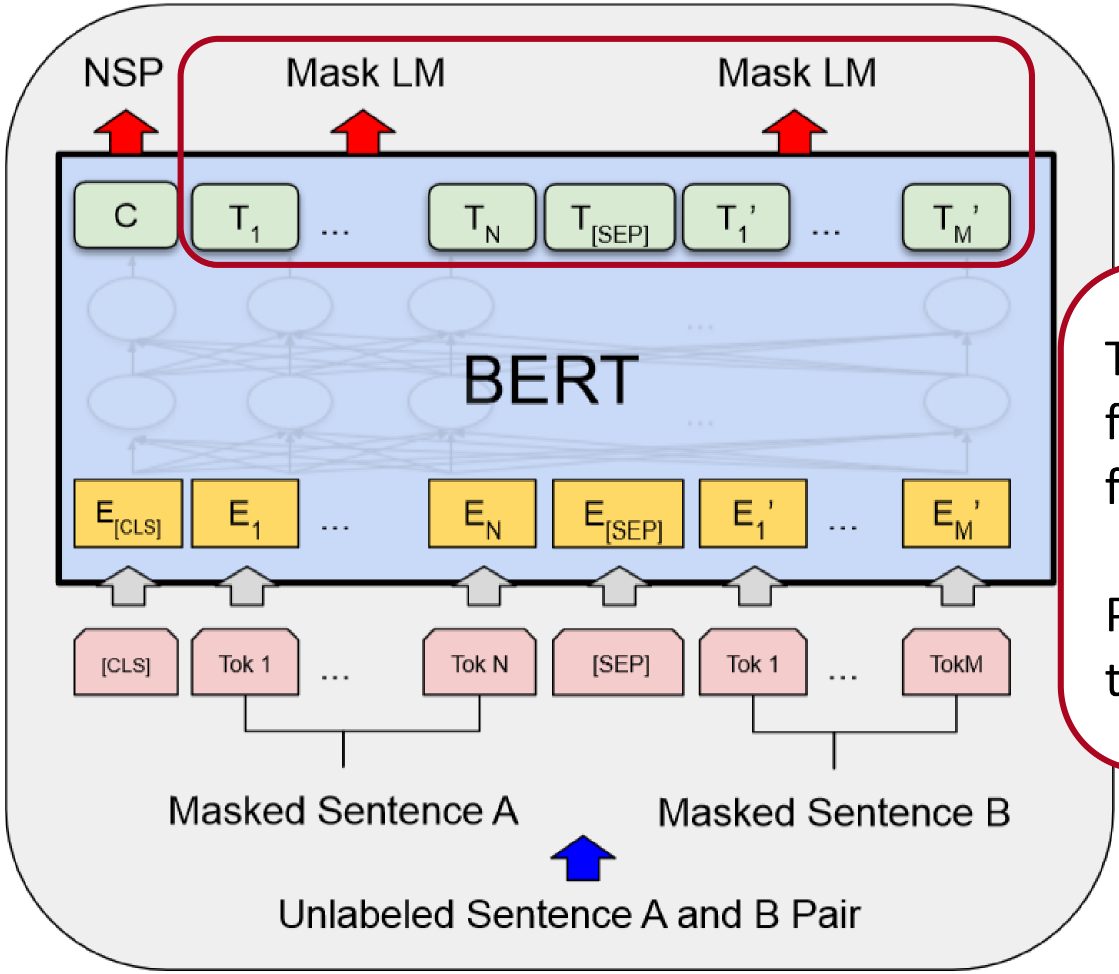


12 (bert-base) or
24 (bert-large)
stacked Encoder blocks.

Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Pre-training (Cont.)



The final hidden states are fed into an output softmax for Masked LM.

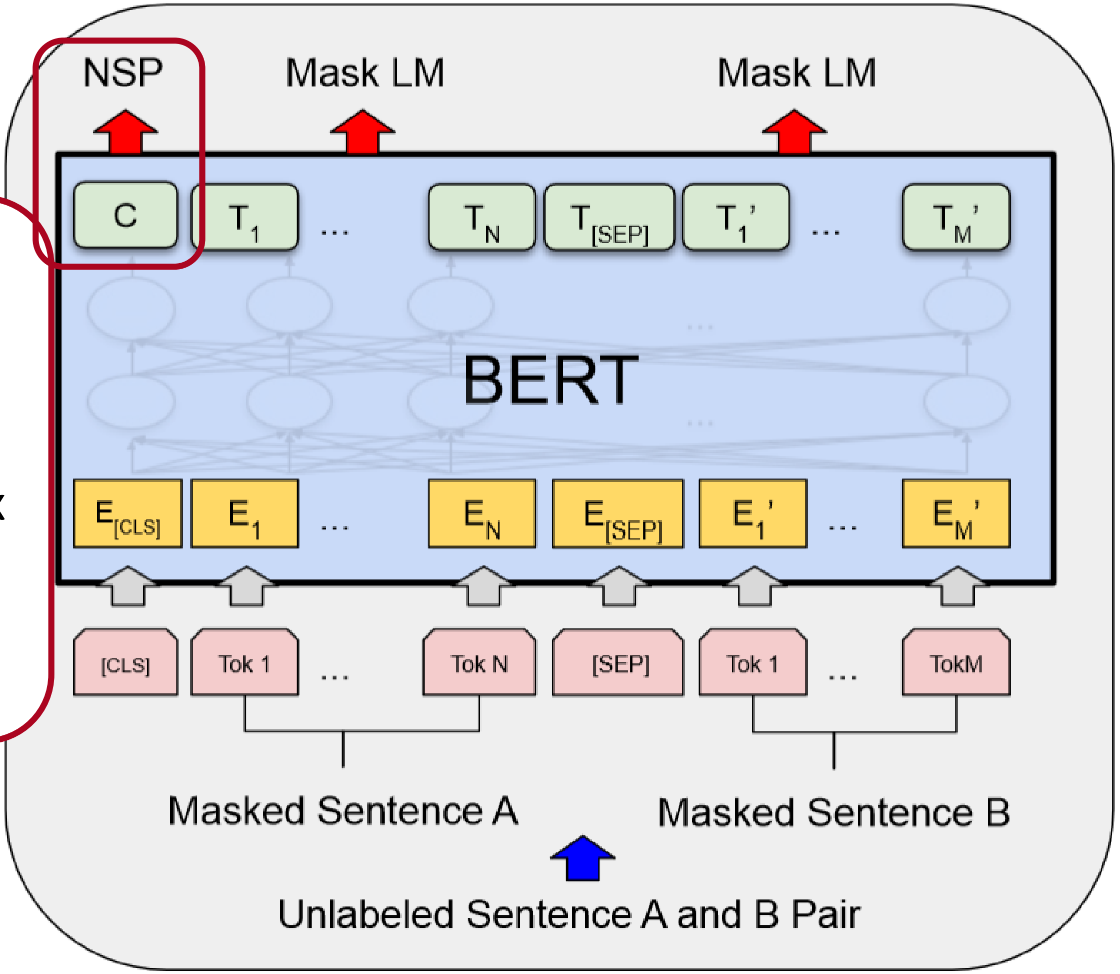
Predictions for not masked tokens are ignored.

Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

BERT: Pre-training (Cont.)

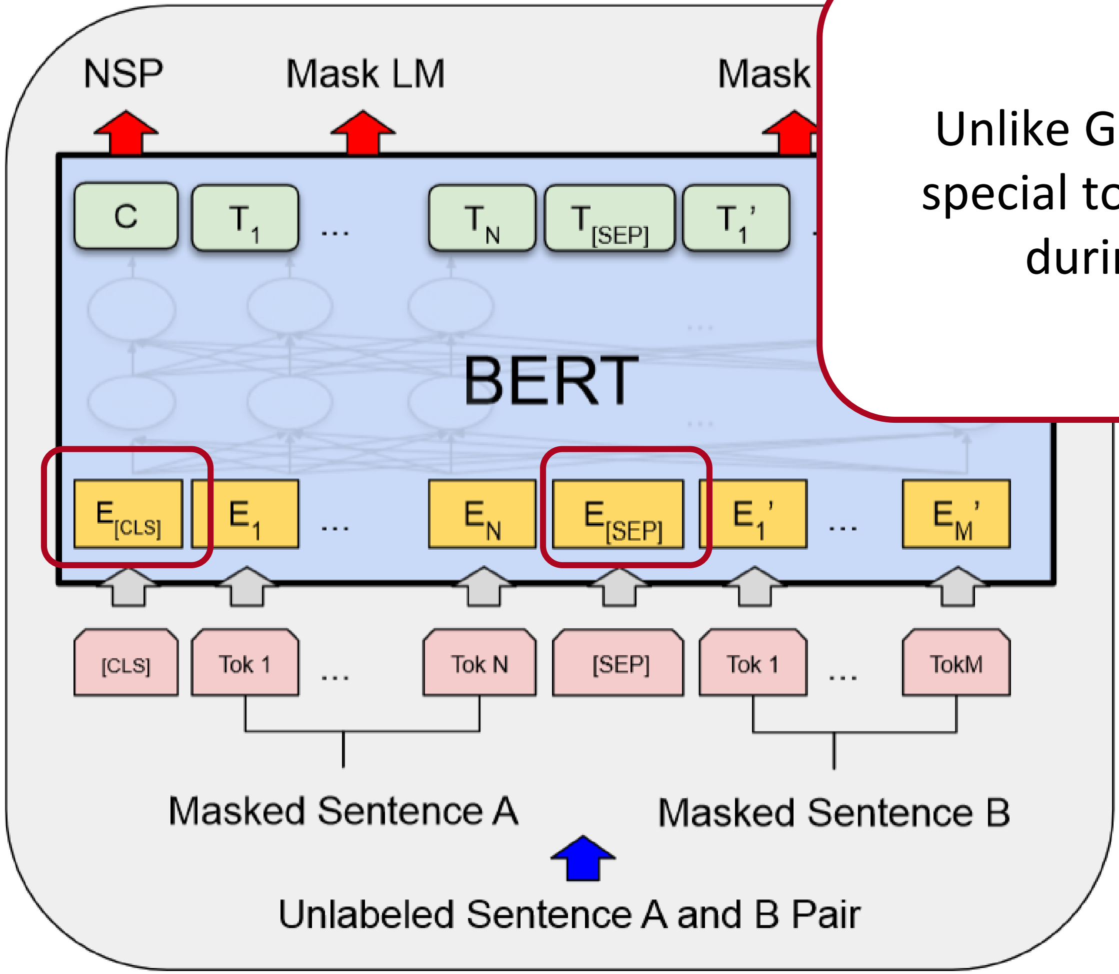
The final hidden state of the [CLS] token is fed into a softmax for Next Sentence Prediction.



Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Pre-training (Cont.)

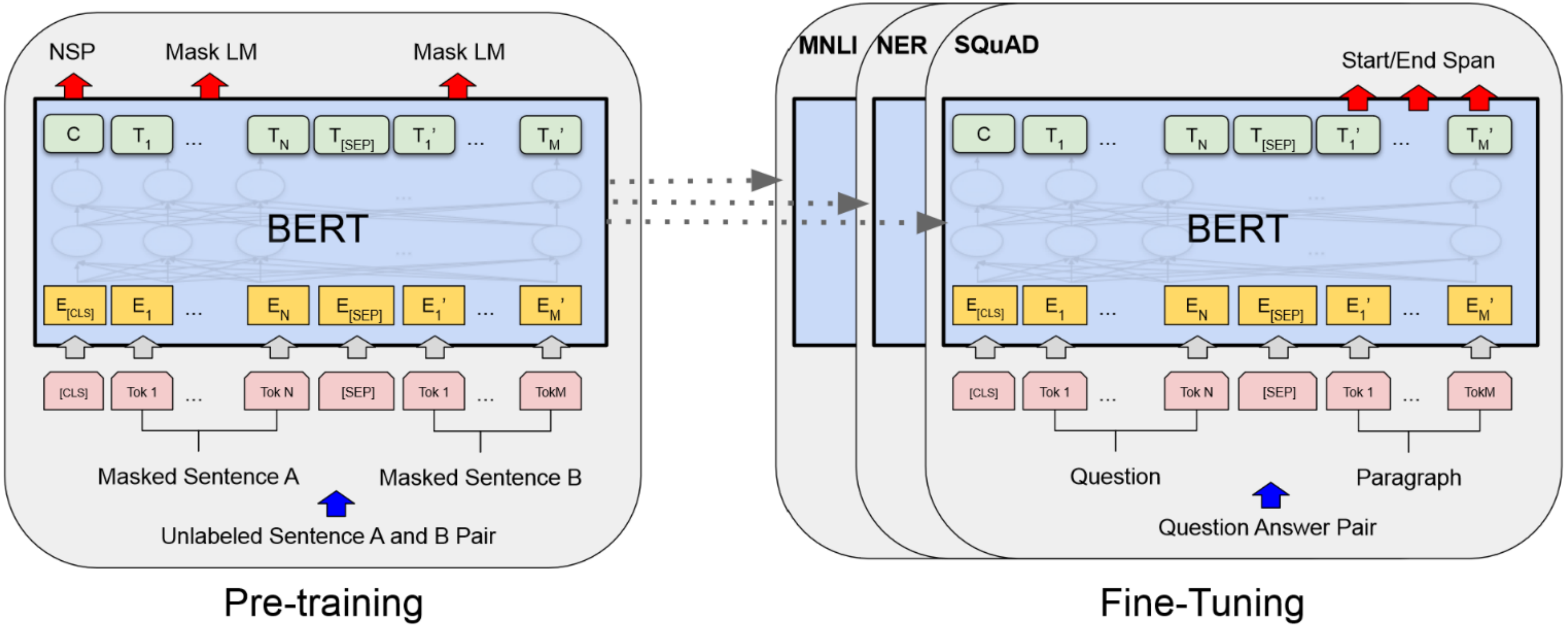


Unlike GPT, embeddings for special tokens are also learnt during Pre-training.

Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

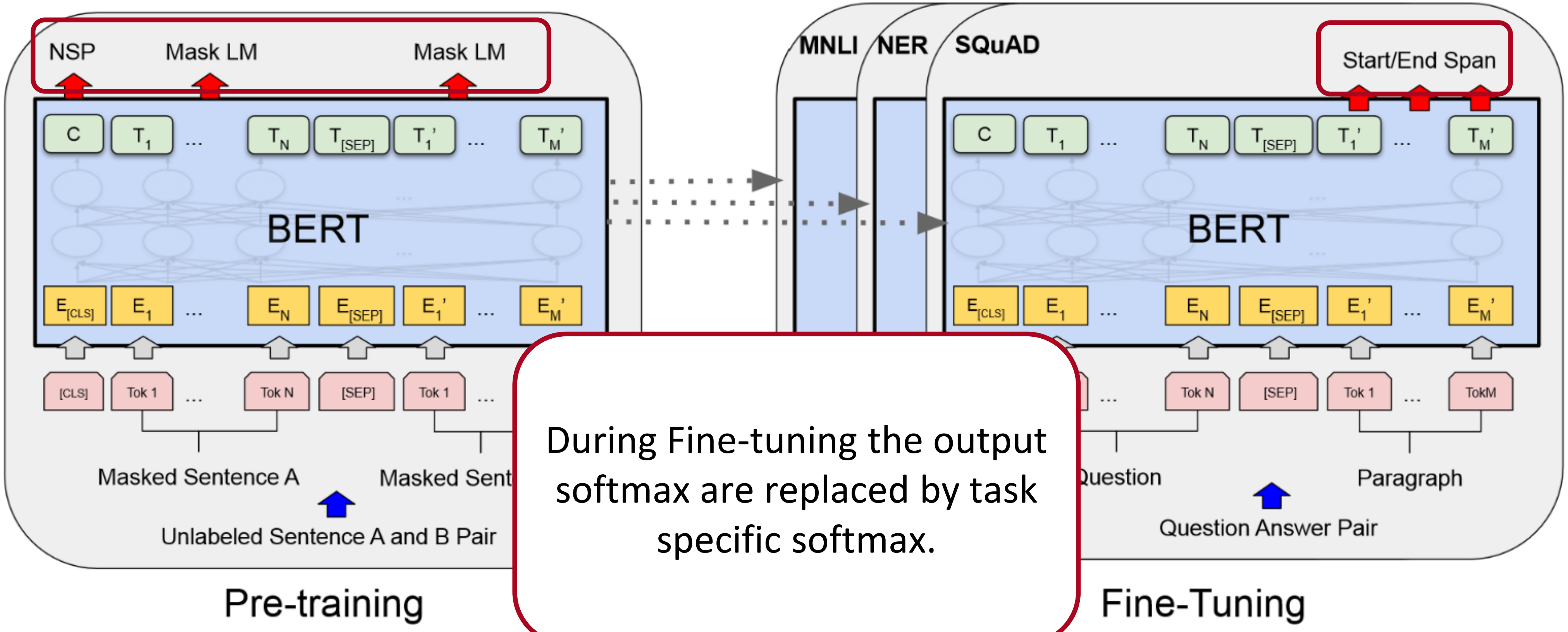
BERT: Fine-tuning (Cont.)



Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Fine-tuning Settings (Cont.)

