Pioneering Transformers

Danna Gurari
University of Colorado Boulder
Spring 2025

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2025/

Review

- Last lecture:
  - Transformer overview
  - Self-attention
  - Common transformer ingredients
  - Pioneering transformer: machine translation
  - Programming tutorial

- Assignments (Canvas):
  - Lab assignment 2 due earlier today
  - Problem set 4 (last one!) due in one week

- Questions?

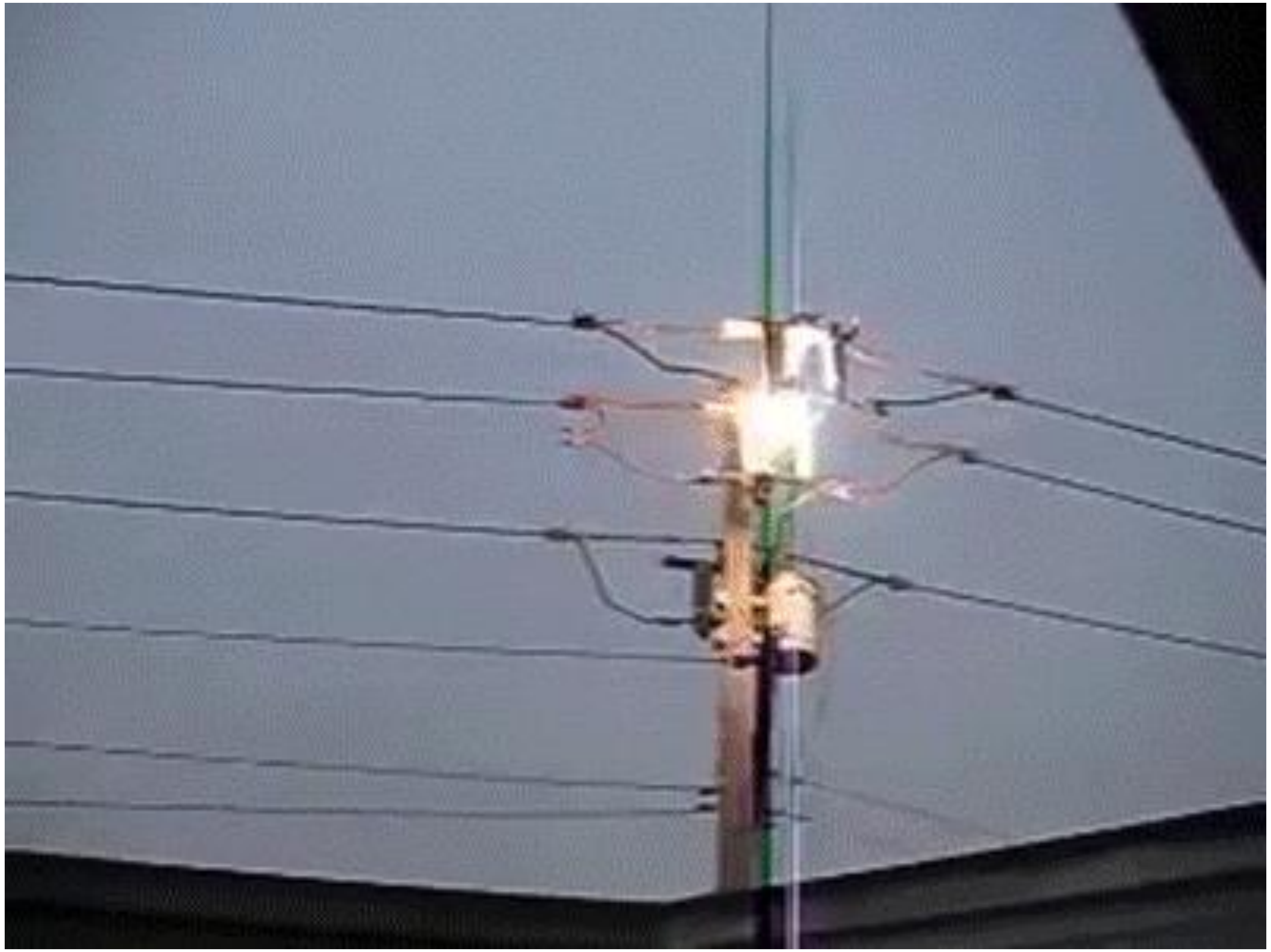
Today’s Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Programming tutorial

Today's Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Programming tutorial

Explosion of Transformers in Society

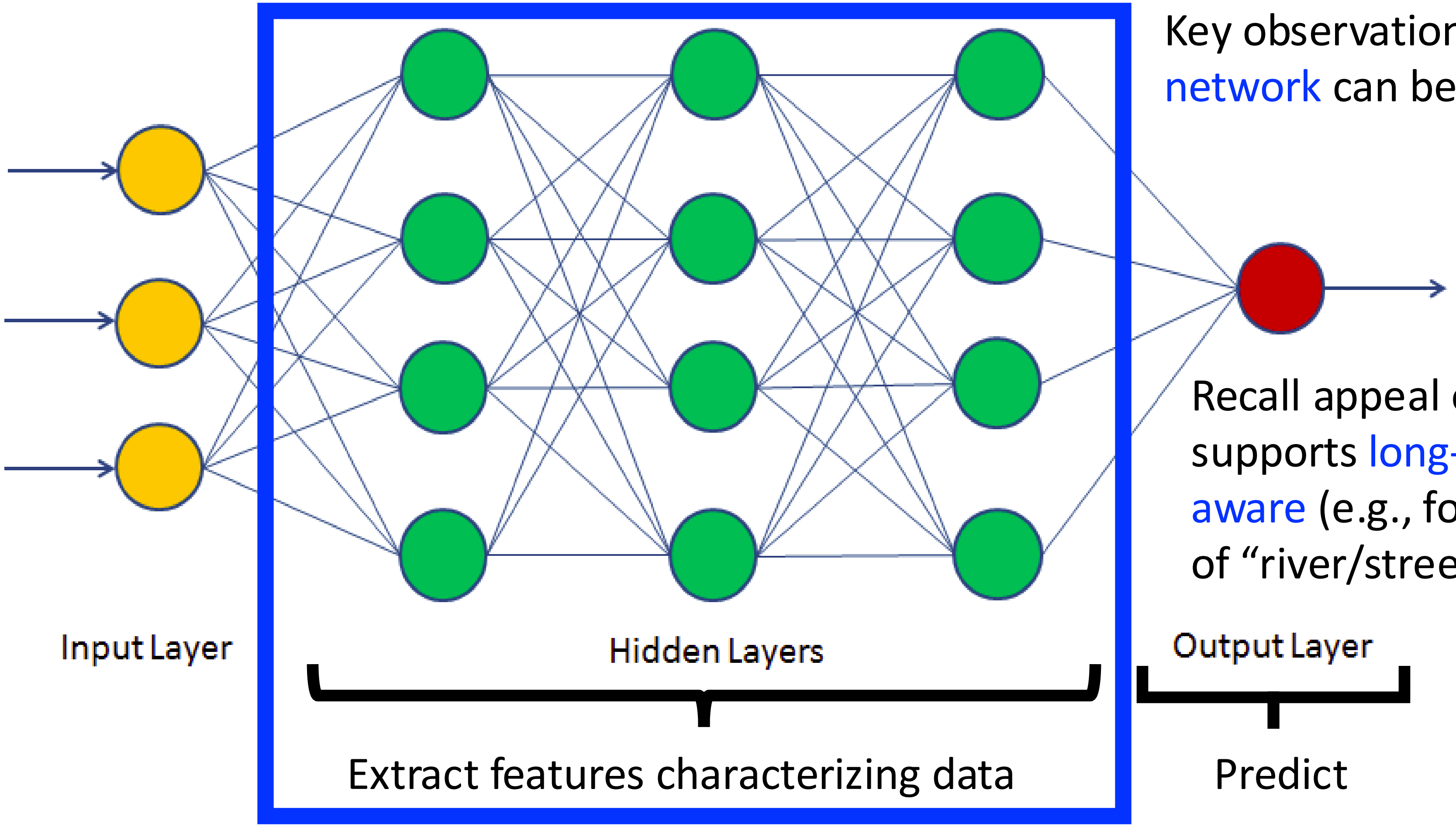


https://www.pinterest.com/pin/521784306804400819/

Initial Focus: Train Models for the Many Language Understanding Tasks

- Named entity recognition
- Recognizing semantically equivalent text, such as for pairs of questions or sentences
- Recognizing whether sentences are grammatically correct in English
- Question answering
- Machine translation
- And many more…