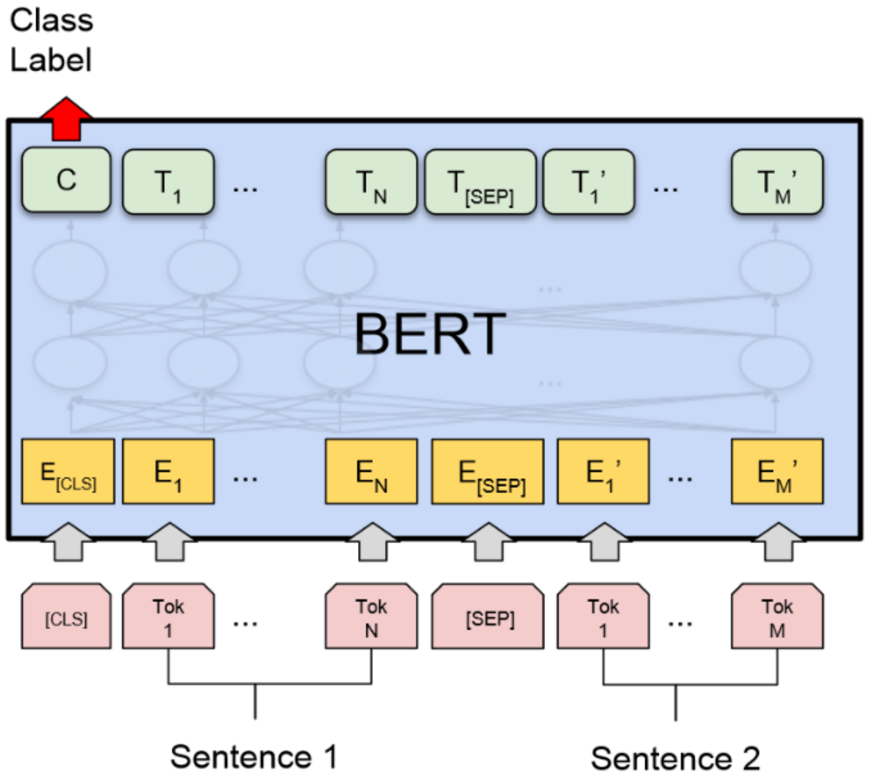


During Fine-tuning the output softmax are replaced by task specific softmax.

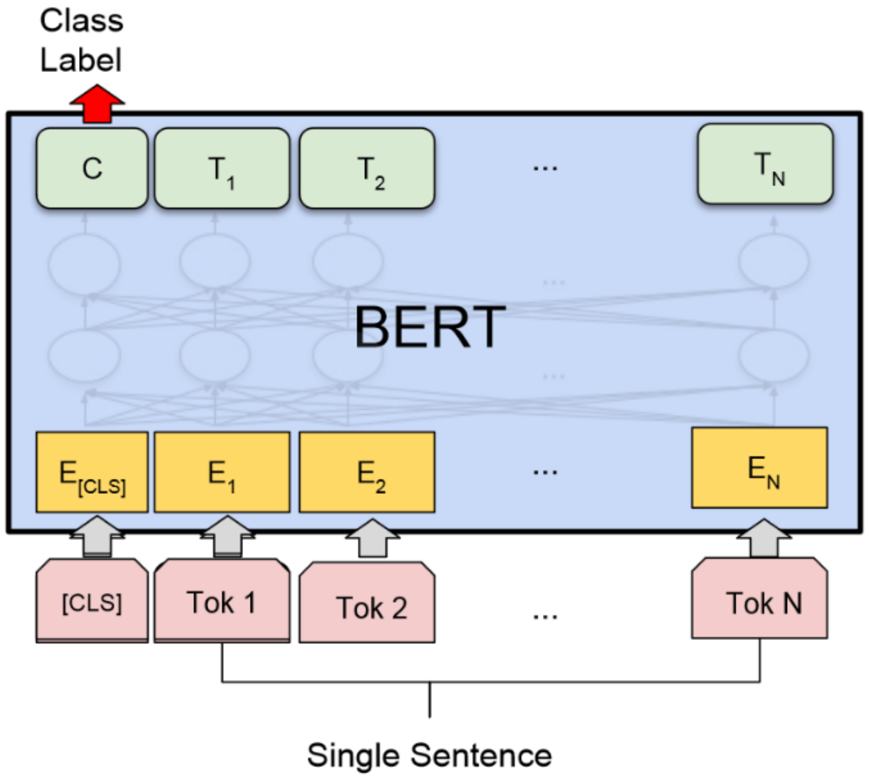
Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Fine-tuning Settings (Cont.)



Text Matching

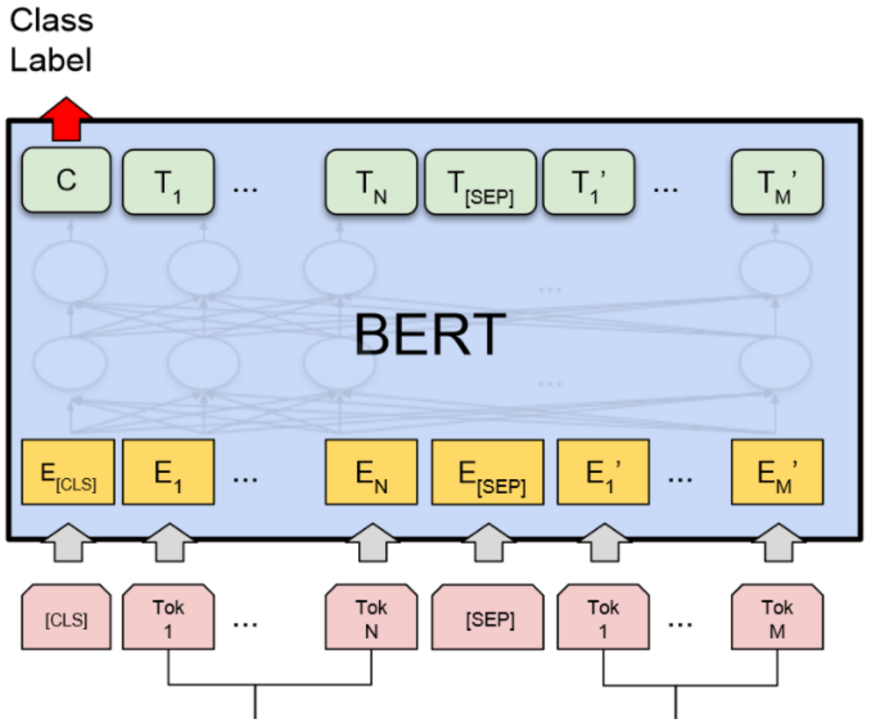


Text Classification

Devlin et al., 2018

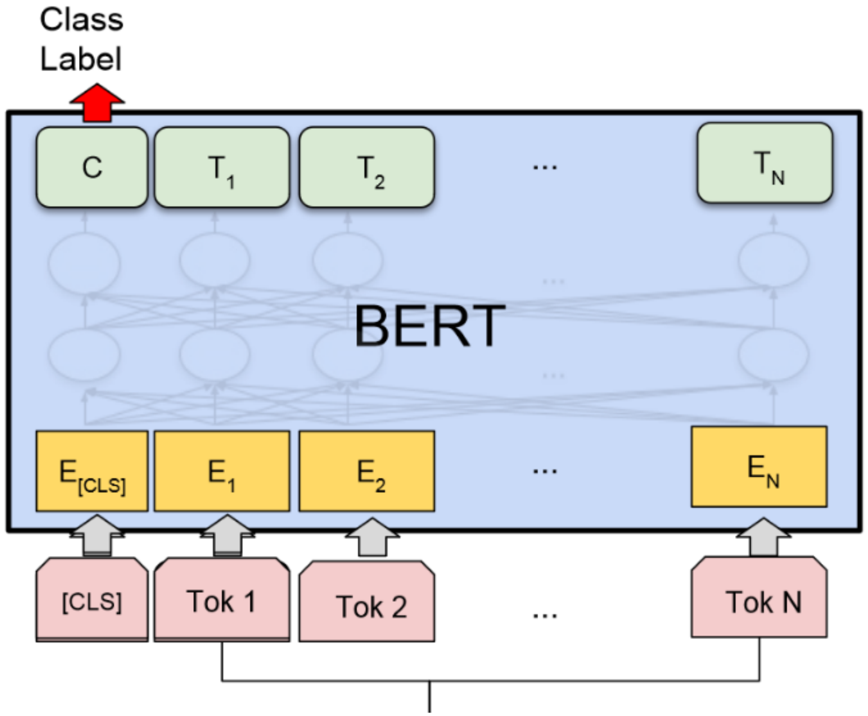
This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

BERT: Fine-tuning Settings (Cont.)



[CLS] This is a premise [SEP] This is a hypothesis

Text Matching



[CLS] This is a sentence

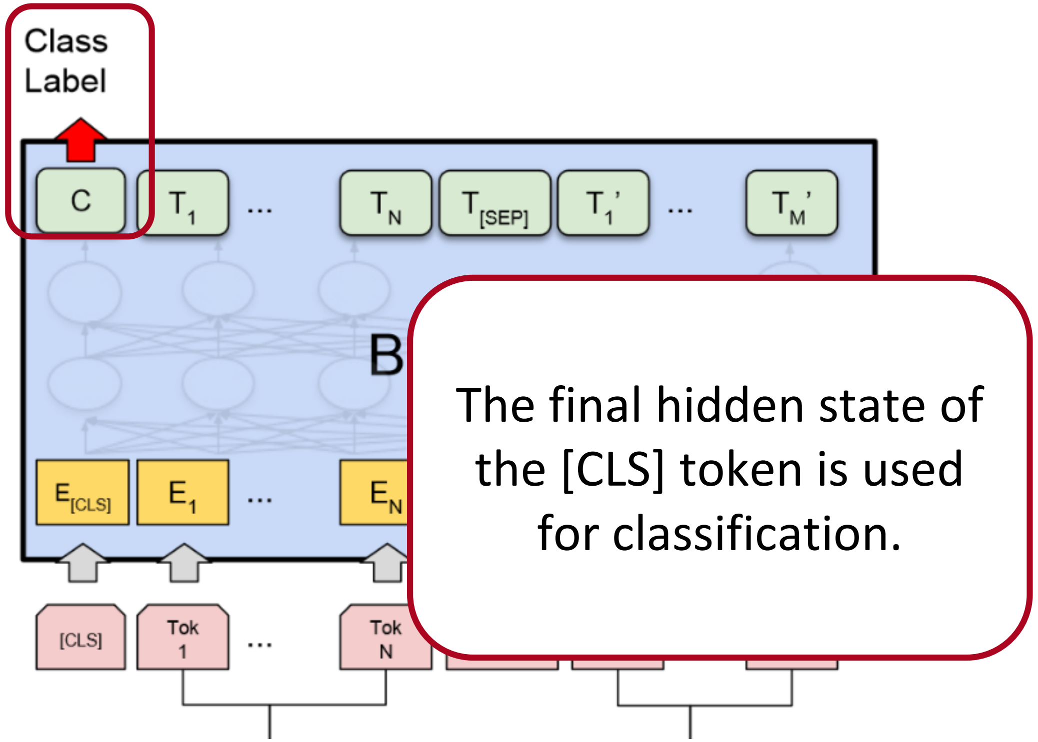
Text Classification

Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

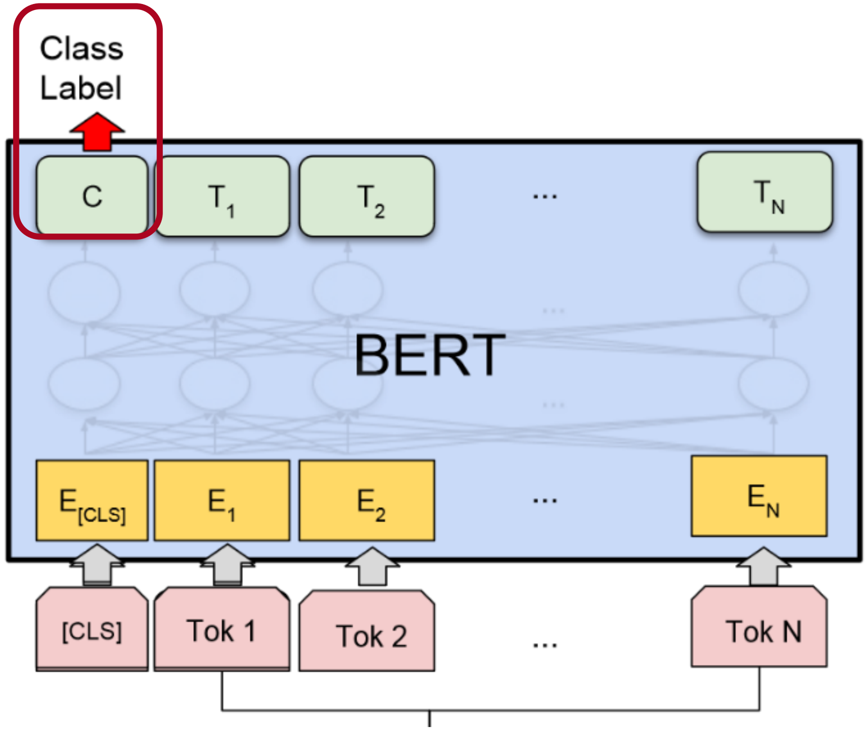
THE UNIVERSITY OF ARIZONA

BERT: Fine-tuning Settings (Cont.)



[CLS] This is a premise [SEP] This is a hypothesis

Text Matching



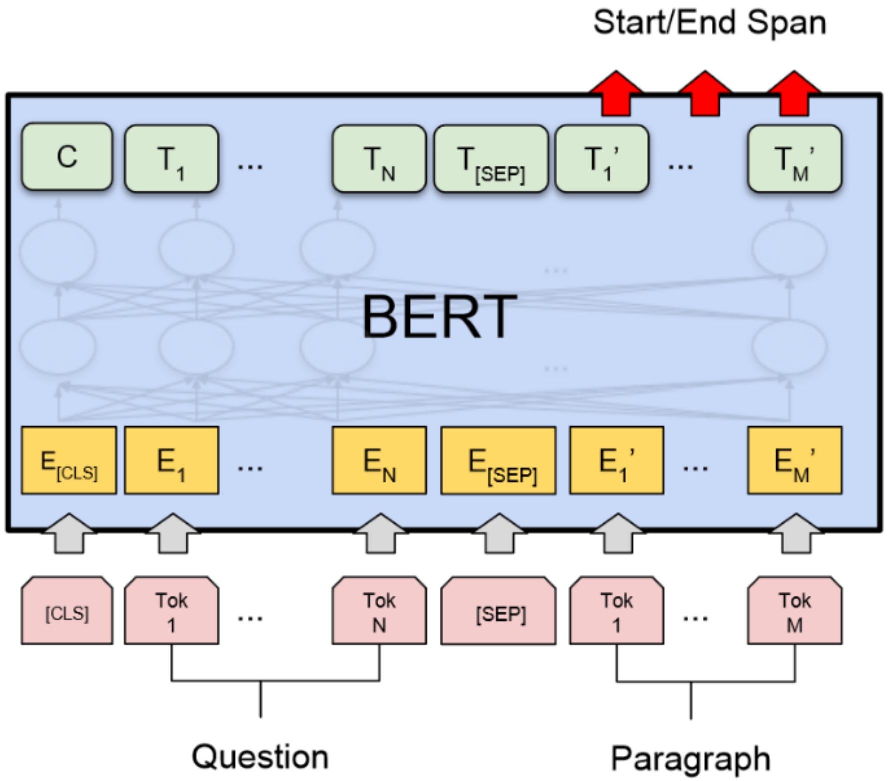
[CLS] This is a sentence

Text Classification

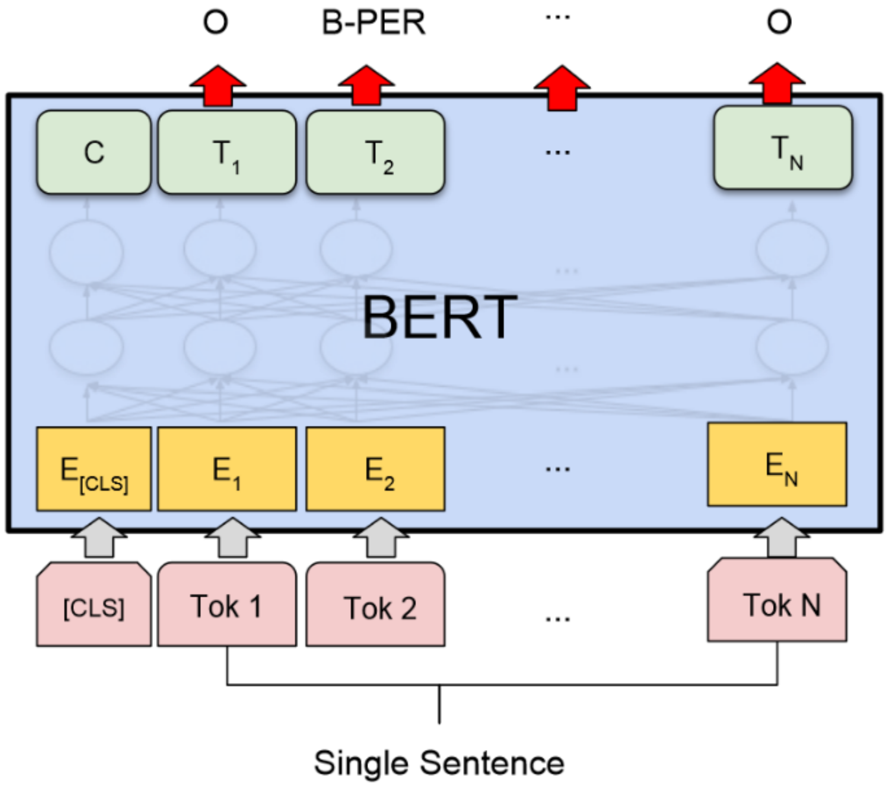
Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

BERT: Fine-tuning Settings (Cont.)



Reading Comprehension

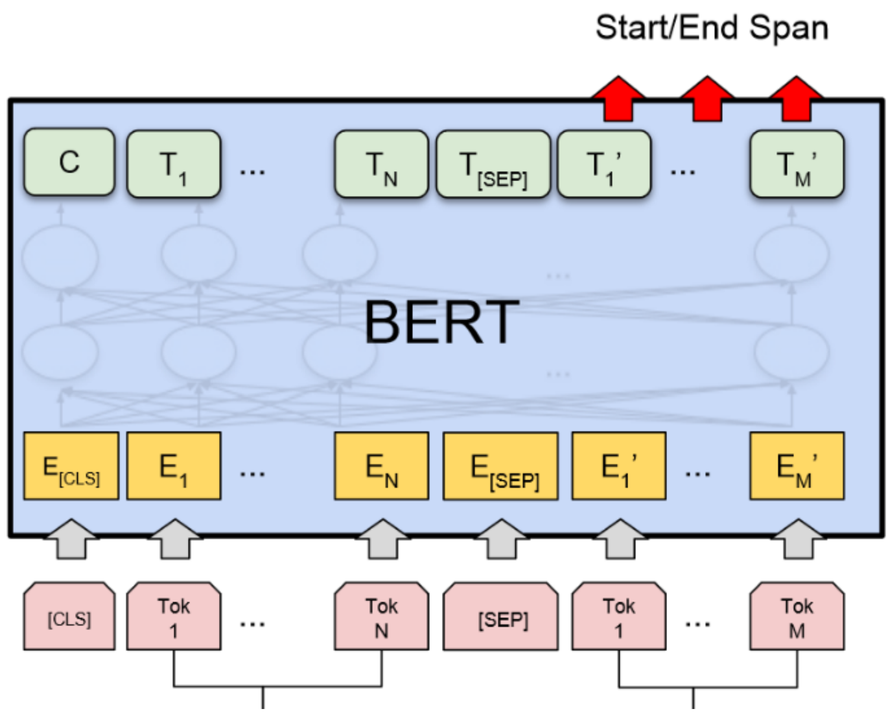


Sequence Labelling

Devlin et al., 2018

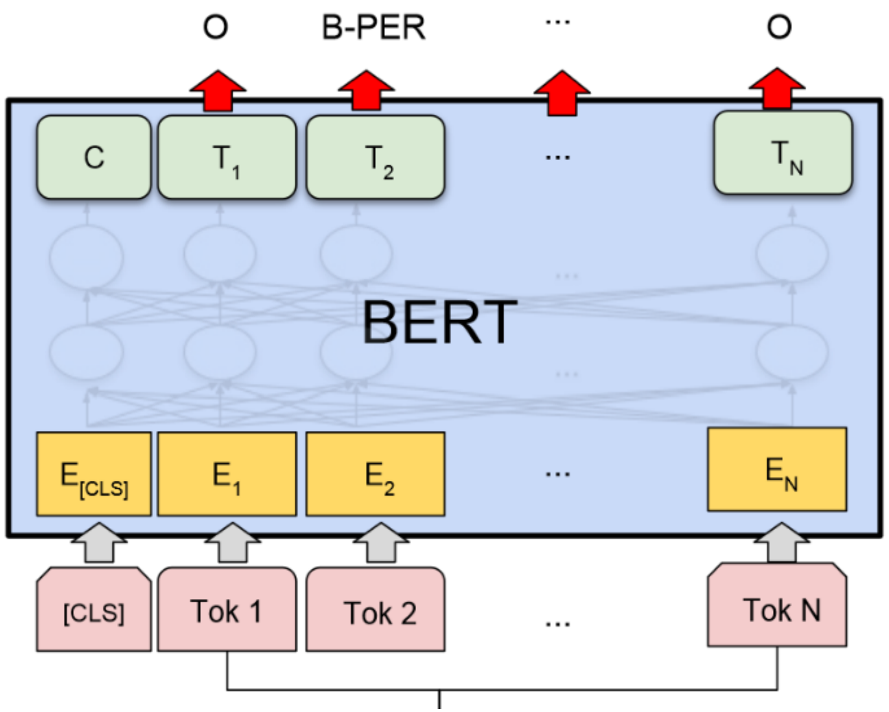
This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

BERT: Fine-tuning Settings (Cont.)



[CLS] This is a question [SEP] This is a context

Reading Comprehension



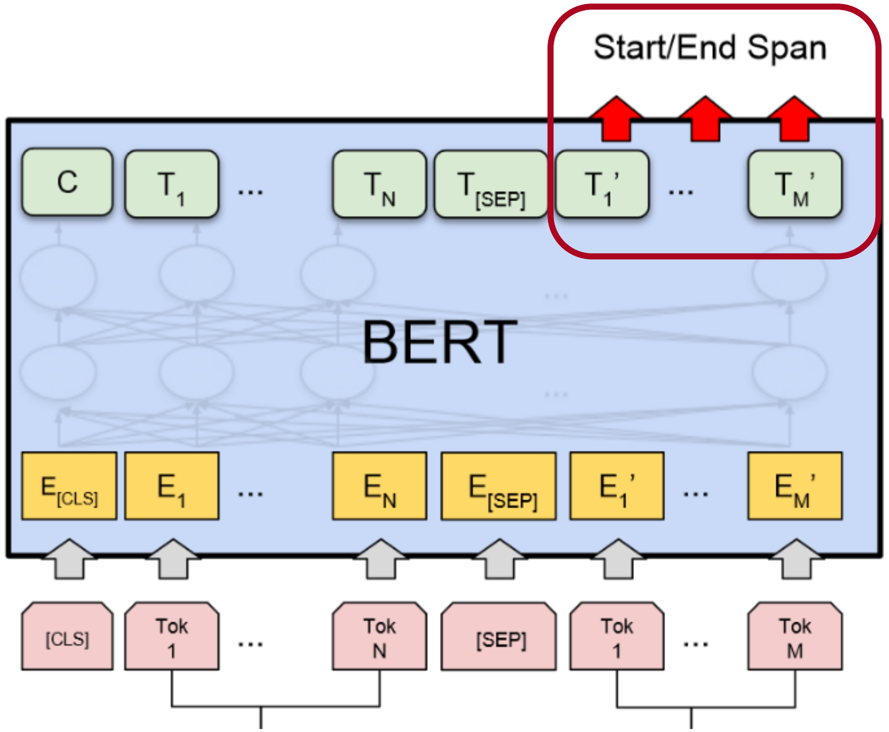
[CLS] This is a sentence

Sequence Labelling

Devlin et al., 2018

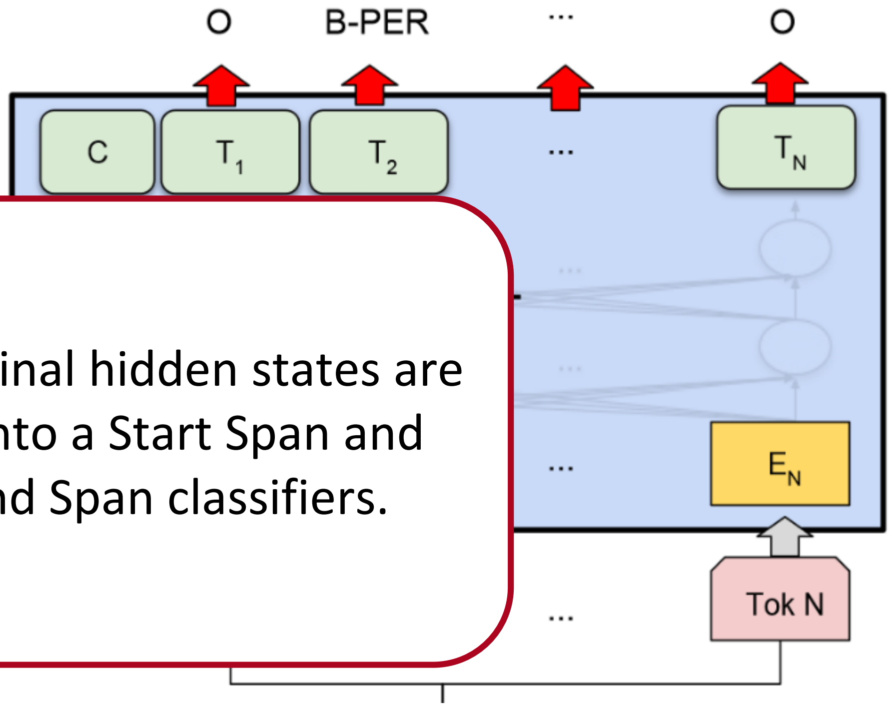
This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

BERT: Fine-tuning Settings (Cont.)



[CLS] This is a question [SEP] This is a context

Reading Comprehension



The final hidden states are fed into a Start Span and an End Span classifiers.

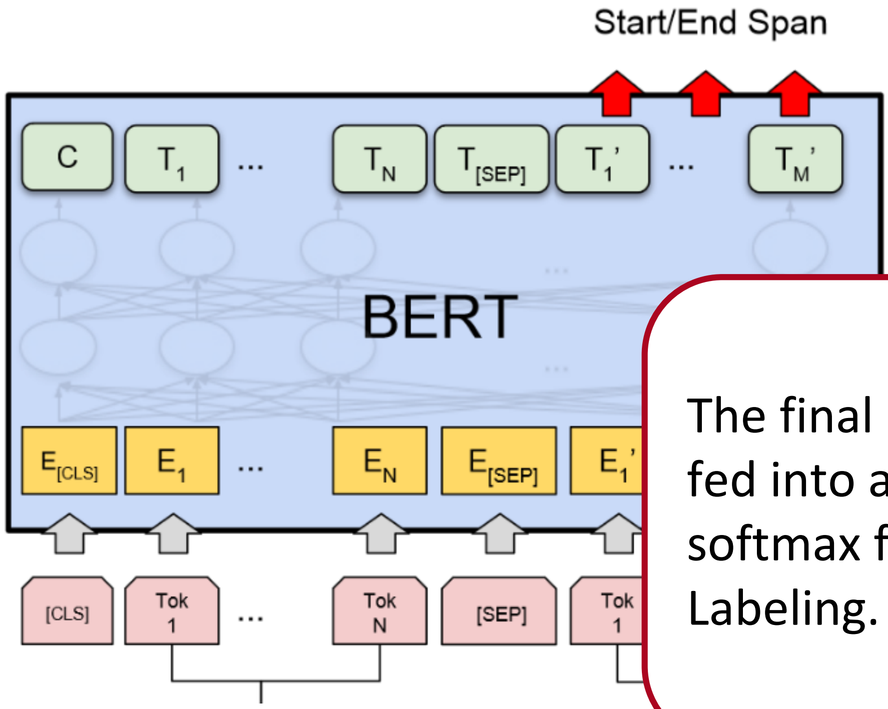
[CLS] This is a sentence

Sequence Labelling

Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

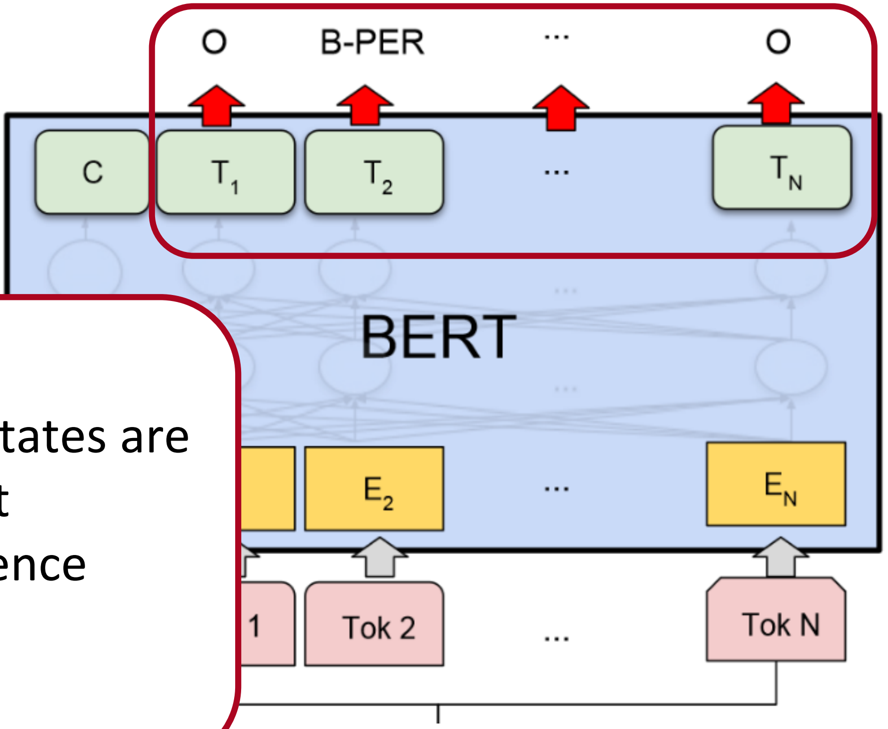
BERT: Fine-tuning Settings (Cont.)



[CLS] This is a question [SEP] This is a context

Reading Comprehension

The final hidden states are fed into an output softmax for Sequence Labeling.



[CLS] This is a sentence

Sequence Labelling

Devlin et al., 2018

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Pre-trained Transformers

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Agenda

In this session, we will discuss:

- Size of Pre-trained Transformers
- Seq-to-Seq Pre-trained Transformer
- Alternative Pre-Training Tasks

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Size of Pre-trained Transformers

- Pre-trained Transformers provide state-of-the-art results for several tasks.
  - The larger the model, the better performance.
- Results in GLUE (General Language Understanding Evaluation) benchmark:

	Average	Acceptability	Sentiment Analysis	Similarity	Paraphrase	Question Paraphrase	Question NLI	NLI
BERT (large)	**82.1**	**60.5**	**94.9**	**86.5**	**89.3**	**72.1**	**92.7**	**70.1**
BERT (base)	79.6	52.1	93.5	85.8	88.9	71.2	90.5	66.4
GPT	75.1	45.4	91.3	80.0	82.3	70.3	87.4	56.0
Previous SOTA	74.0	35.0	93.2	81.0	86.0	66.1	82.3	61.7

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Size of Pre-trained Transformers (Cont.)

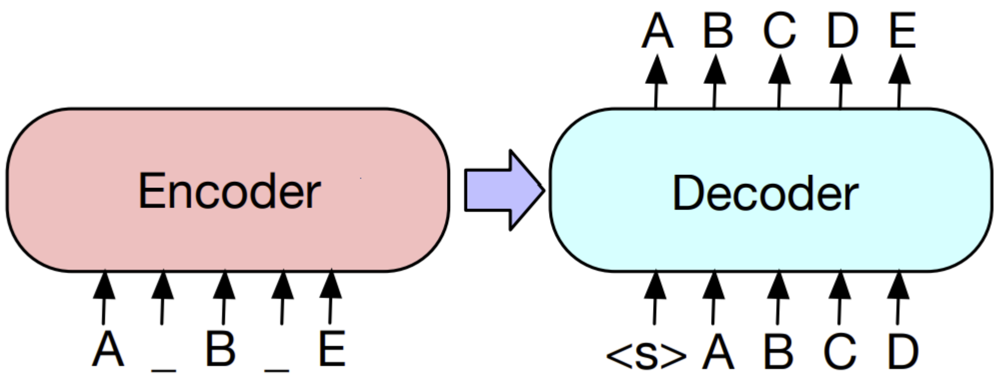
- Large pre-trained Transformers are extremely large.

Model	Params	Corpus	Corpus Size	
BERT	110M-340M	WikiEn+BookCorpus	16GB	~3.3 Billion tokens
GPT	117M	BookCorpus	4.6GB	~1.3 Billion tokens
GPT2	117M-1.5B	WebText	40GB	~15 Billion tokens
GPT3	125M-175B	WikiEn+BookCorpus+ WebText+CommonCrawl	570GB	~400 Billion tokens

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Seq-to-Seq Pre-trained Transformer

- GPT only uses the Decoder part of the Transformer.
  - It is not bi-directional.
- BERT only uses the Encoder part of the Transformer.
  - Cannot be easily used for text generation.
- We can pre-train a whole Transformer (e.g., BART):



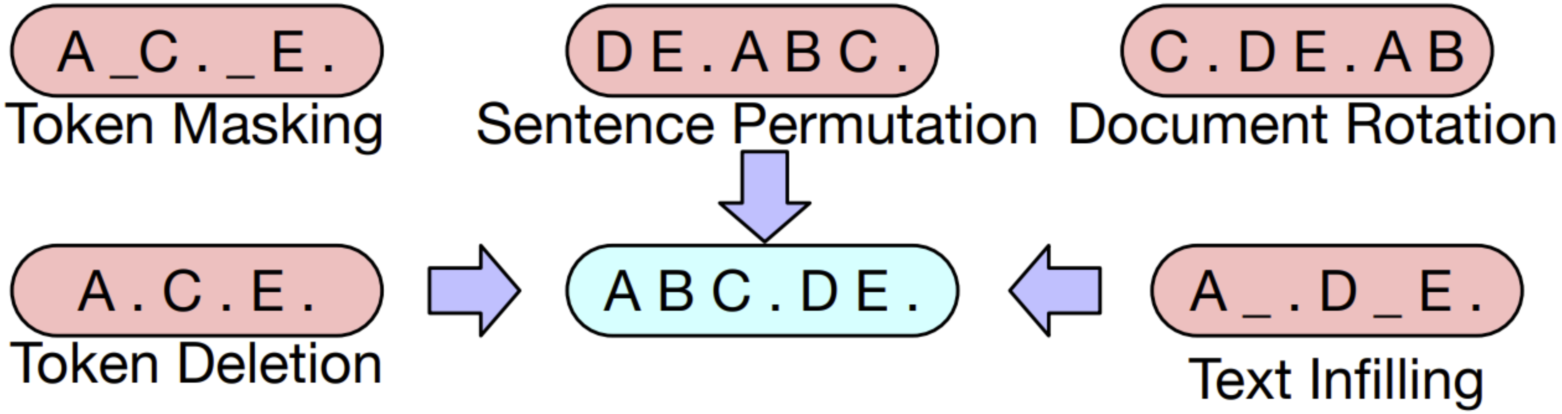
Lewis et al., 2019

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

THE UNIVERSITY OF ARIZONA

Seq-to-Seq Pre-trained Transformer: Fine-tuning

- The Pre-training task consists of reconstructing a corrupted input:

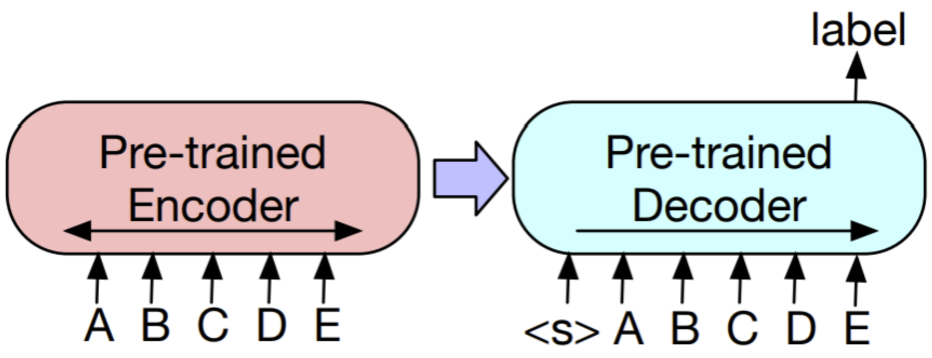


Lewis et al., 2019

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

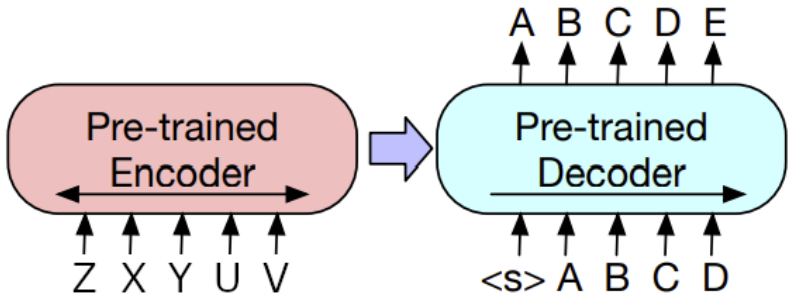
Seq-to-Seq Pre-trained Transformer: Fine-tuning (Cont.)