Key Idea 1: Fine-Tune Pre-Trained Models

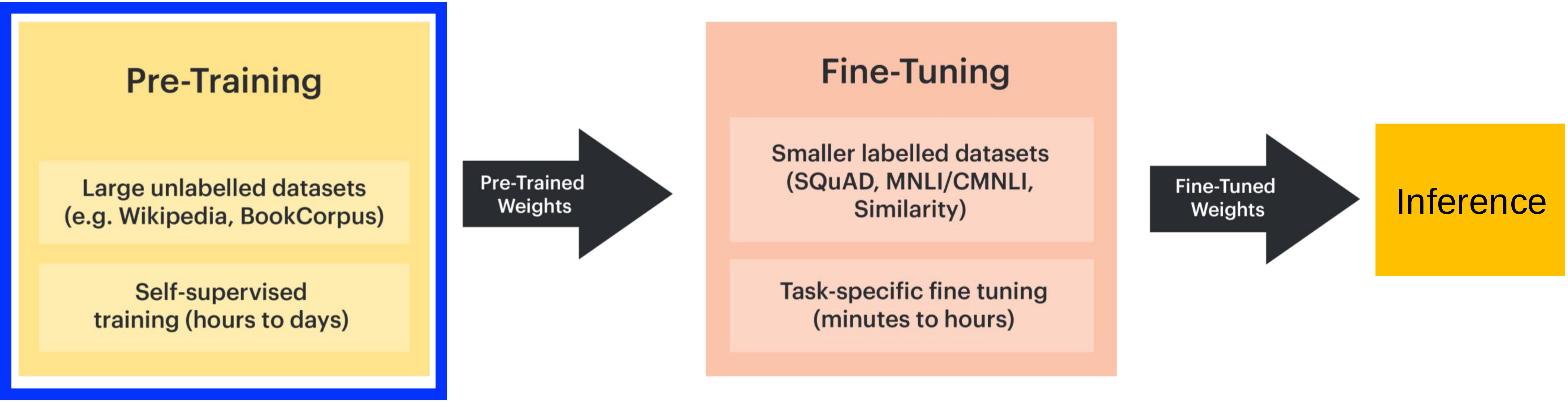


Key observation: features from a pretrained network can be useful for other datasets/tasks

Recall appeal of transformer representations: supports long-term dependencies AND context-aware (e.g., for multiple "bank" options, context of "river/street" determines choice)

https://www.datacamp.com/community/tutorials/neural-network-models-r

Key Idea 2: Data Is Supervision for Pretraining



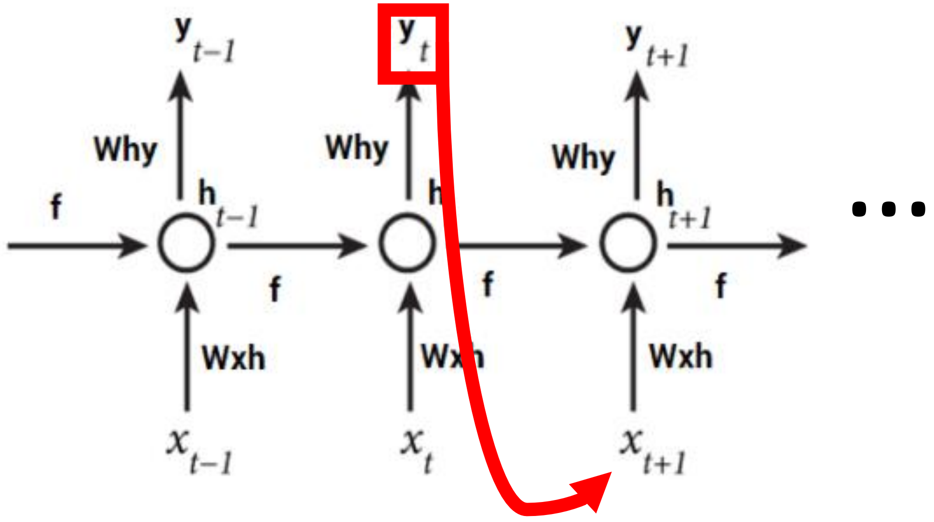
Since VERY challenging to collect large-scale human-annotated datasets, we don't!

https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Key Idea 2: Data Is Supervision (Self-Supervised)

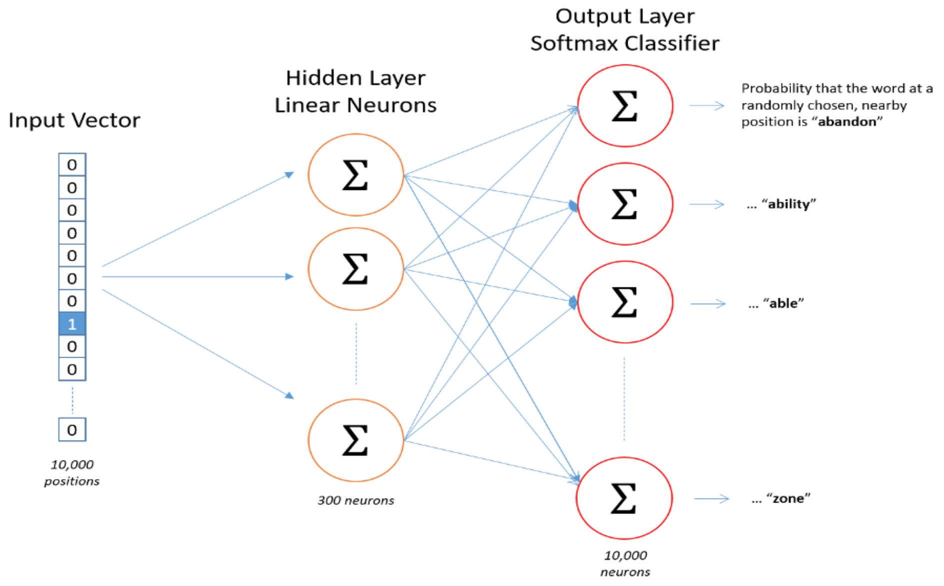
Recall, we already have seen self-supervised learning in our NLP-focused lectures:

RNNs
(e.g., predict next character)



https://www.analyticsvidhya.com/blog/2017/12
/introduction-to-recurrent-neural-networks/

Word embeddings
(e.g., predict nearby word for given word for word2vec)



https://towardsdatascience.com/word2vec-skip-
gram-model-part-1-intuition-78614e4d6e0b

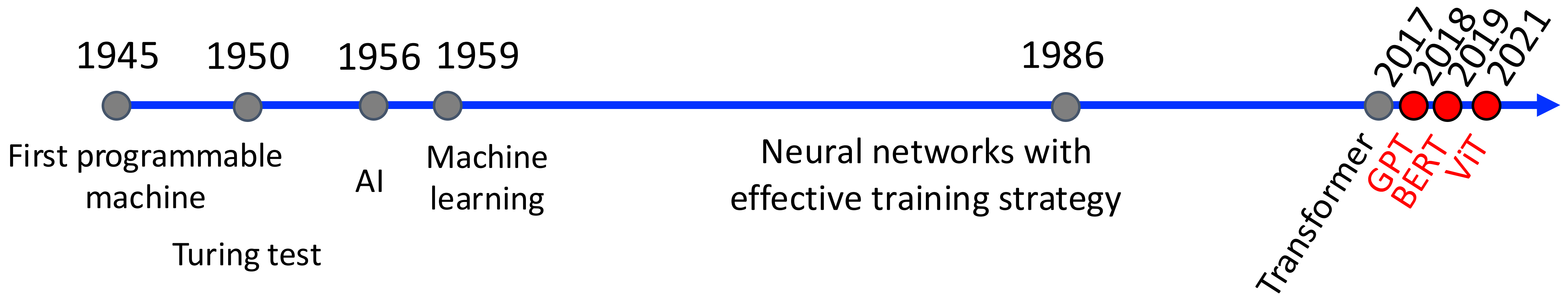
Key Idea 2: Data Is Supervision (Self-Supervised)

- Relatively Cheap
- Can Collect Data Fast

https://lovevery.com/community/blog/child-development/the-surprising-learning-power-of-a-household-mirror/

https://www.rockettes.com/blog/how-to-use-the-mirror-in-dance-class/

How to Develop Pre-trained Models?