Classification tasks



- Same input is fed into the Encoder and Decoder.
- The final output is fed into the Classifier.

Seq2Seq tasks



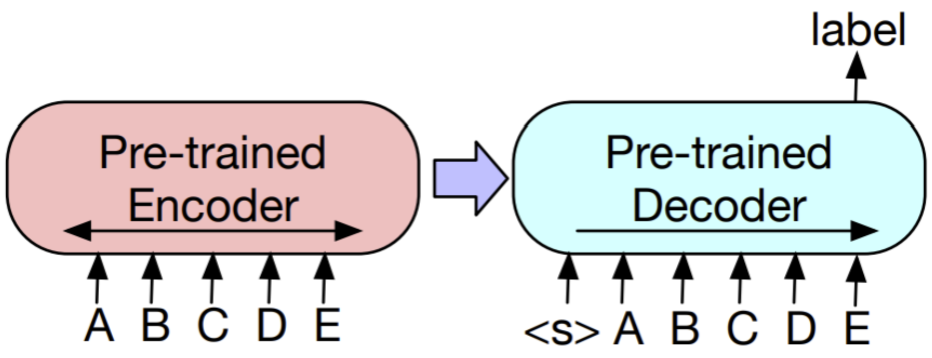
- An uncorrupted input is fed into both the Encoder and Decoder.

Lewis et al., 2019

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

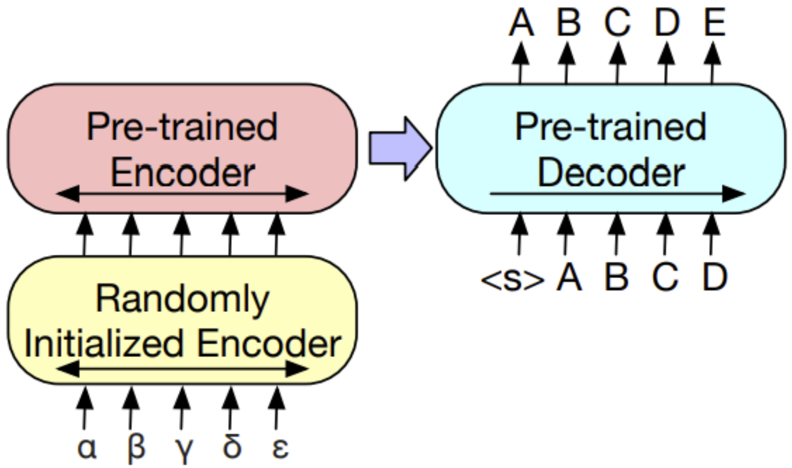
Seq-to-Seq Pre-trained Transformer: Fine-tuning (Cont.)

Classification tasks



- Same input is fed into the Encoder and Decoder.
- The final output is fed into the Classifier.

Machine Translation



- Additional Encoder that replaces the word embeddings

Lewis et al., 2019

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Alternative Pre-training Tasks

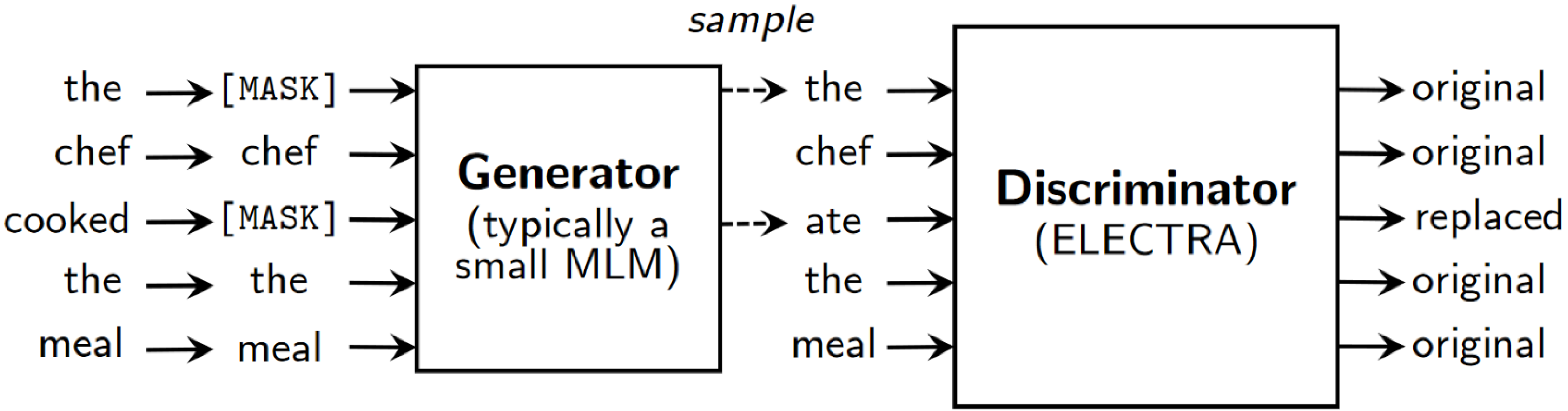
- Autoregressive Language Models are unidirectional.
- Masked Language Modeling is bidirectional but creates a mismatch between Pre-training and Fine-tuning.
  - The [MASK] token is not seen during Fine-tuning.

- Alternatives:
  - Replaced Token Detection (ELECTRA)
  - Permutation Language Modeling (XLNET)

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Alternative Pre-training Tasks (Cont.)

Replaced Token Detection

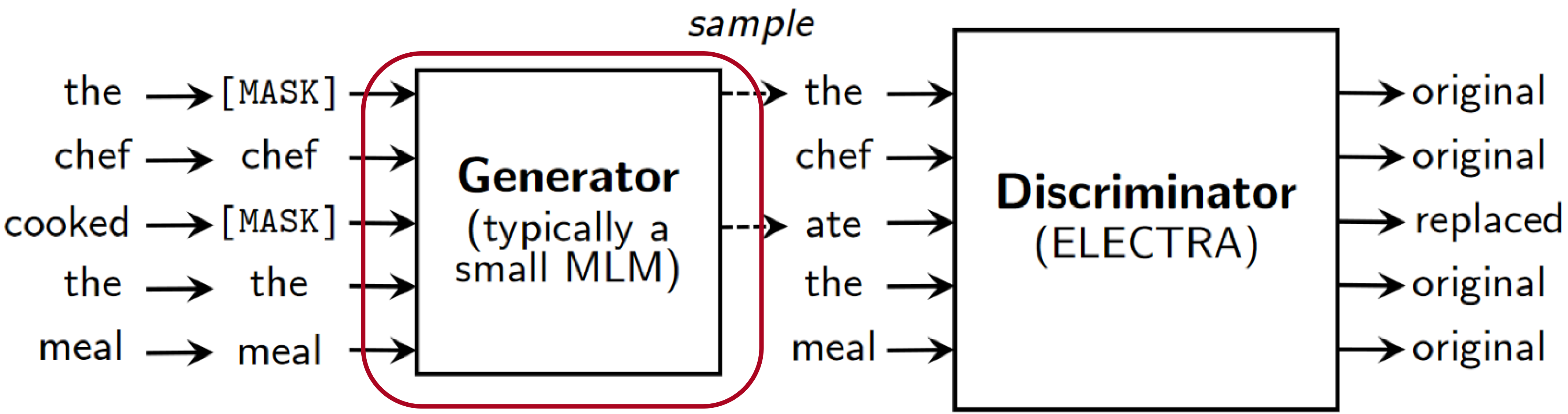


Clark et al., 2020

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Alternative Pre-training Tasks (Cont.)

Replaced Token Detection



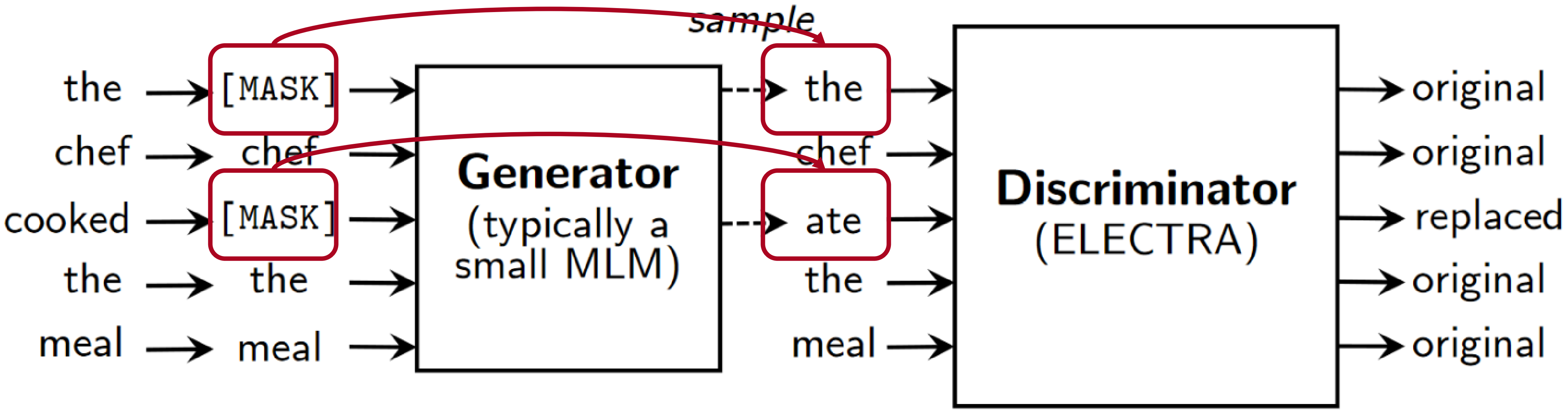
Train a Small Transformer for Masked Language Modeling.

Clark et al., 2020

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Alternative Pre-training Tasks (Cont.)

Replaced Token Detection



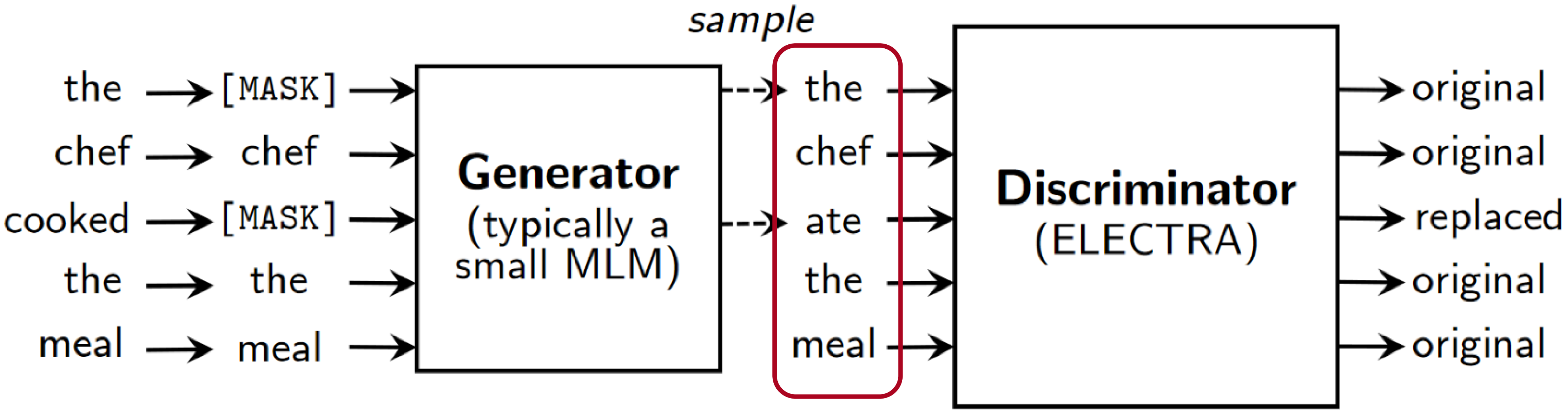
Use it to replace [MASK] with predicted tokens.

Clark et al., 2020

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Alternative Pre-training Tasks (Cont.)

Replaced Token Detection



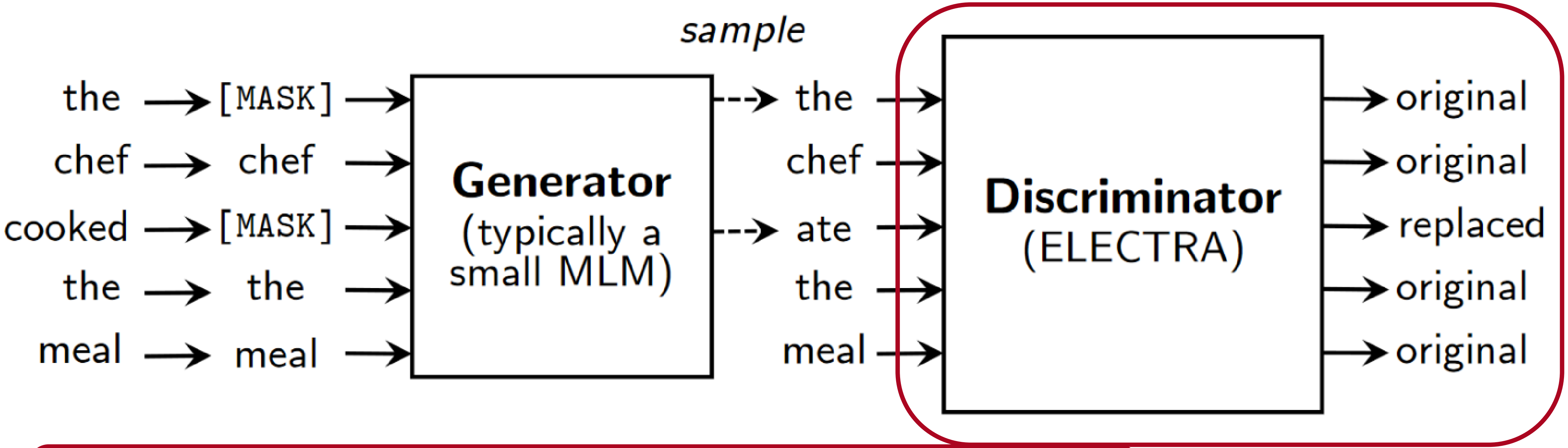
Unlike random replacement, the corrupted sequence is plausible.

Clark et al., 2020

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Alternative Pre-training Tasks (Cont.)

Replaced Token Detection



Train a Large Transformer Encoder to predict the replaced tokens.

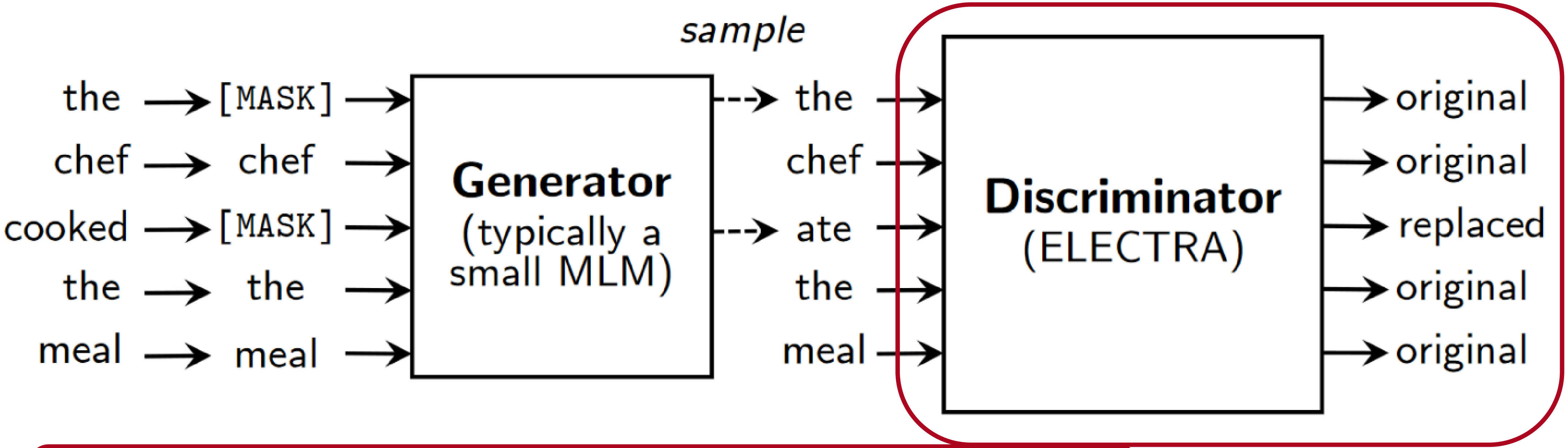
Clark et al., 2020

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Alternative Pre-training Tasks (Cont.)

Replaced Token Detection

The model does not see [MASK] token during Pre-training.



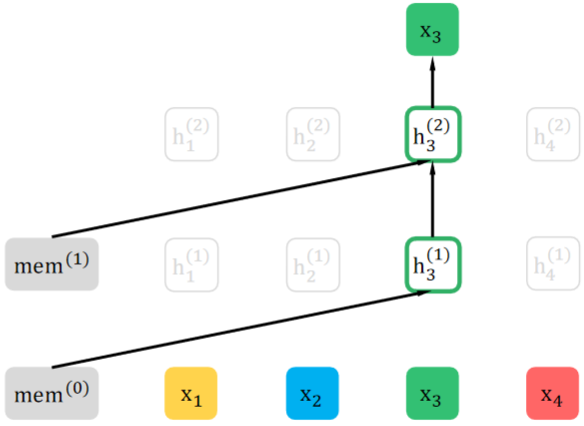
Train a Large Transformer Encoder to predict the replaced tokens.

Clark et al., 2020

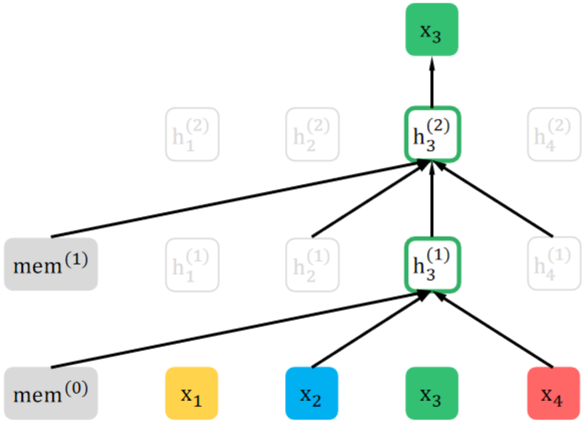
This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Alternative Pre-training Tasks (Cont.)

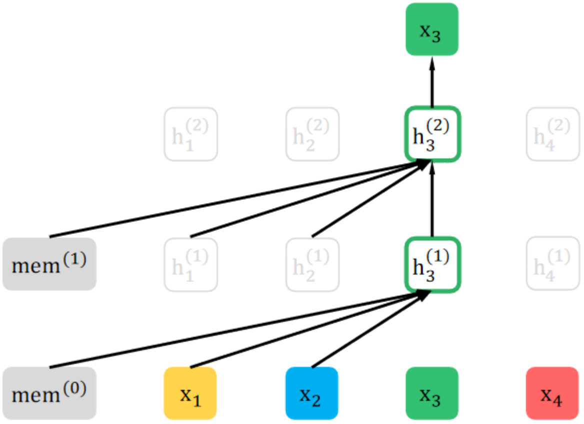
Permutation Language Modeling



Factorization order: 3 → 2 → 4 → 1

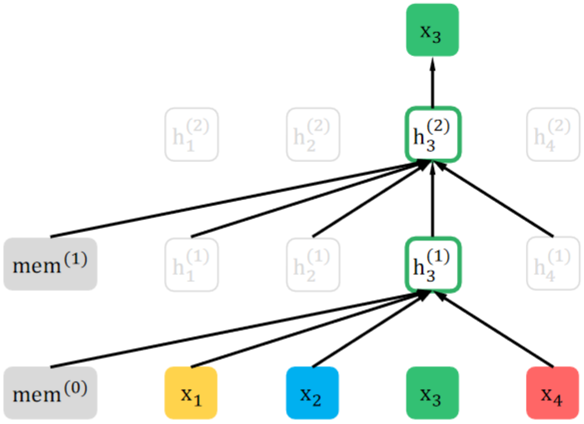


Factorization order: 2 → 4 → 3 → 1

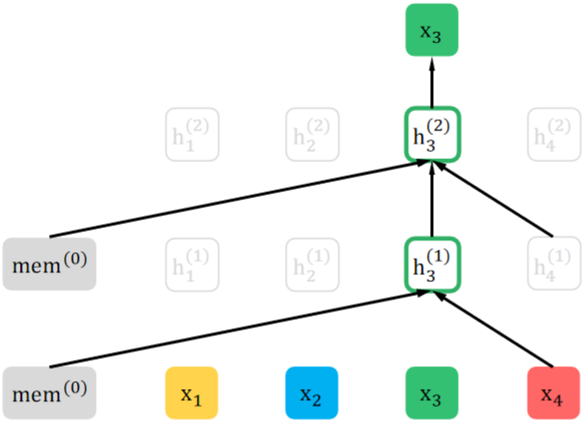


Factorization order: 1 → 2 → 3 → 4

Original Order



Factorization order: 1 → 4 → 2 → 3



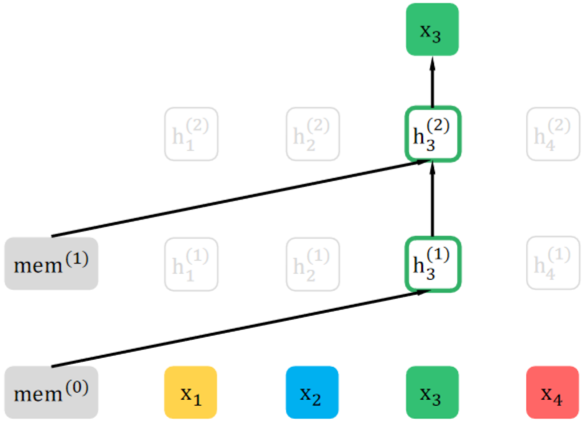
Factorization order: 4 → 3 → 1 → 2

Yang et al., 2019

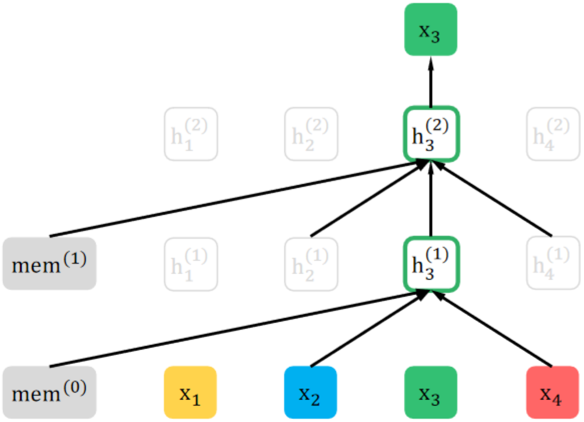
This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Alternative Pre-training Tasks (Cont.)

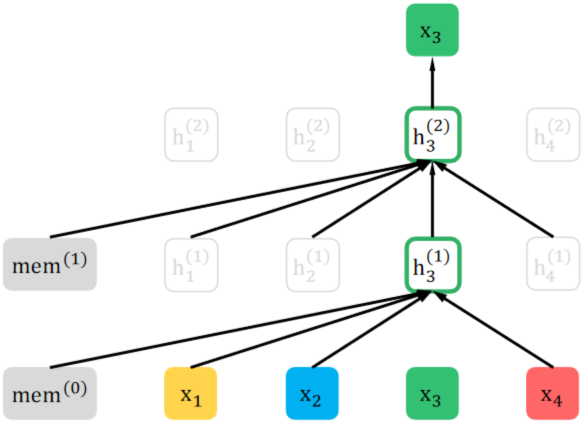
Permutation Language Modeling



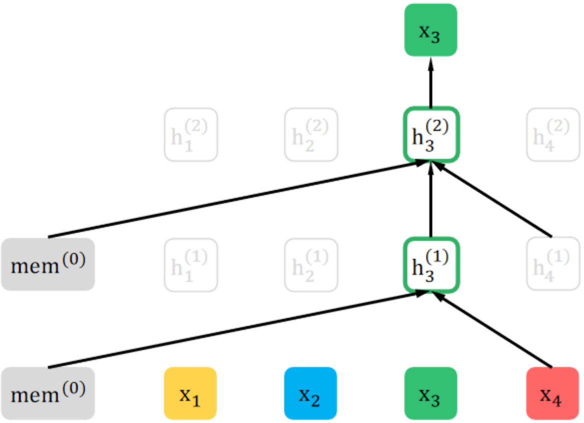
Factorization order: 3 → 2 → 4 → 1



Factorization order: 2 → 4 → 3 → 1

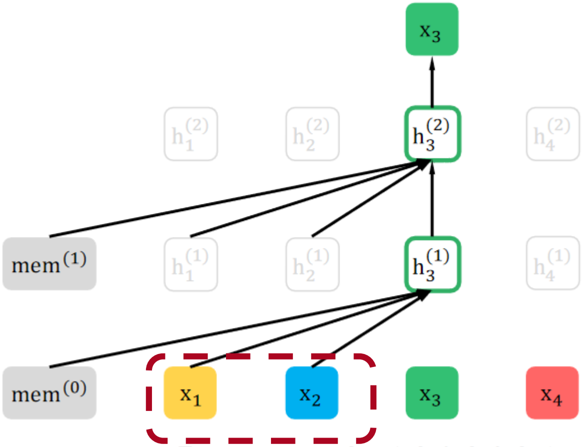


Factorization order: 1 → 4 → 2 → 3



Factorization order: 4 → 3 → 1 → 2

Autoregressive Language Model attends only to previous tokens.



Factorization order: 1 → 2 → 3 → 4

Original Order

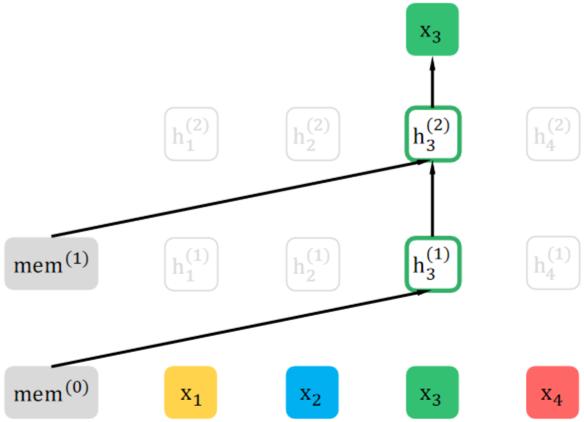
Yang et al., 2019

This file is meant for personal use by libingeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

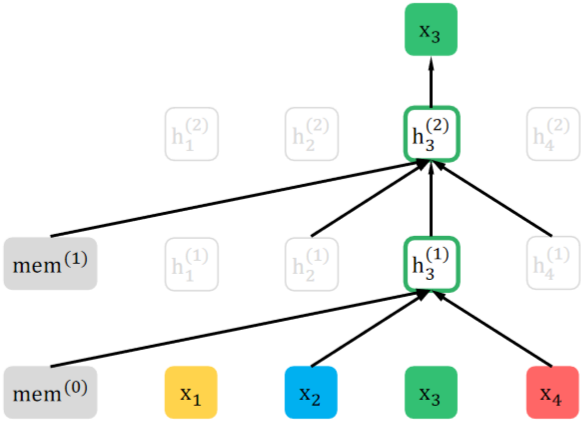
Alternative Pre-training Tasks (Cont.)

Permutation Language Modeling

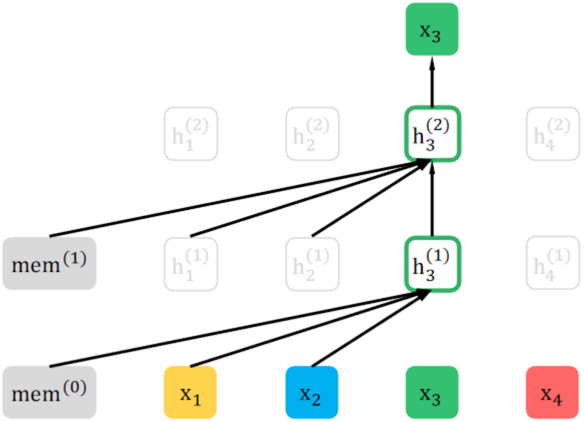
In Permutation Language Modeling, a token is predicted according to all possible permutations of the order.



Factorization order: 3 → 2 → 4 → 1

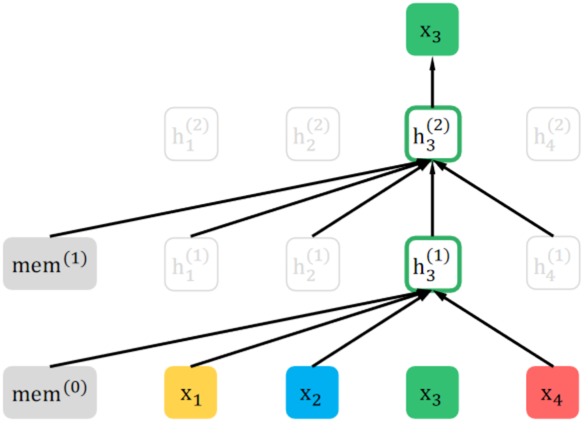


Factorization order: 2 → 4 → 3 → 1

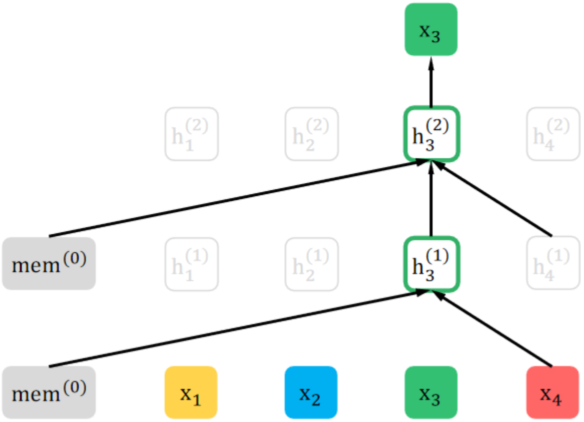


Factorization order: 1 → 2 → 3 → 4

Original Order



Factorization order: 1 → 4 → 2 → 3



Factorization order: 4 → 3 → 1 → 2

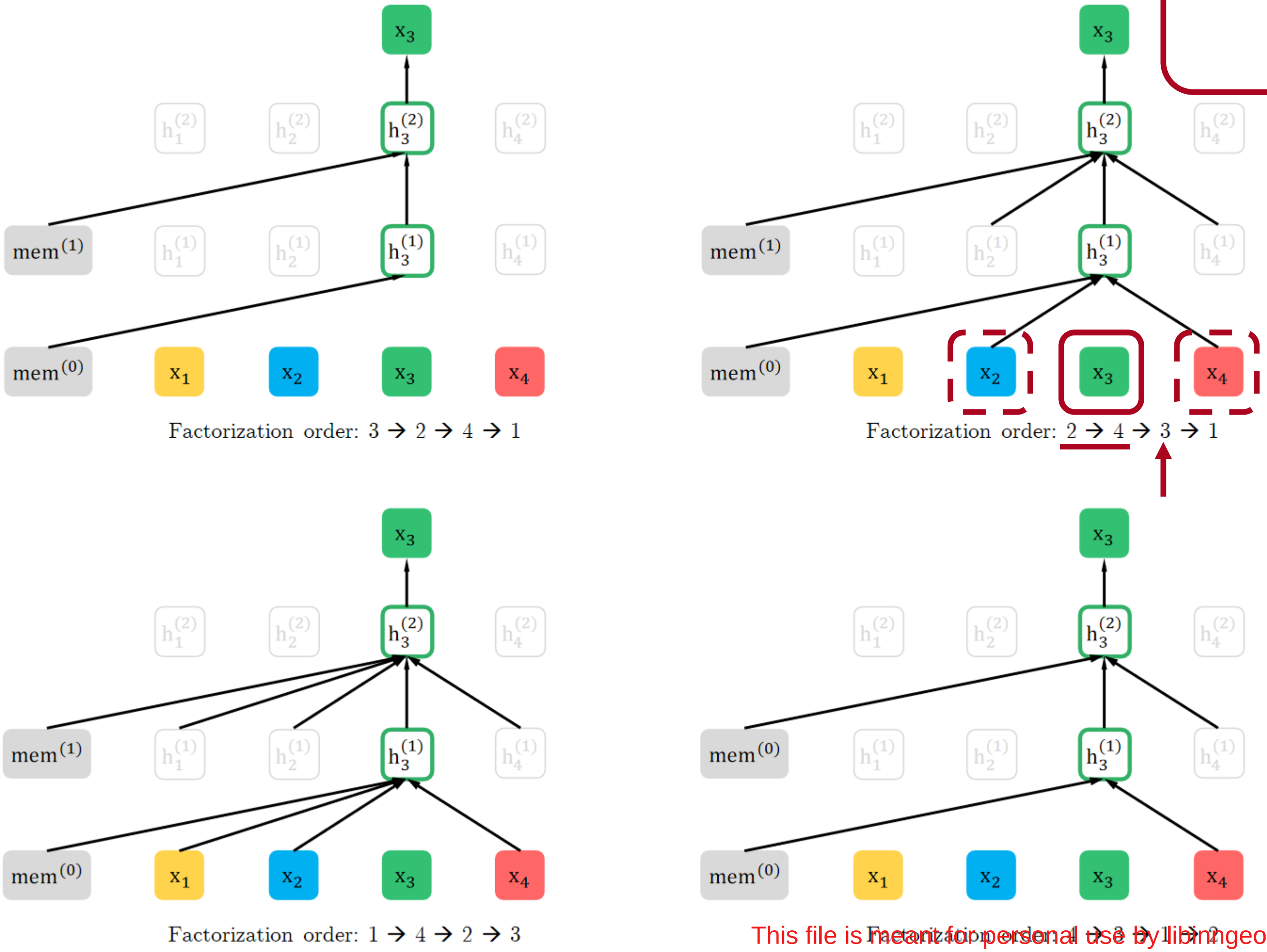
Yang et al., 2019

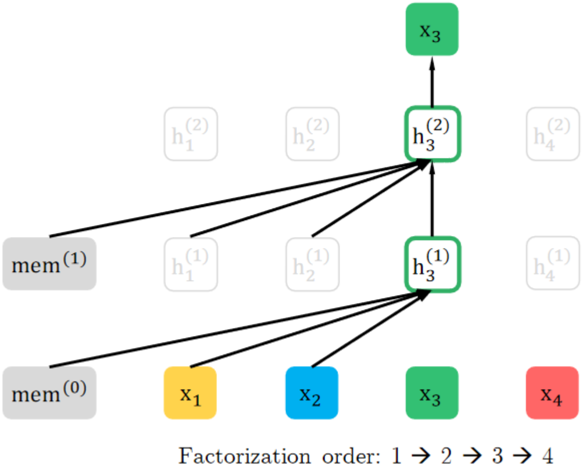
This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Alternative Pre-training Tasks (Cont.)