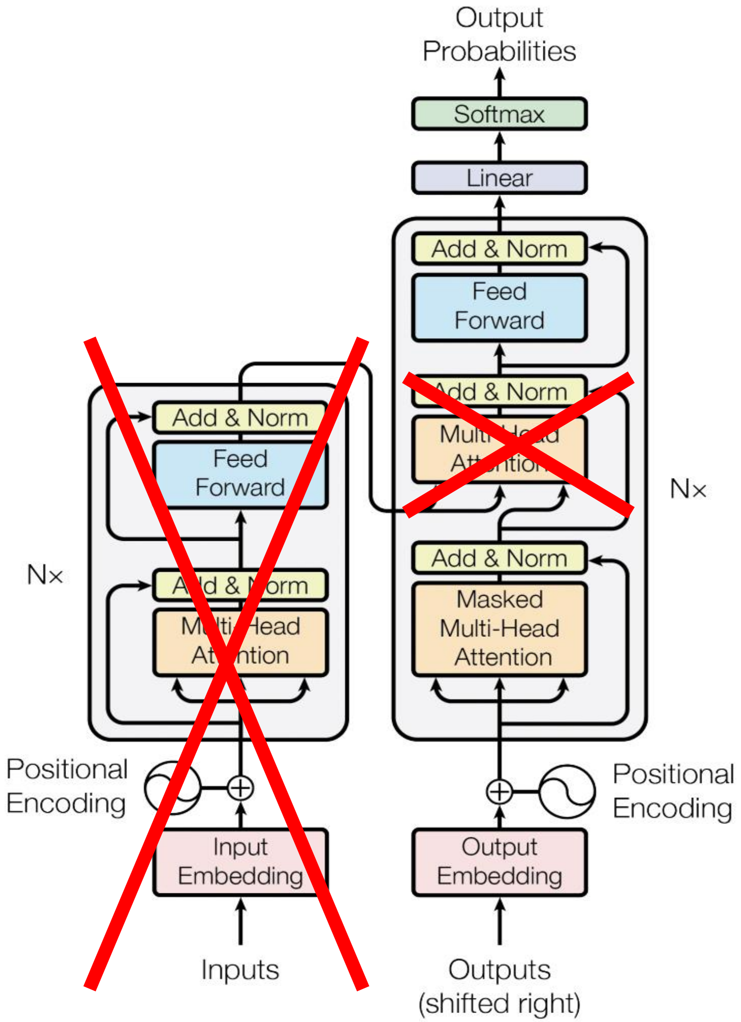


Today’s Topics

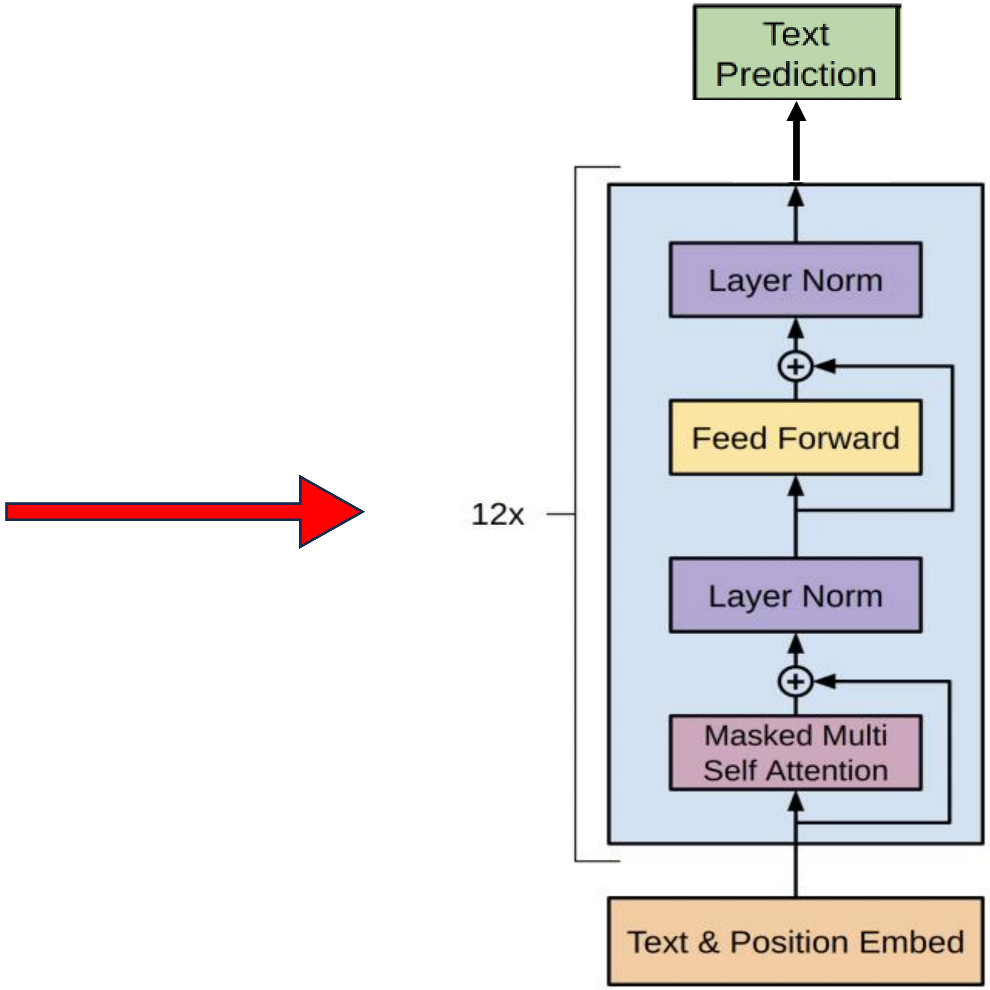
- Explosion of transformers
- GPT
- BERT
- ViT
- Programming tutorial

Architecture: Decoder of Pioneering Transformer

No encoder or cross-attention:



Vaswani et al. Attention Is All You Need. Neurips 2017.



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

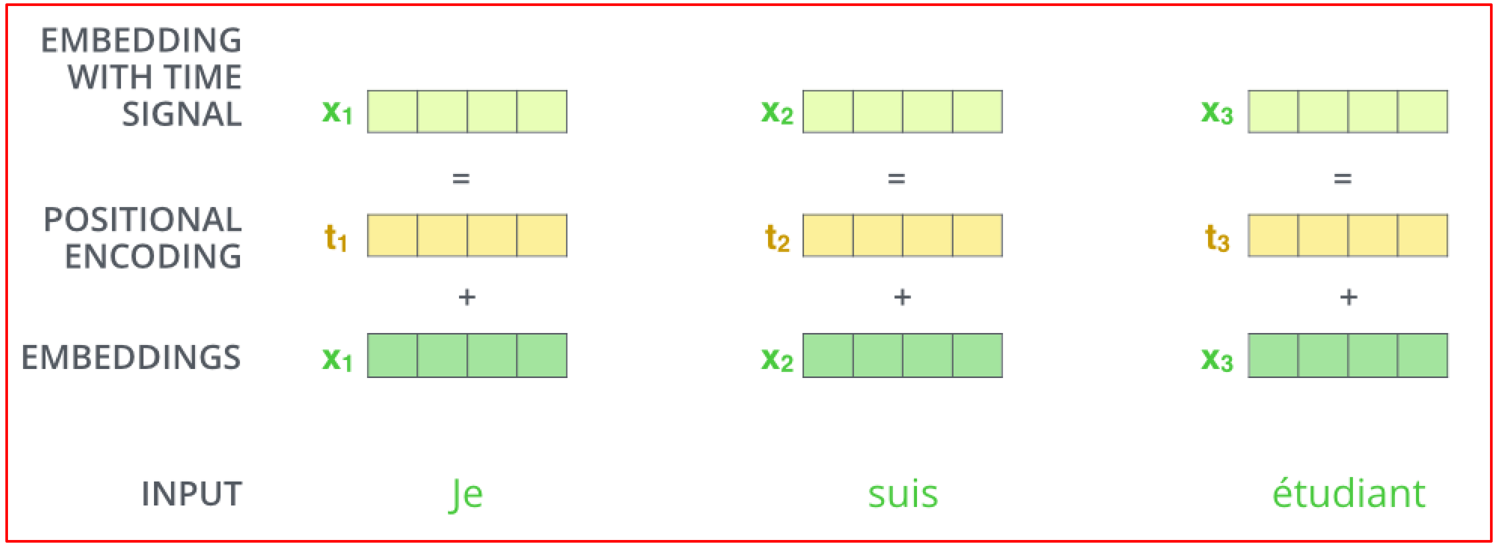
Recall: Input and Output

Softmax layer:

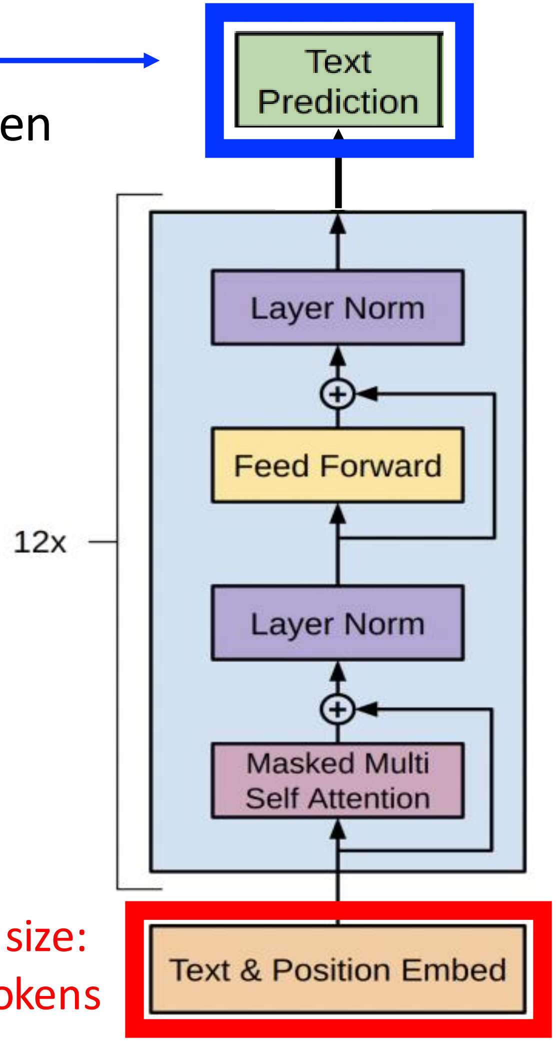
- Autoregressive: predicts next word token based on previously generated tokens

Vocabulary: created with byte pair encoding

Embedding matrix: learned during training



http://jalammar.github.io/illustrated-transformer/



Input size: 512 tokens

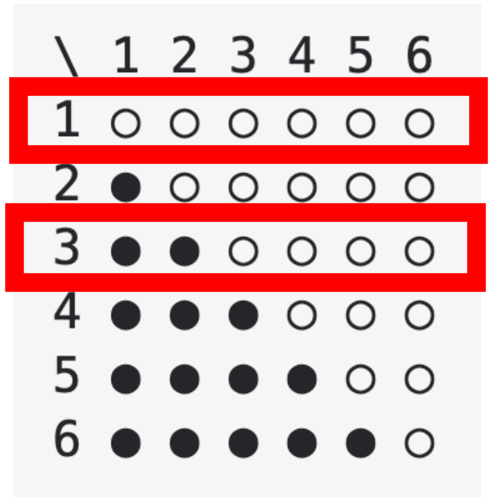
Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Recall: Masked Attention

Limit each word’s new representation to only reflect earlier words (mimics inference time when only previous tokens can be seen):

Query x Key

Query

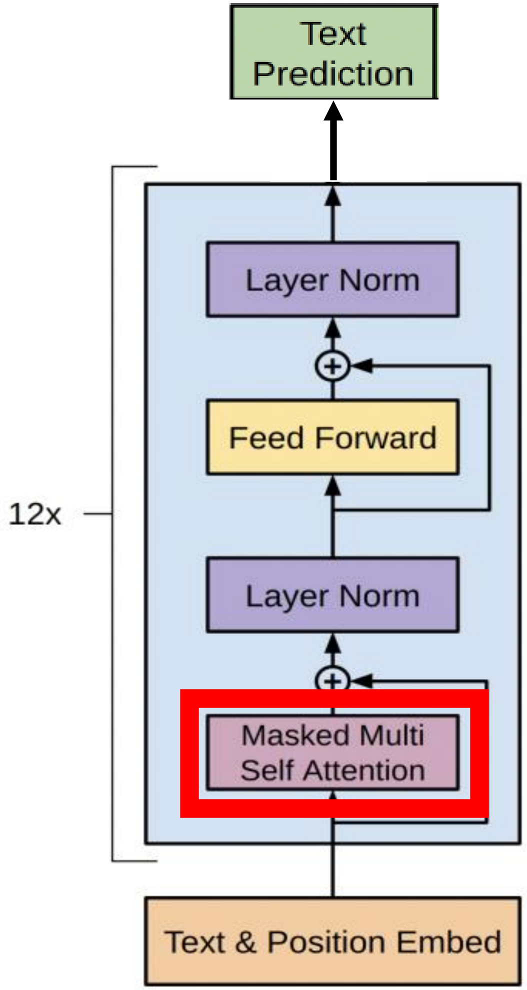


e.g., at <sos> token, no previous inputs

e.g., at 3rd token, two previous inputs

Masked scores are negative infinity so softmax result (i.e., attention weight) is 0

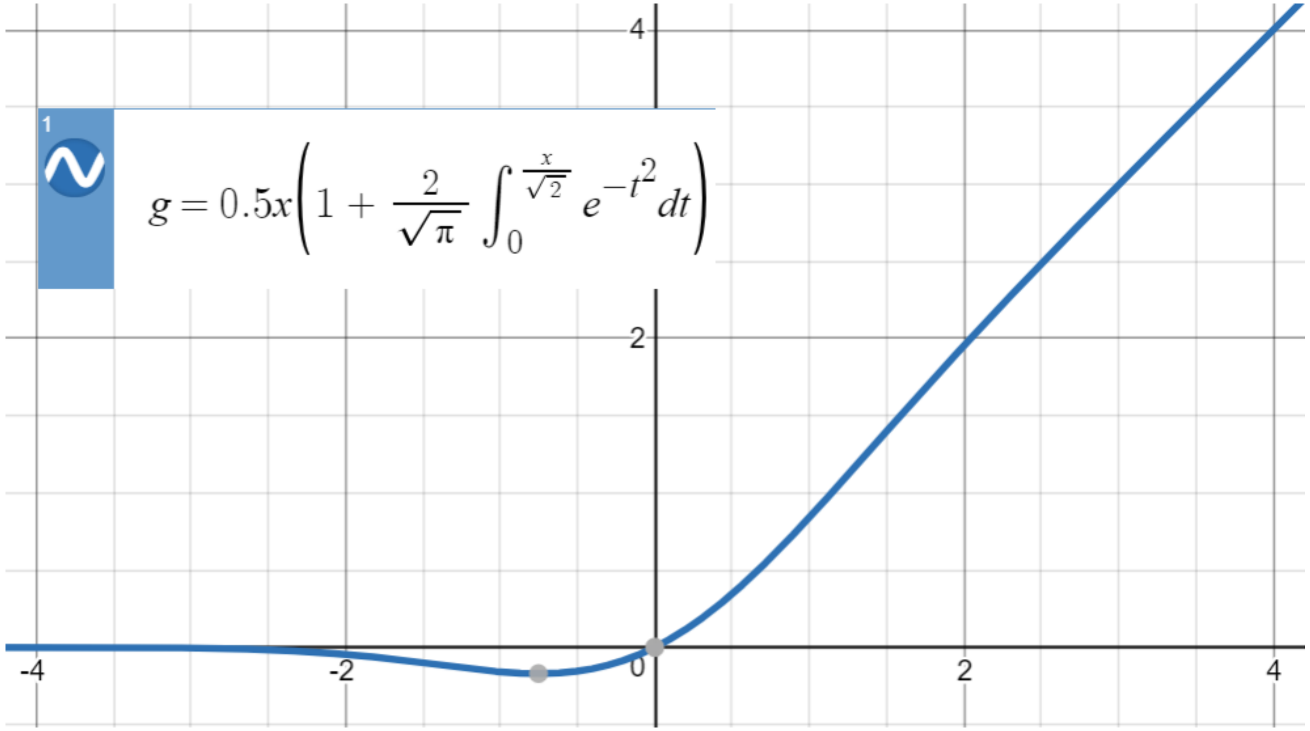
https://stackoverflow.com/questions/64799622/how-is-the-gpts-masked-self-attention-is-utilized-on-fine-tuning-inference



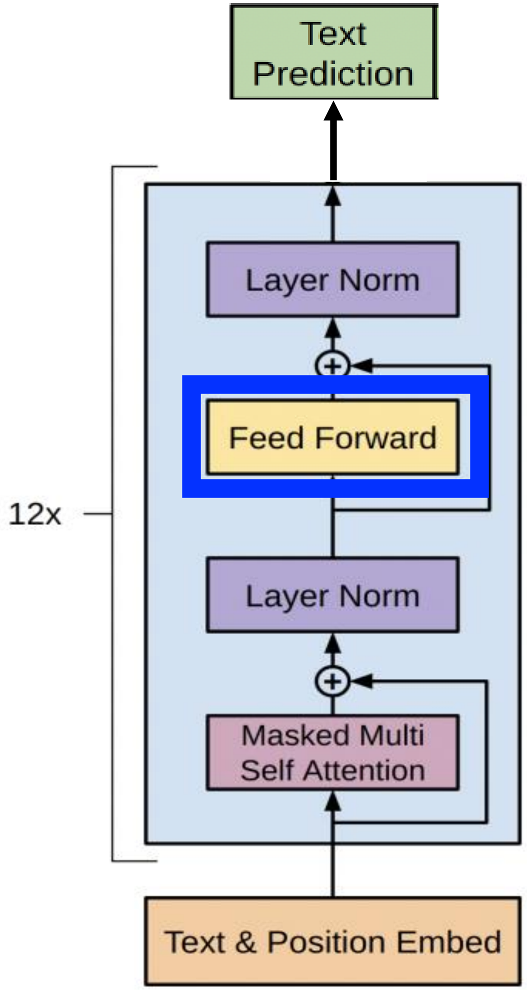
Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Architecture: Minor Tweak