Permutation Language Modeling

The permutation happens in the Masked Self-Attention.
E.g., in this order, token 1 is not attended.





Original Order

Yang et al., 2019

This file is meant for personal use by libingeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Pre-trained Transformers For Long Sequences

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Agenda

In this session, we will discuss:

- Pre-trained Transformers for Long Sequences

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

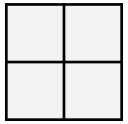
THE UNIVERSITY OF ARIZONA

Pre-trained Transformers for Long Sequences

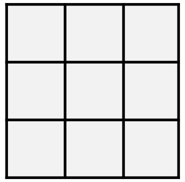
- Memory and computational requirements of self-attention grow quadratically:

Self-Attention Matrix:

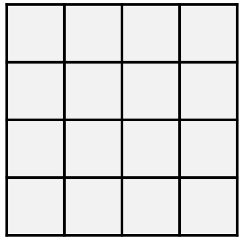
2 token sequence



3 token sequence



4 token sequence



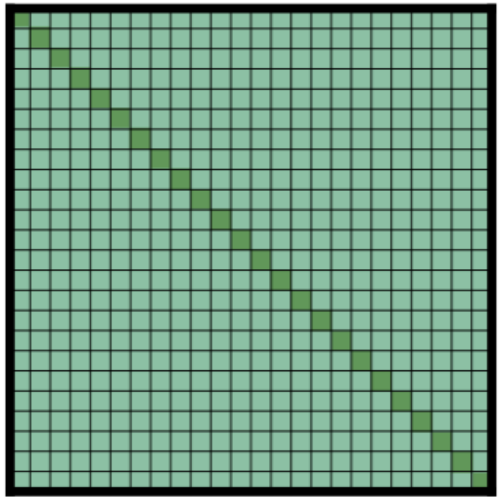
- Pre-trained Transformers are typically limited to an input of 512 tokens.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

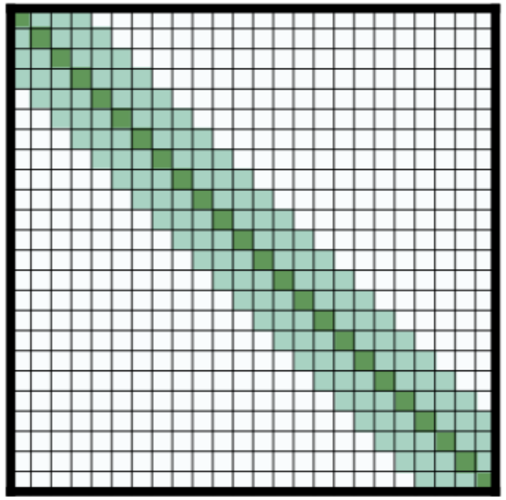
Pre-trained Transformers for Long Sequences (Cont.)

- Sparse self-attention patterns:

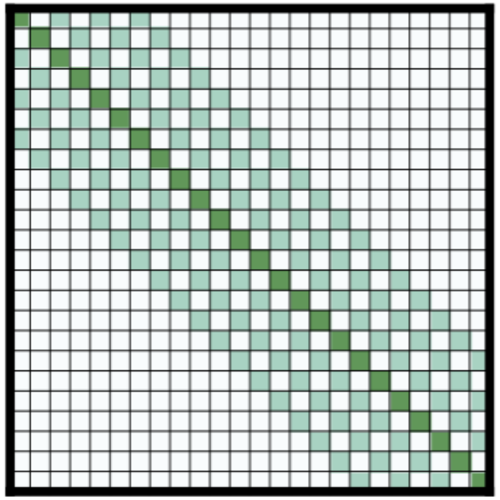
Beltagi et al., 2020



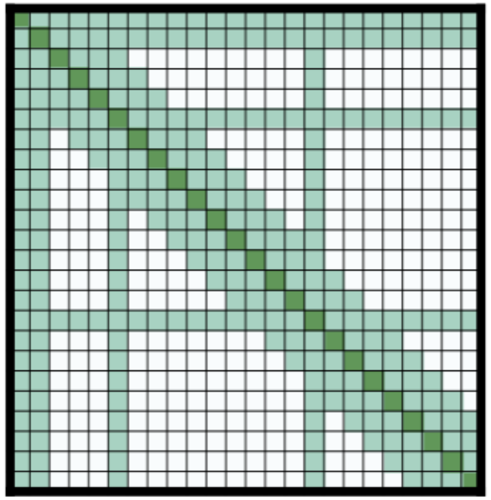
Original full attention



Sliding window



Dilated sliding window



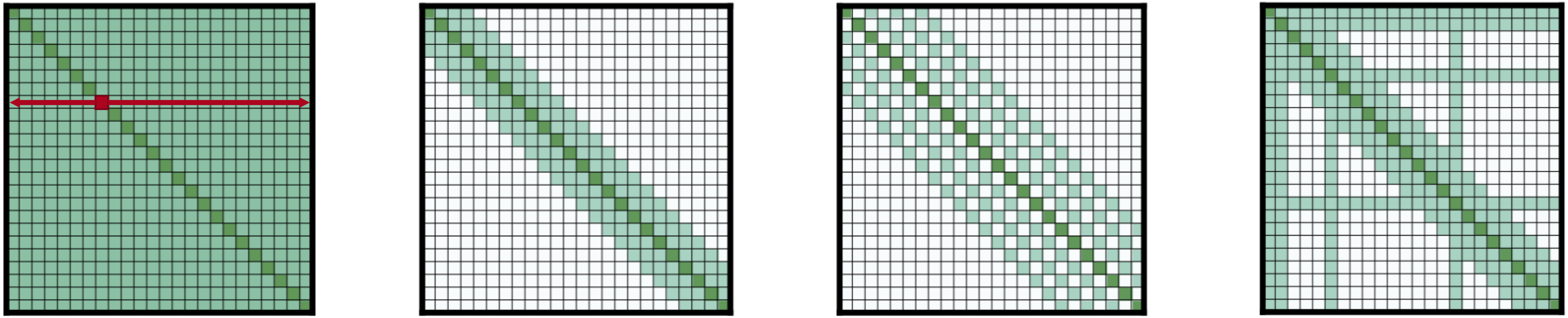
Global + sliding window

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Pre-trained Transformers for Long Sequences (Cont.)

- Sparse self-attention patterns:

Beltagi et al., 2020



Original full attention | Sliding window | Dilated sliding window | Global + sliding window

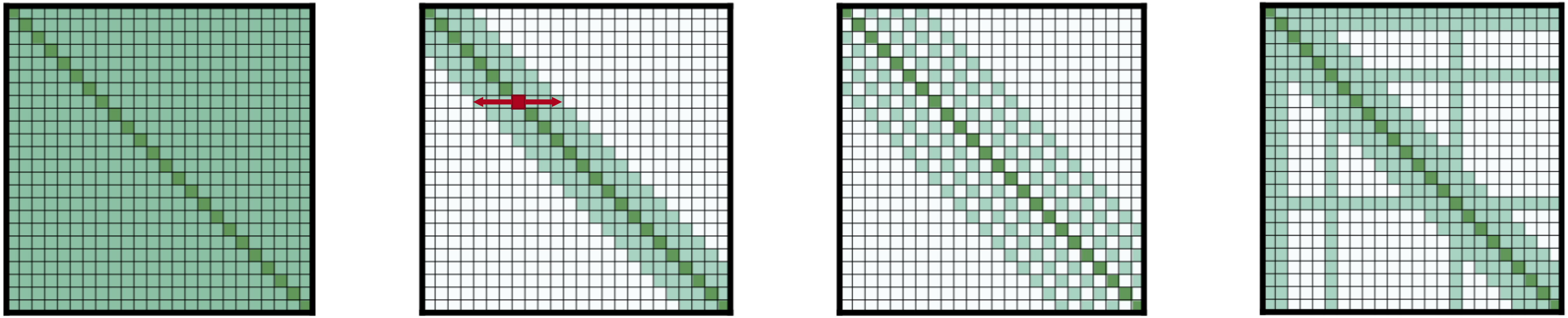
Each token attends to every token in the sequence.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Pre-trained Transformers for Long Sequences (Cont.)

- Sparse self-attention patterns:

Beltagi et al., 2020



Original full attention | Sliding window | Dilated sliding window | Global + sliding window

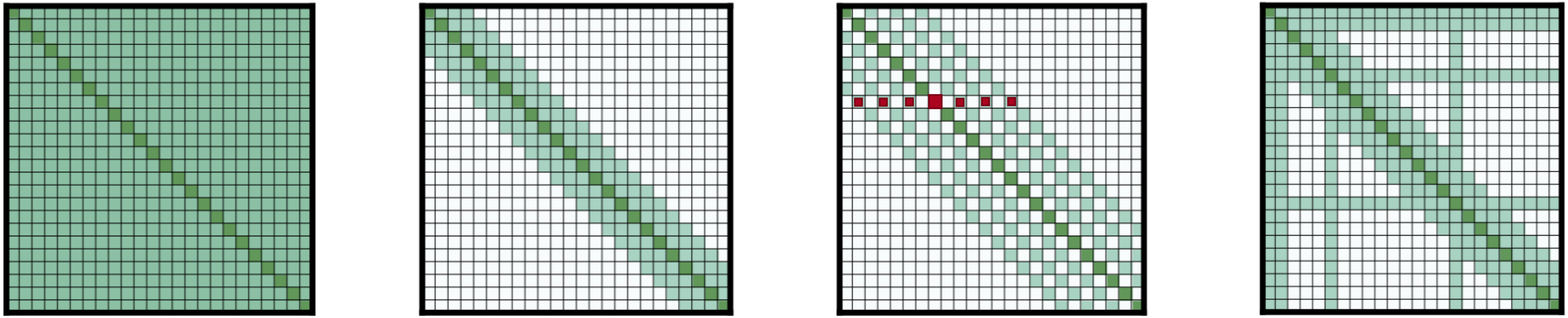
Each token attends only to a window of w tokens.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Pre-trained Transformers for Long Sequences (Cont.)

- Sparse self-attention patterns:

Beltagi et al., 2020



Original full attention | Sliding window | Dilated sliding window | Global + sliding window

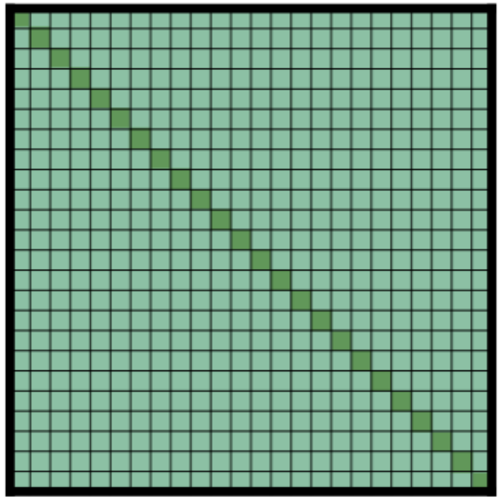
Each token attends only to a window of w tokens separated by a distance d.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

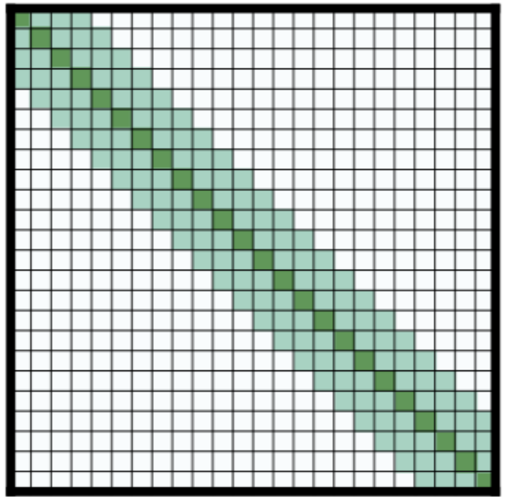
Pre-trained Transformers for Long Sequences (Cont.)

- Sparse self-attention patterns:

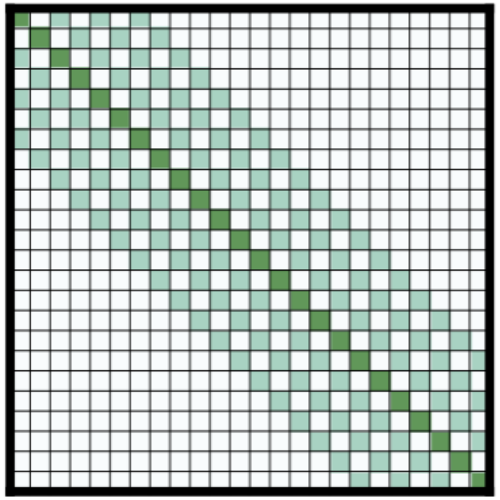
Beltagi et al., 2020



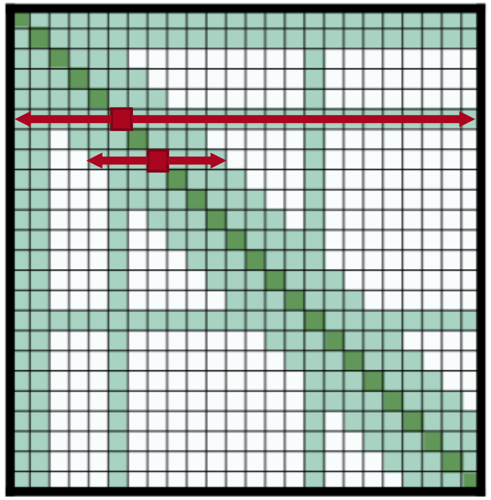
Original full attention



Sliding window



Dilated sliding window



Global + sliding window

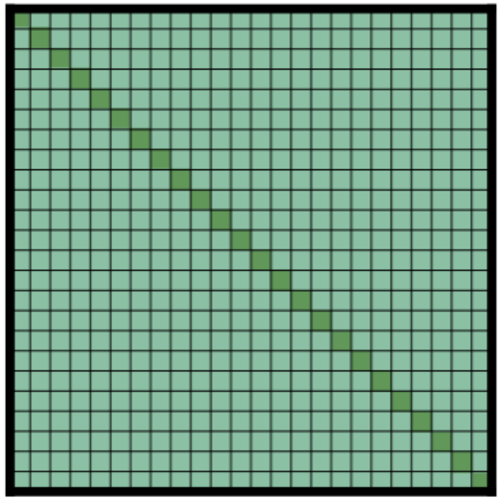
Use full attention for a few selected tokens (e.g., [CLS]) and sliding window for the rest.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

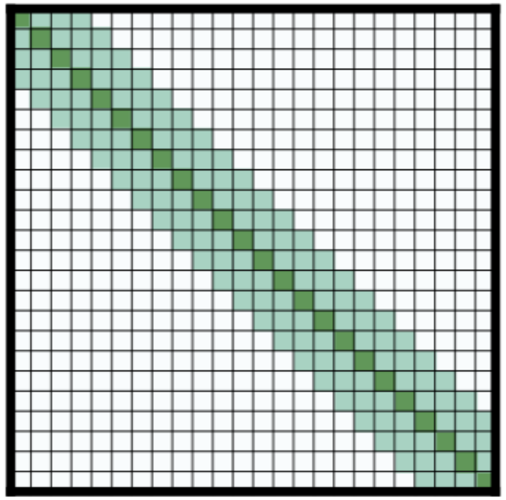
Pre-trained Transformers for Long Sequences (Cont.)

Beltagi et al., 2020

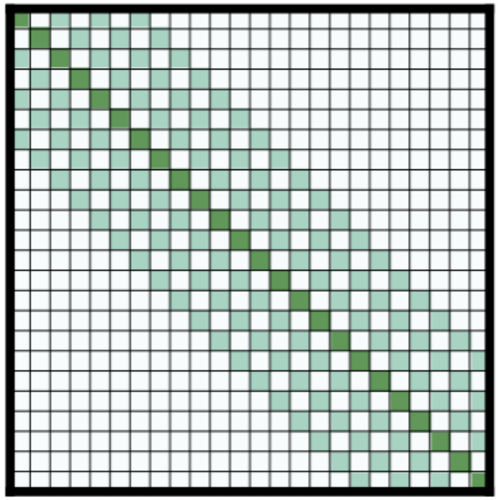
- Sparse self-attention patterns:
  - Memory and computational usage scales linearly



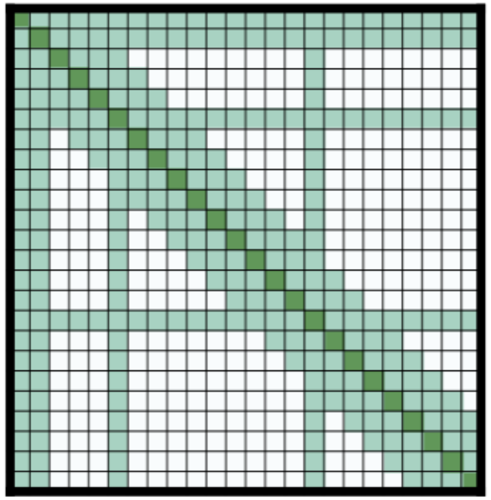
Original full attention



Sliding window



Dilated sliding window



Global + sliding window

- E.g., LongFormer, BigBird

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Efficient Fine-tuning

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

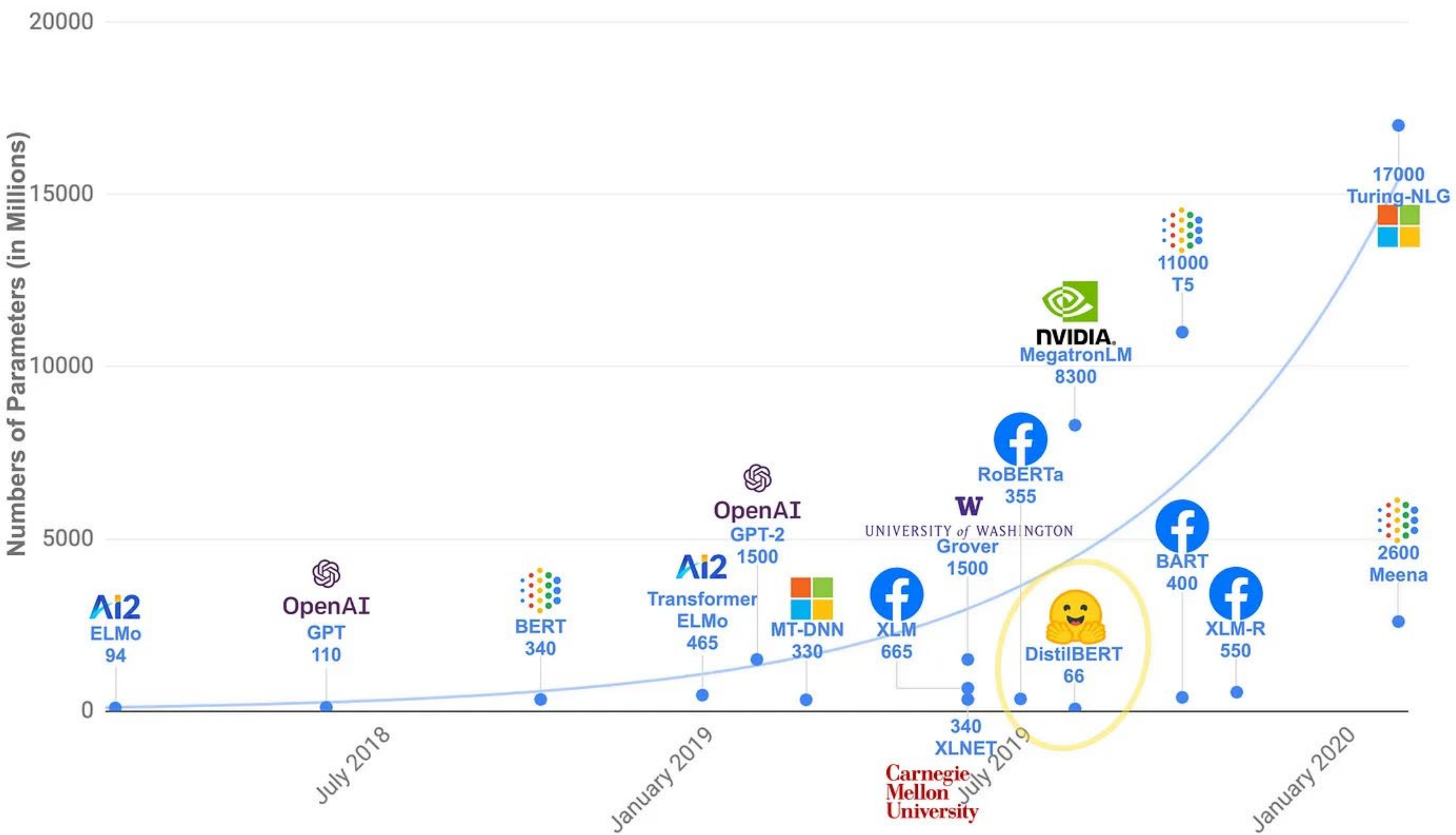
Agenda

In this session, we will discuss:

- Efficient Fine-tuning – Distillation and Adapter

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Efficient Fine-tuning

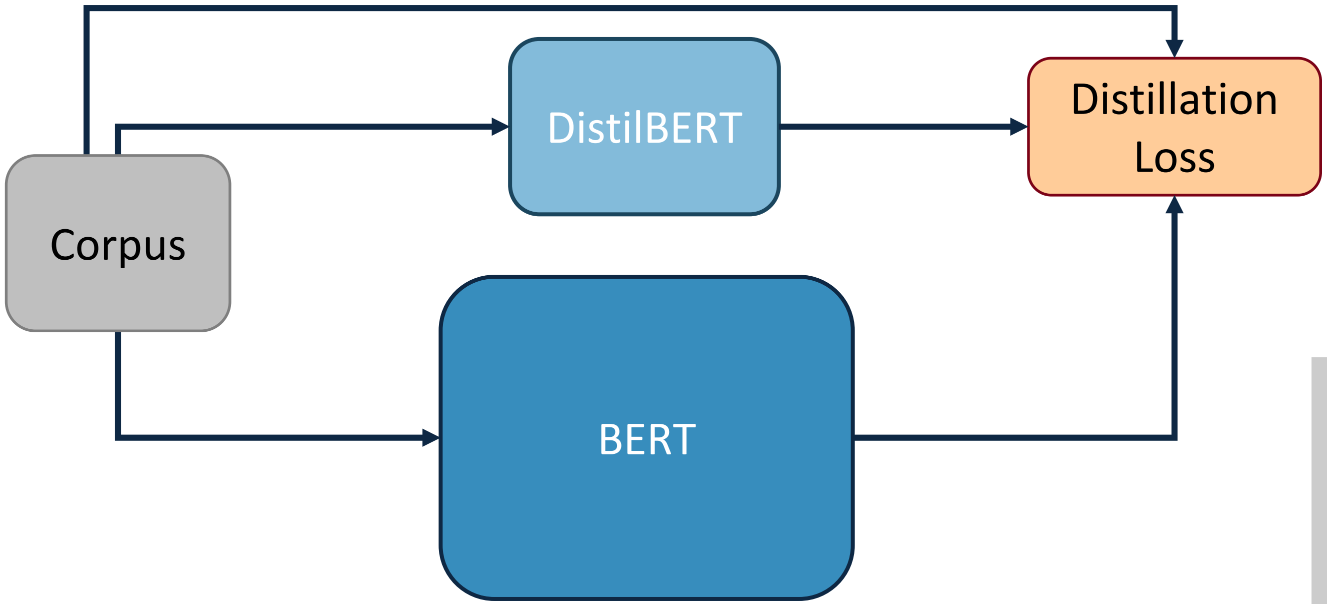


Sanh et al., 2020

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Efficient Fine-tuning: Distillation

- Fine-tuning a Large Language Model is expensive.
- Knowledge Distillation:
  - Obtain a compact version of a larger model while keeping the same performance.

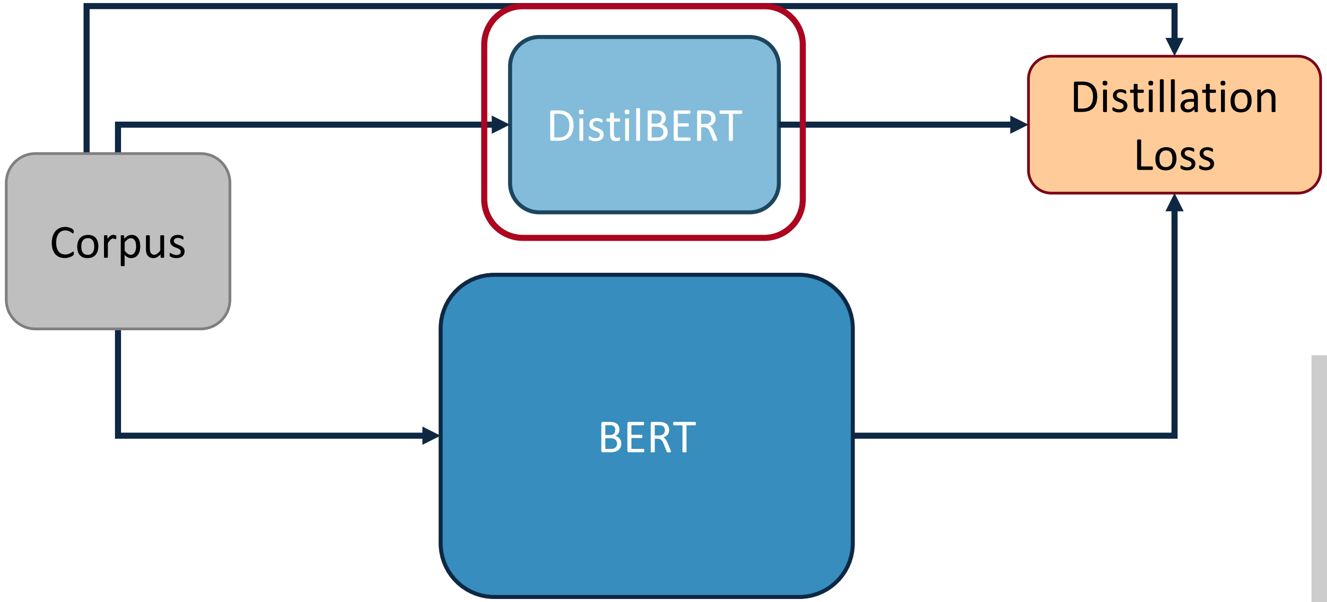


This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Efficient Fine-tuning: Distillation (Cont.)

- Fine-tuning a Large Language Model is expensive.
- Knowledge Distillation:
  - Obtain a compact version of a larger model while keeping the same performance.

Starting from a pre-trained BERT, pre-train a smaller version: same architecture but half of the layers.

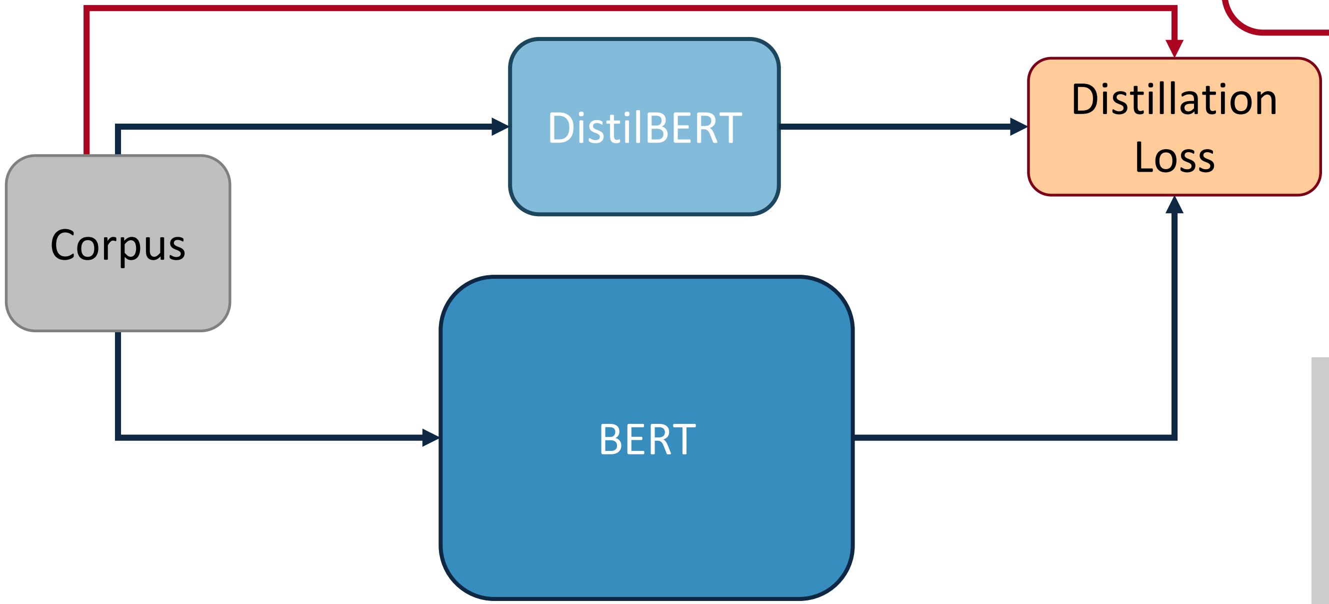


This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Efficient Fine-tuning: Distillation (Cont.)

- Fine-tuning a Large Language Model is expensive.
- Knowledge Distillation:
  - Obtain a compact version of a larger model while keeping the same performance.

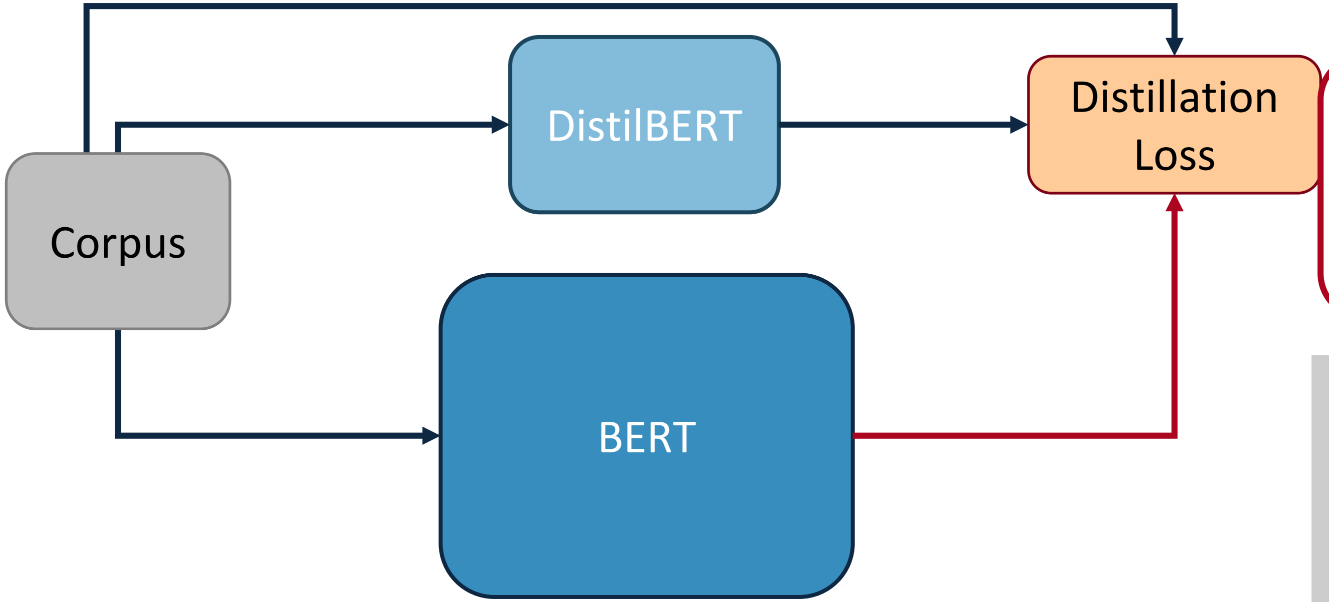
The model is trained for Masked Language Modeling.



This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Efficient Fine-tuning: Distillation (Cont.)

- Fine-tuning a Large Language Model is expensive.
- Knowledge Distillation:
  - Obtain a compact version of a larger model while keeping the same performance.



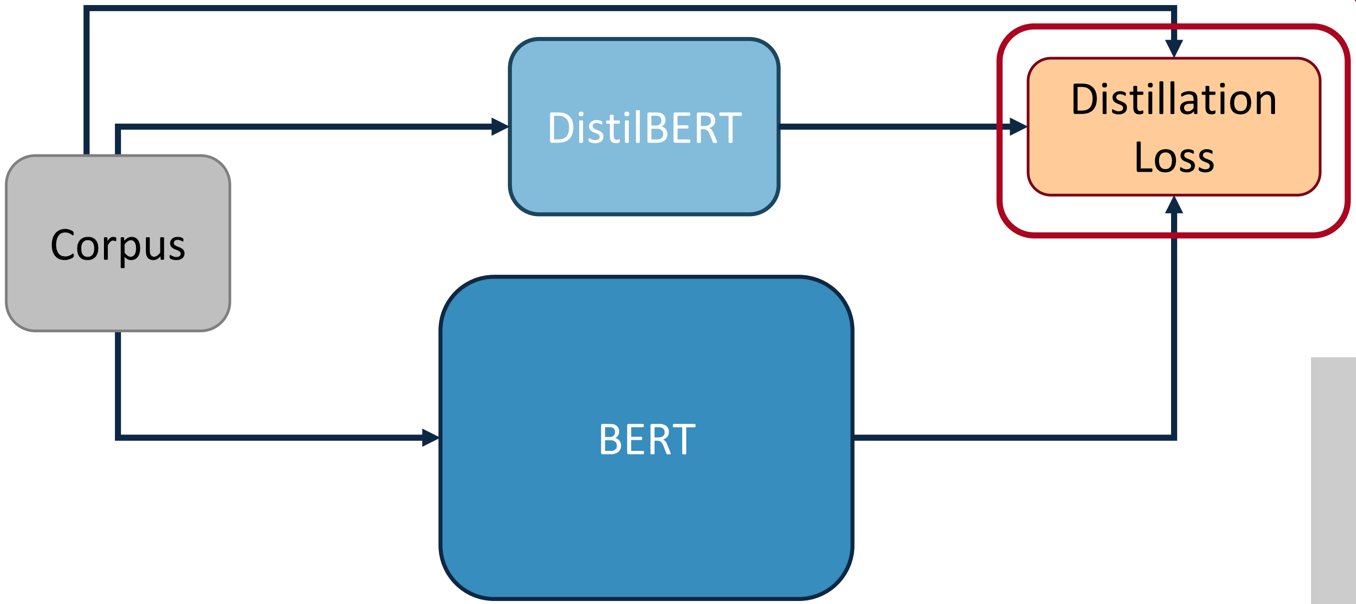
The model is trained to mimic BERT’s predicted probabilities.