Activation function: Gaussian error linear unit (GELU)



https://datascience.stackexchange.com/questions/49522/what-is-gelu-activation



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

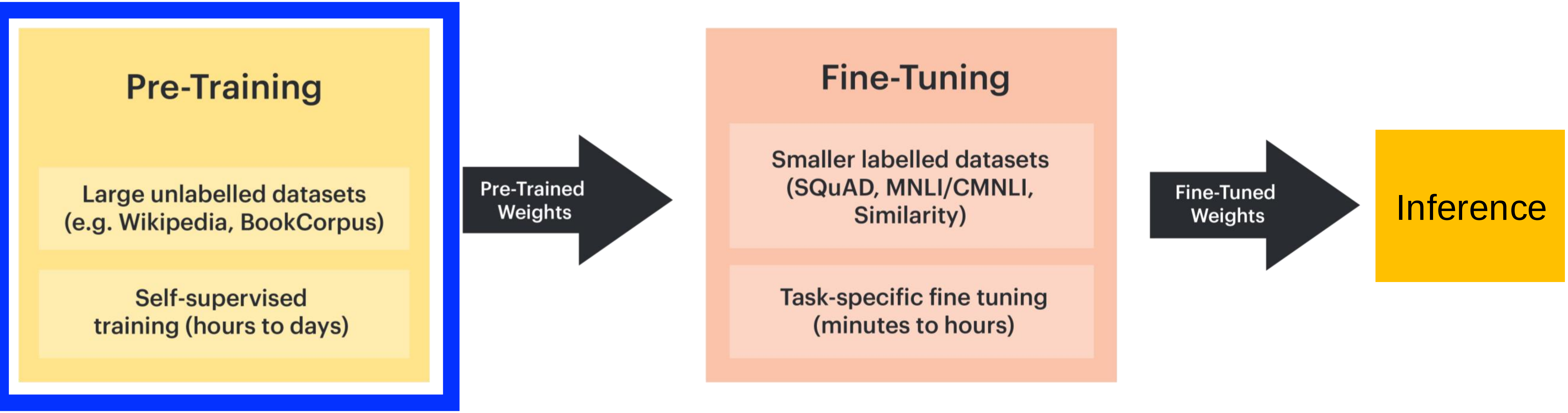
Key Ideas

- Learn good representations for downstream tasks: pretraining objective function
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

Key Ideas

- Learn good representations for downstream tasks: pretraining objective function
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

Why GPT? Generative Pre-Training



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Task: Predict Next Word Given Previous Ones

e.g.,

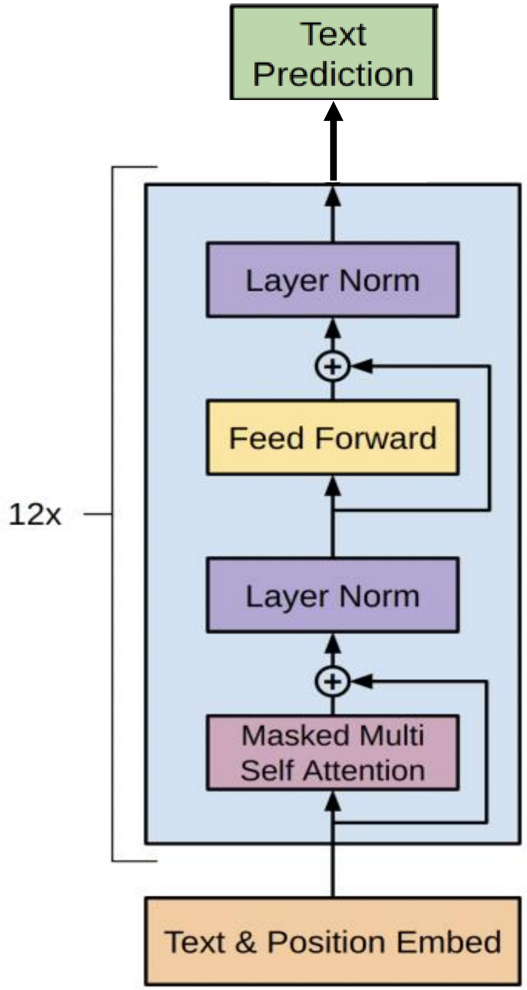
1. Background music from a ________
2. Many people danced around the ________
3. I practiced for many years to learn how to play the ________

Key Ideas

- Learn good representations for downstream tasks: pretraining objective function
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

Training

- Dataset/Task: predict next word using 800M words in BooksCorpus (>7,000 unpublished books)
- Mini-batch size: 64 sequences of 512 tokens each
- Regularization: dropout and L2 norm penalty
- Optimizer: Adam
- Training duration: 100 epochs

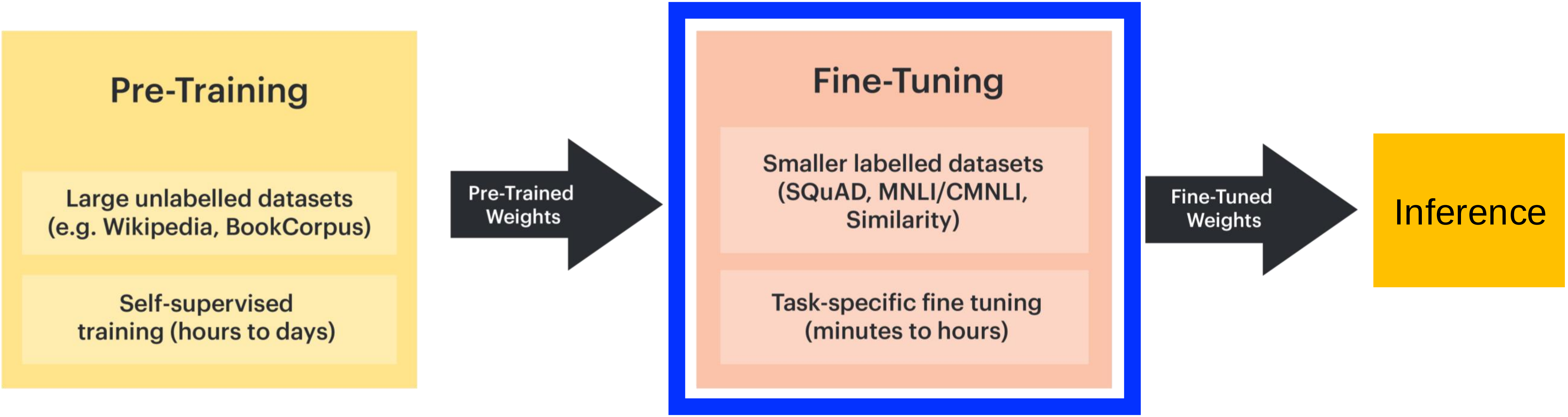


Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Key Ideas

- Learn good representations for downstream tasks: pretraining objective function
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