This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Efficient Fine-tuning: Distillation (Cont.)

- Fine-tuning a Large Language Model is expensive.
- Knowledge Distillation:
  - Obtain a compact version of a larger model while keeping the same performance.

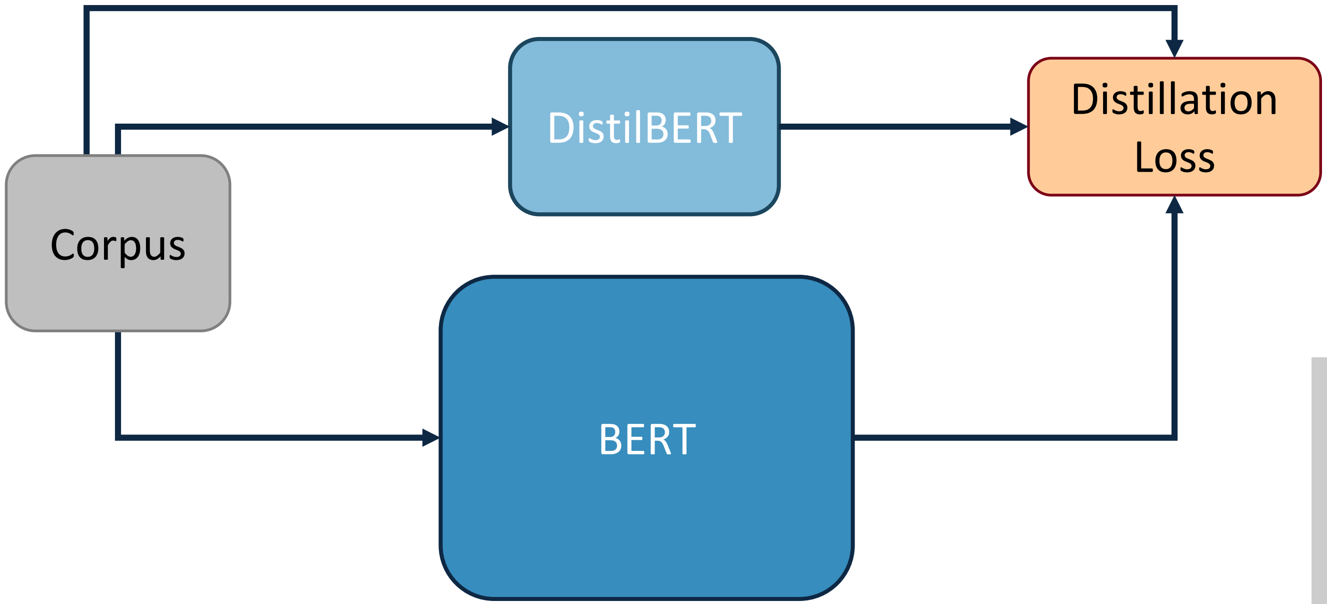
Knowledge Distillation transfers knowledge from the large model to the smaller.



This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Efficient Fine-tuning: Distillation (Cont.)

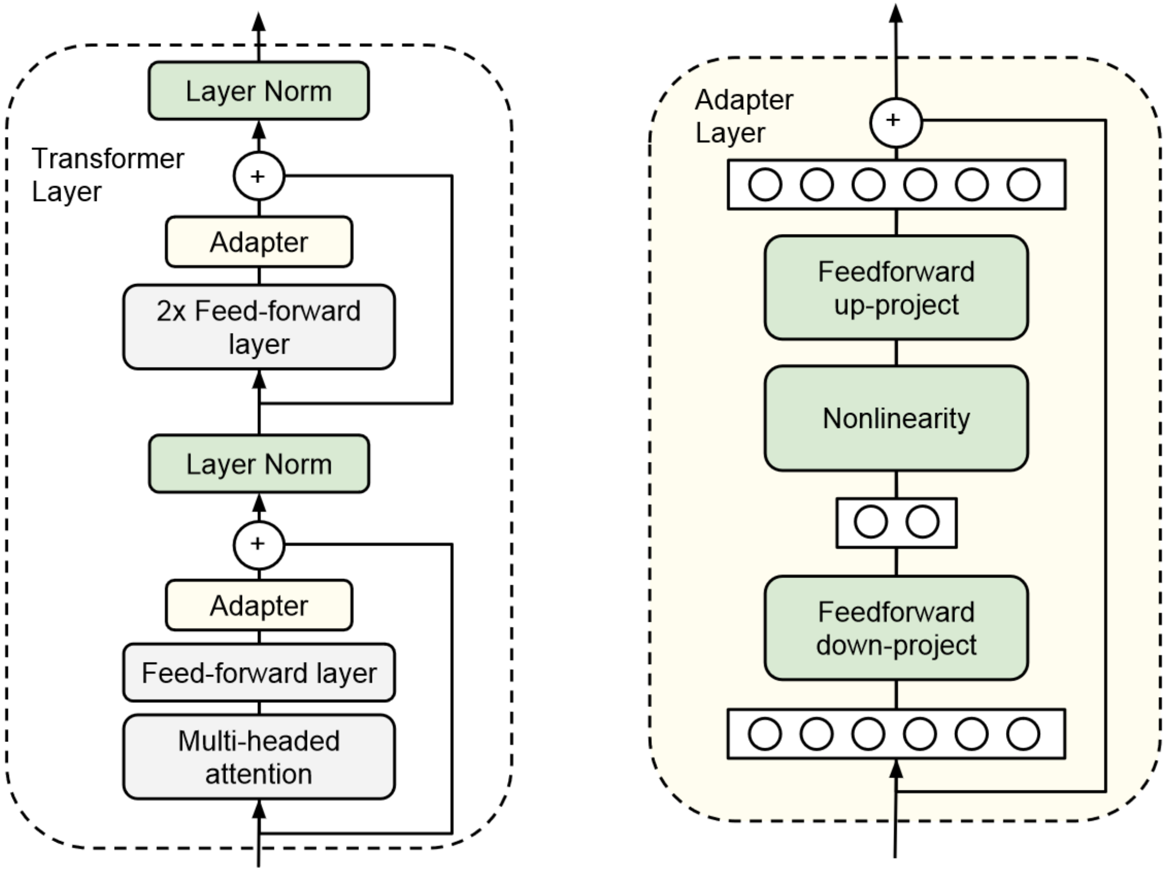
- Fine-tuning a Large Language Model is expensive.
- Knowledge Distillation:
  - Obtain a compact version of a larger model while keeping the same performance.
  - E.g.: DistilBERT, TinyBERT, DistilGPT2



This file is meant for personal use by libinngeorge@gmail.com only.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.
Sharing or publishing the contents in part or full is liable for legal action.

Efficient Fine-tuning: Adapters

- Fine-tuning an entire Large Language Model is parameter inefficient.
- Adapter modules:



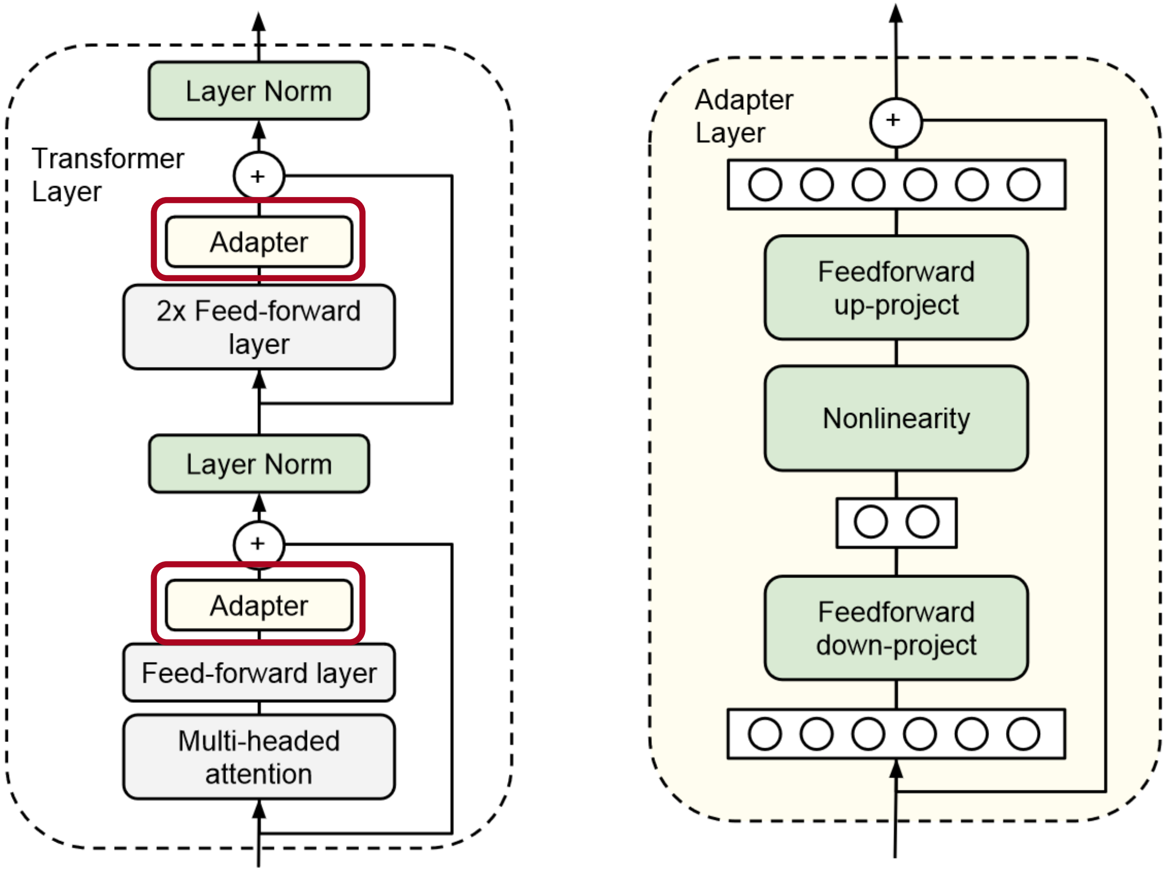
Houlsby et al., 2019

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Efficient Fine-tuning: Adapters (Cont.)

- Fine-tune an entire Large Language Model is parameter inefficient.
- Adapter modules:

During fine-tuning, add adapter modules to the pre-trained model.

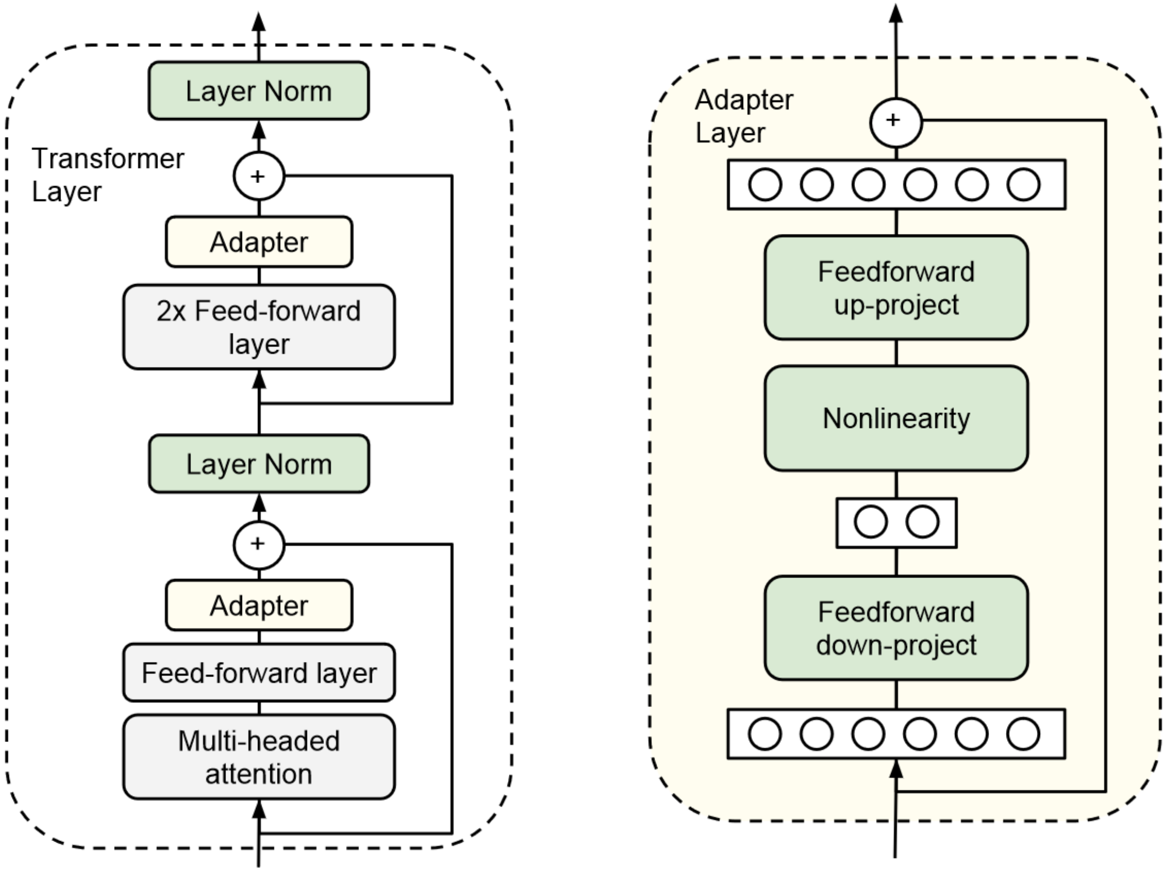


Houlsby et al., 2019

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Efficient Fine-tuning: Adapters (Cont.)

- Fine-tune an entire Large Language Model is parameter inefficient.
- Adapter modules:



The adapter contains fewer parameters than the attention.

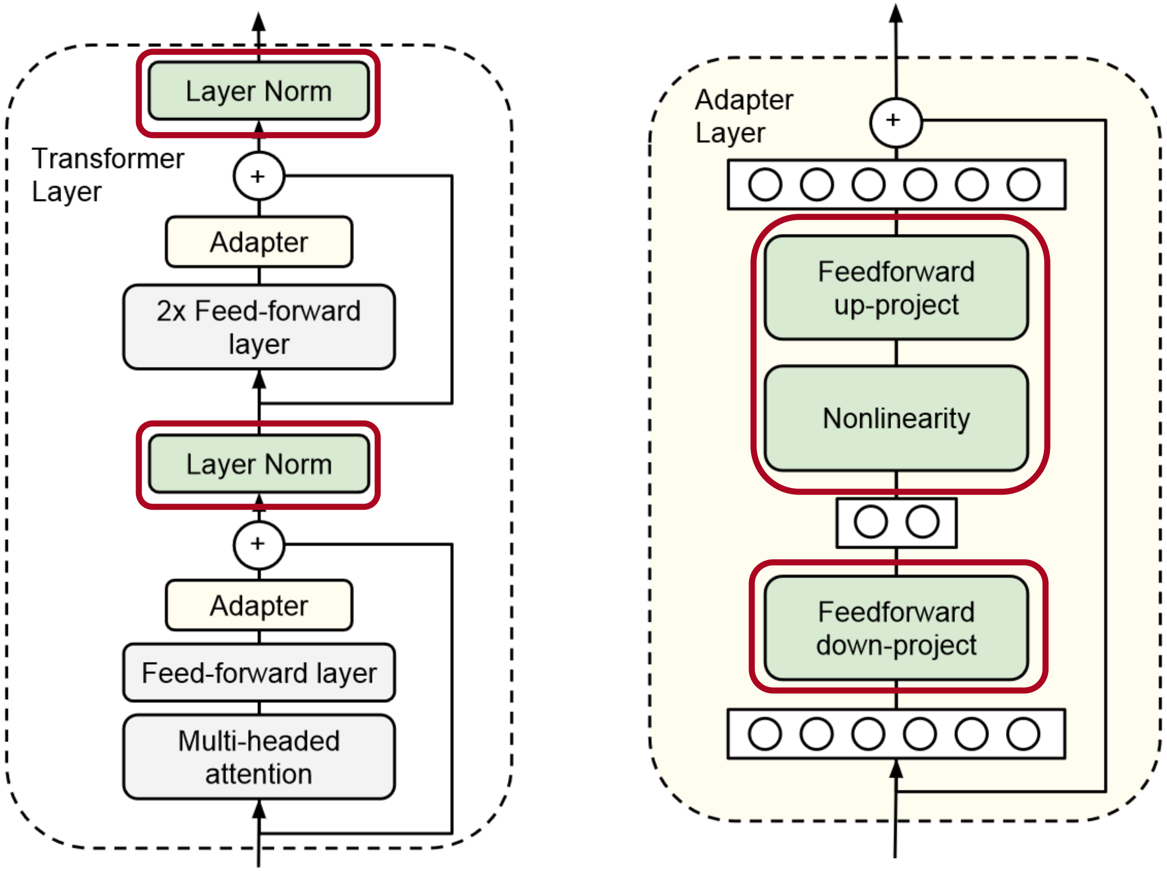
Houlsby et al., 2019

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.

Efficient Fine-tuning: Adapters (Cont.)

- Fine-tune an entire Large Language Model is parameter inefficient.
- Adapter modules:

Only these layers are trained during fine-tuning.



Houlsby et al., 2019

This file is meant for personal use by libinngeorge@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.
Proprietary content. ©University of Arizona. All Rights Reserved. Unauthorized use or distribution prohibited.