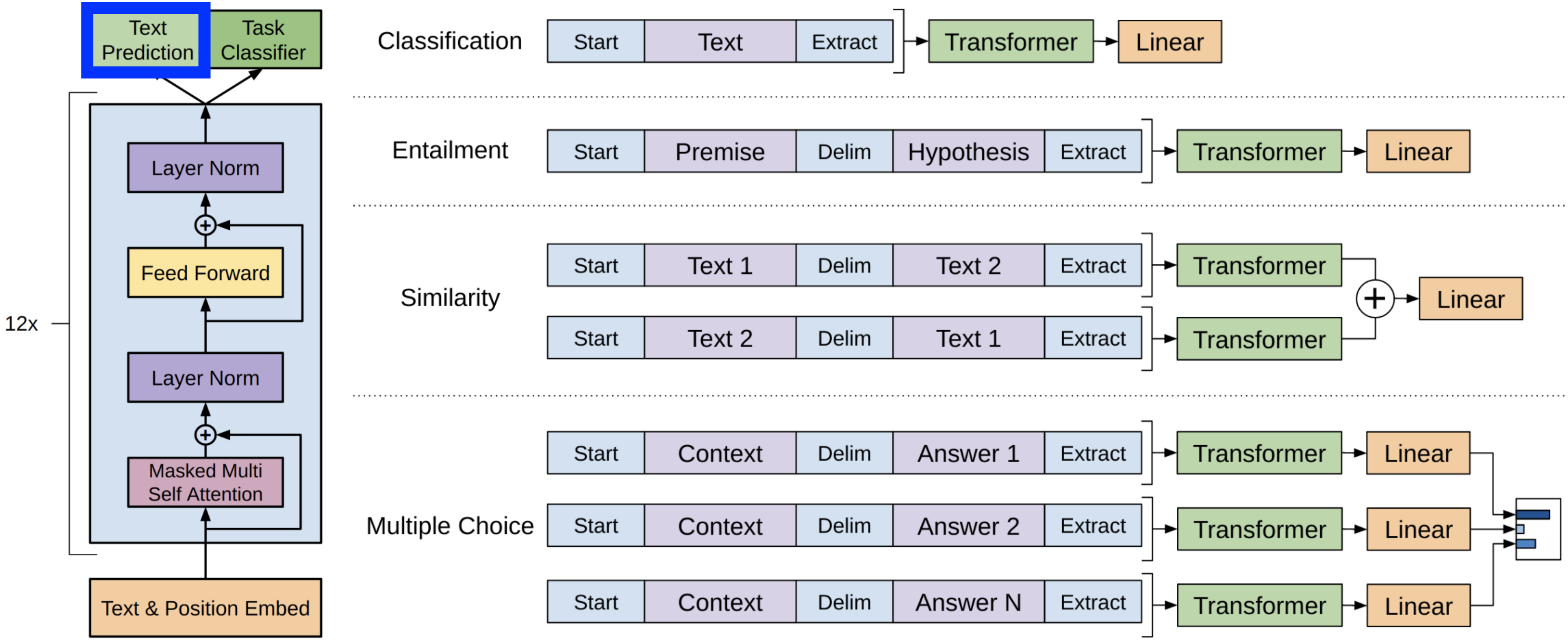
Fine-Tuning to Target Tasks



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Fine-Tuning to Target Tasks (Softmax Output)

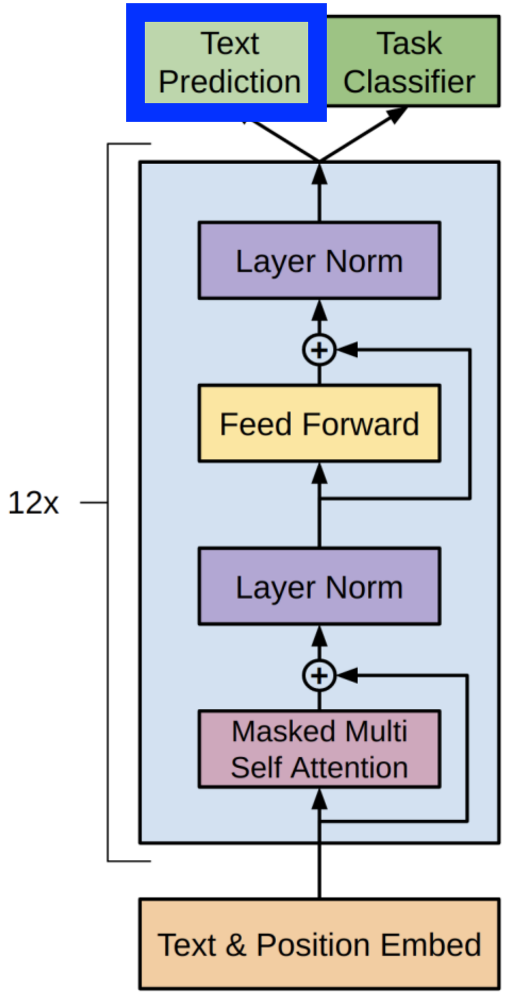
Auxiliary function improves performance

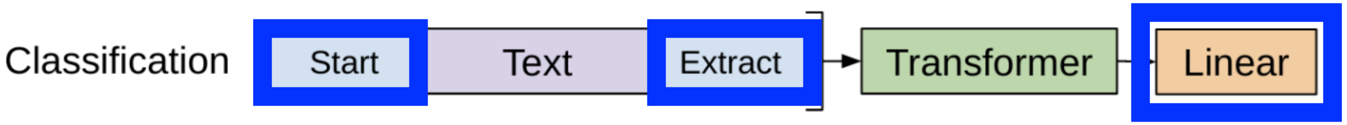


Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Fine-Tuning to Target Tasks (Softmax Output)

Auxiliary function improves performance





Start token

End token

Only new parameters

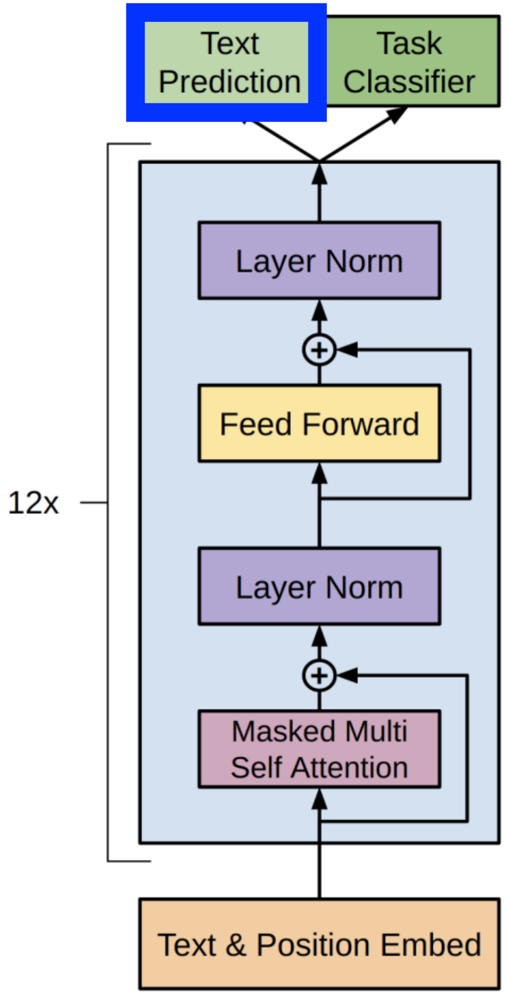
Supervised training to predict whether, for example, a:

- sentence is grammatical
- sentence from movie review has a positive vs negative sentiment

Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Fine-Tuning to Target Tasks (Softmax Output)

Auxiliary function improves performance



Start token

$ token

End token

Only new parameters

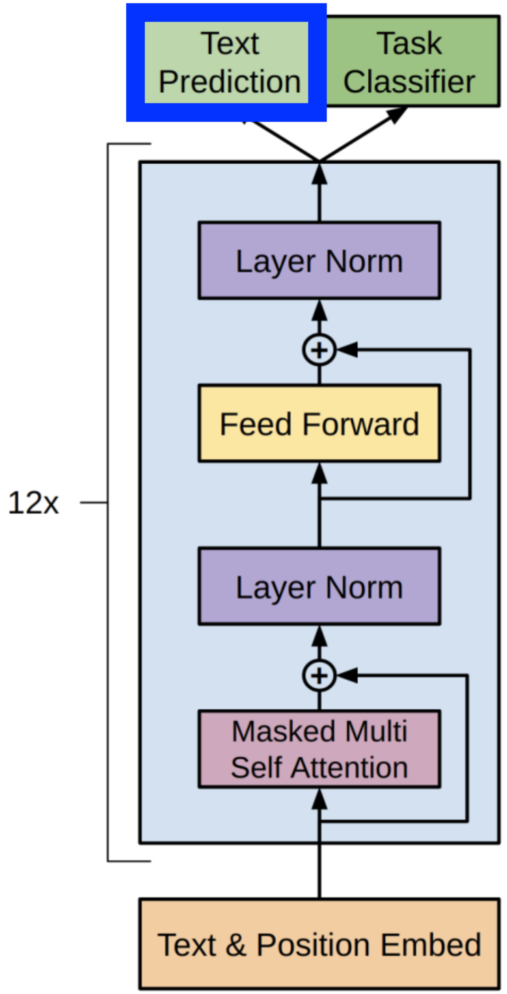
Supervised training for the natural language inference task, to classify the relationship of hypothesis to premise into three options; e.g.,

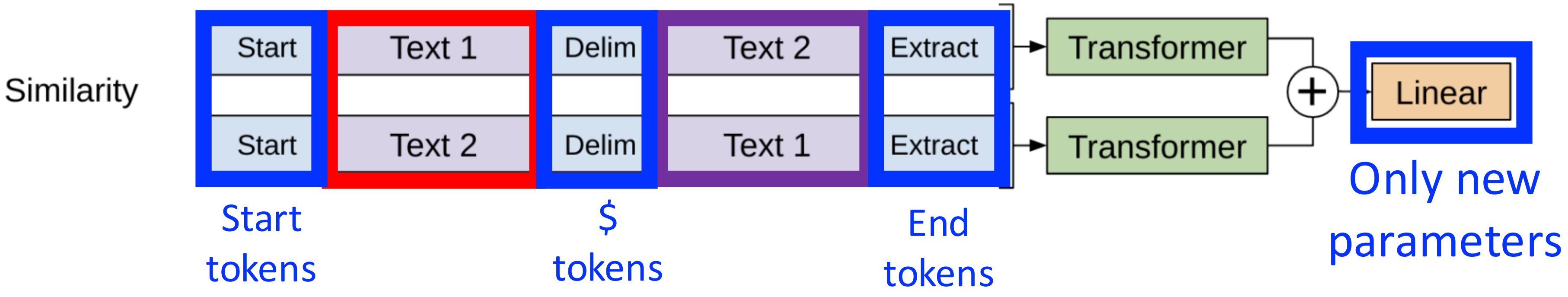
Relationship	**Premise**	**Hypothesis**
Entailment	Students quietly listen to the lecture	The lecturer is speaking.
Neutral	Students quietly listen to the lecture	The projector equipment is broken.
Contradict	Students quietly listen to the lecture	The students are lecturing.

Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Fine-Tuning to Target Tasks (Softmax Output)

Auxiliary function improves performance



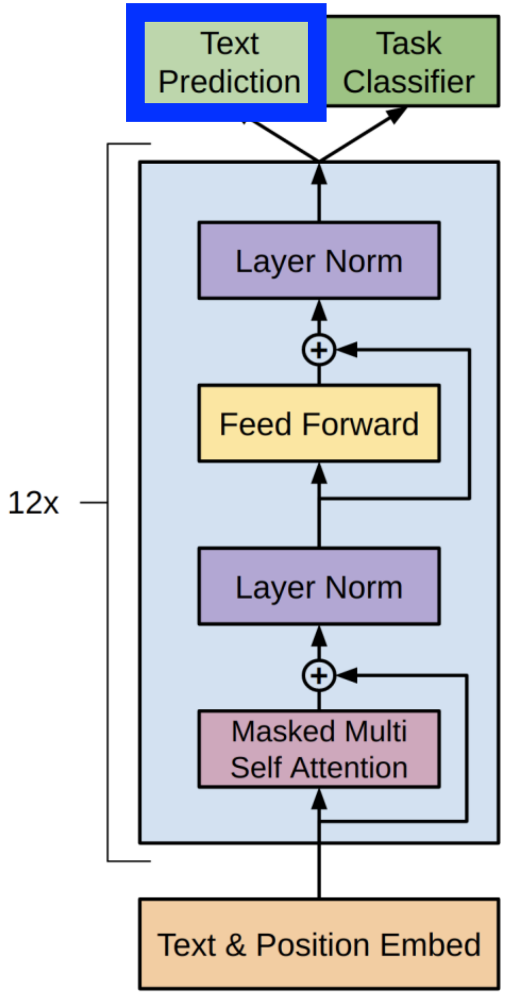


Supervised training to decide if two sentences are similar (given no inherent ordering of the two sentences, both orders are used)

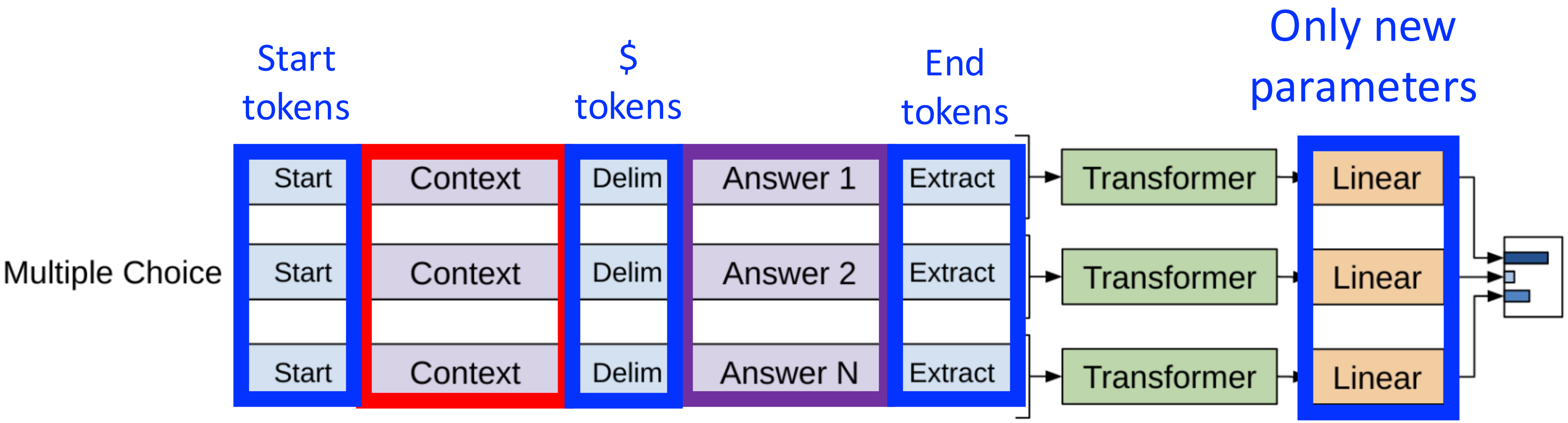
Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Fine-Tuning to Target Tasks (Softmax Output)

Auxiliary function improves performance



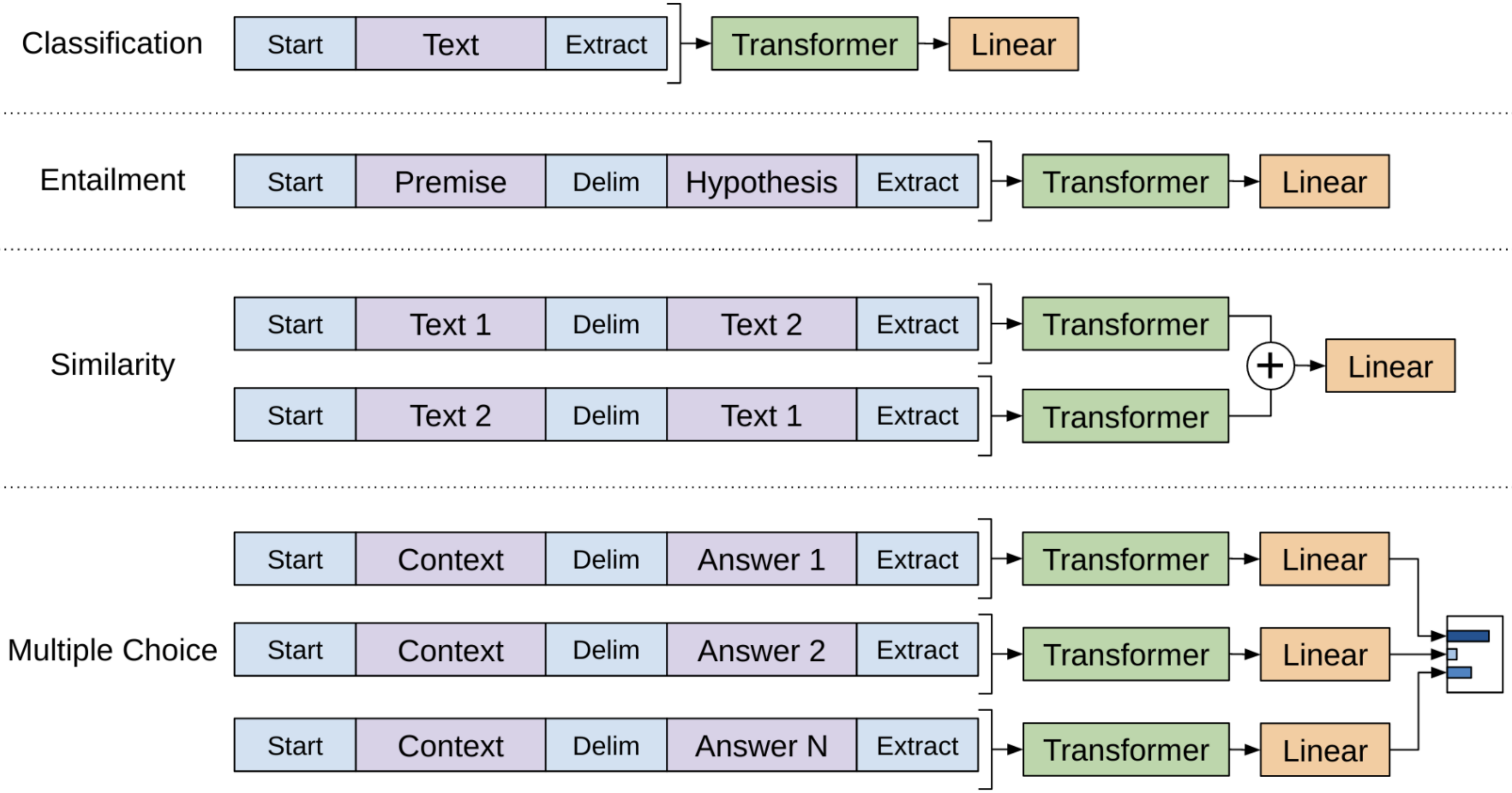
Supervised training to decide which answer results from the context (e.g., question about a document, concatenated); softmax layer generates probabilities for all options



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Fine-Tuning to Target Tasks (Softmax Output): Discriminative (Instead of Generative)

- Task-specific datasets
- Mini-batch size: usually 32
- Regularization: dropout
- Optimizer: Adam
- Training duration: usually 3 epochs
- All parameters fine-tuned



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

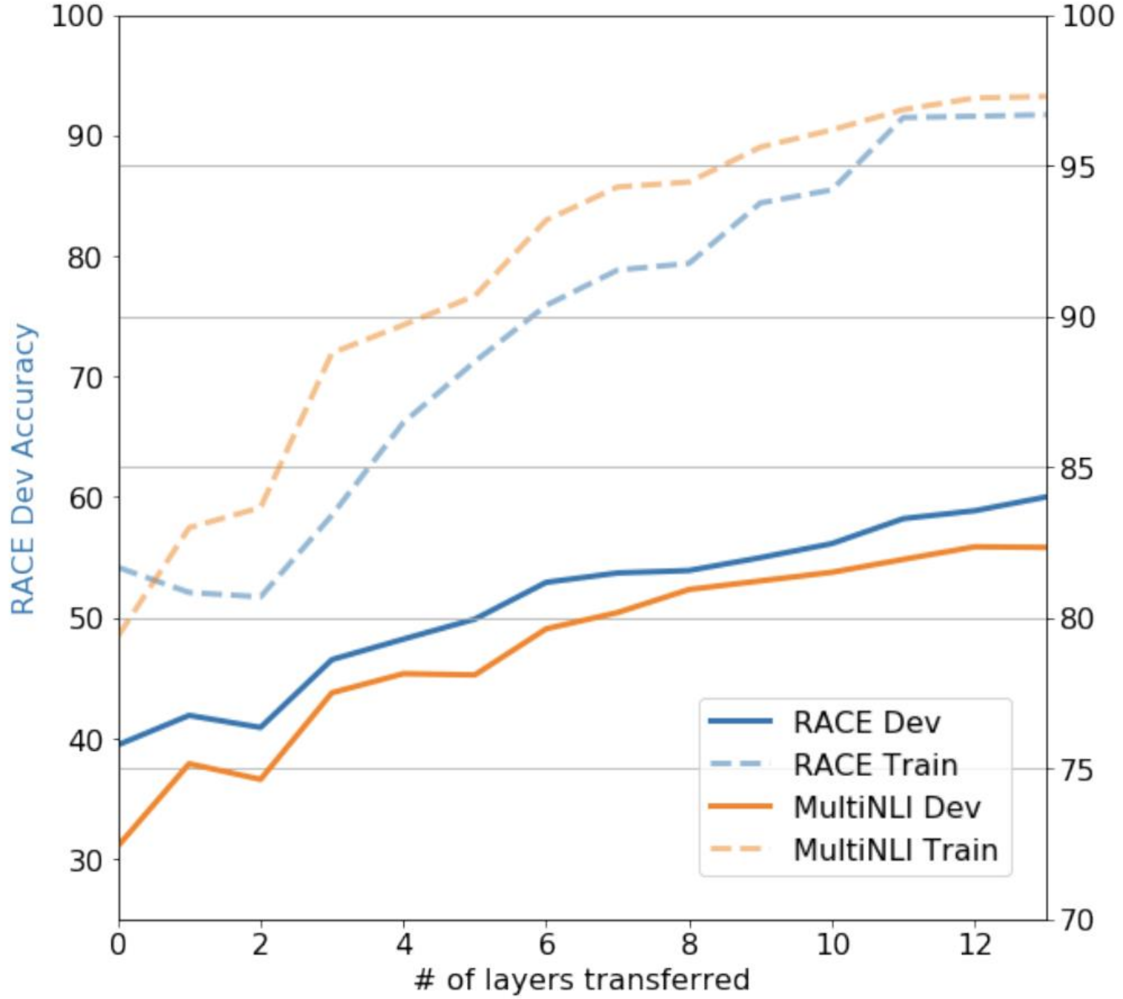
Experimental Findings

Achieved state-of-the-art performance on
9 of 12 tested NLP dataset challenges

Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Experimental Findings: Importance of Design Decisions?

- How many pretrained layers should be used for fine-tuning?
  - All



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Experimental Findings: Importance of Design Decisions?

- How many pretrained layers should be used for fine-tuning? - All
- Does pre-training help?
- Does the auxiliary language modeling task for fine-tuning help?
- Does using a transformer rather than LSTM help?

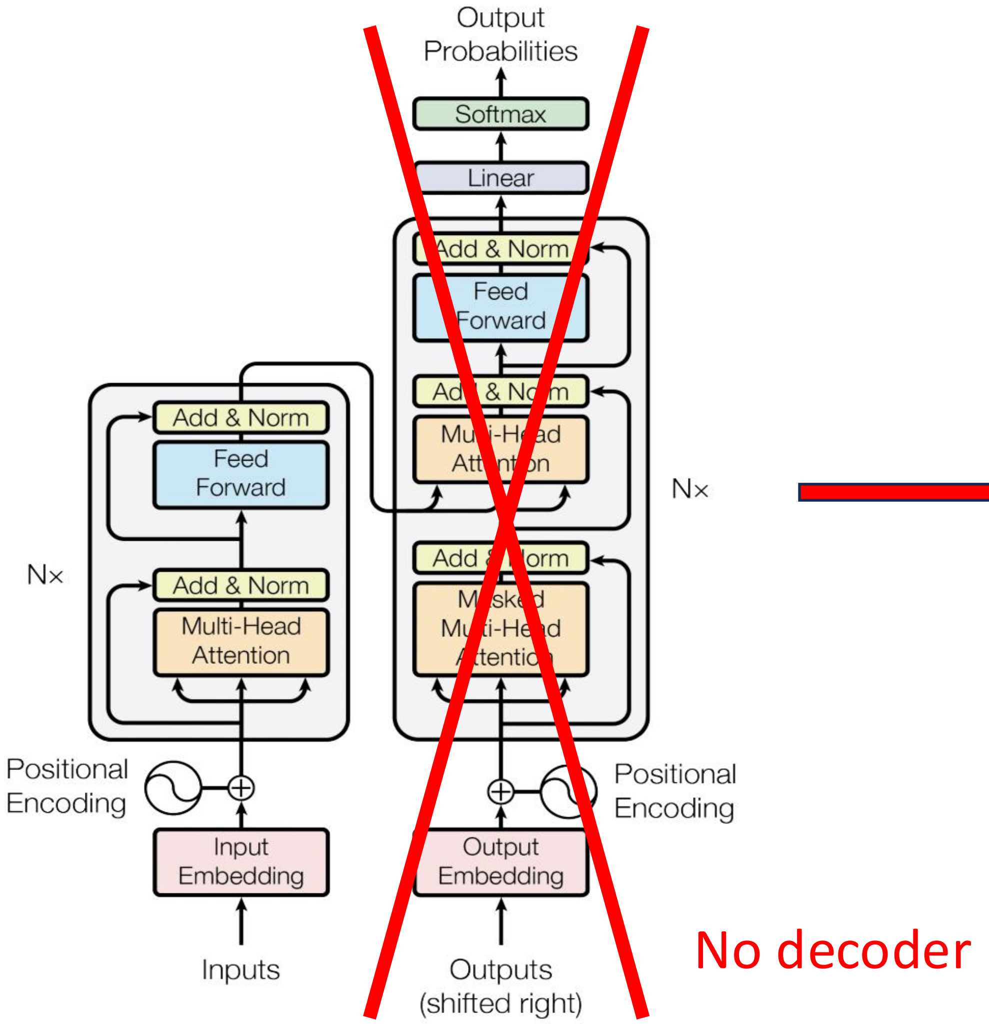
Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	MNLI (acc)	QNLI (acc)	RTE (acc)
Transformer w/ aux LM (full)	74.7	45.4	91.3	82.3	82.0	**70.3**	**81.8**	**88.1**	**56.0**
Transformer w/o pre-training	59.9	18.9	84.0	79.4	30.9	65.5	75.7	71.2	53.8
Transformer w/o aux LM	**75.0**	**47.9**	**92.0**	**84.9**	**83.2**	69.8	81.1	86.9	54.4
LSTM w/ aux LM	69.1	30.3	90.5	83.2	71.8	68.1	73.7	81.1	54.6

Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

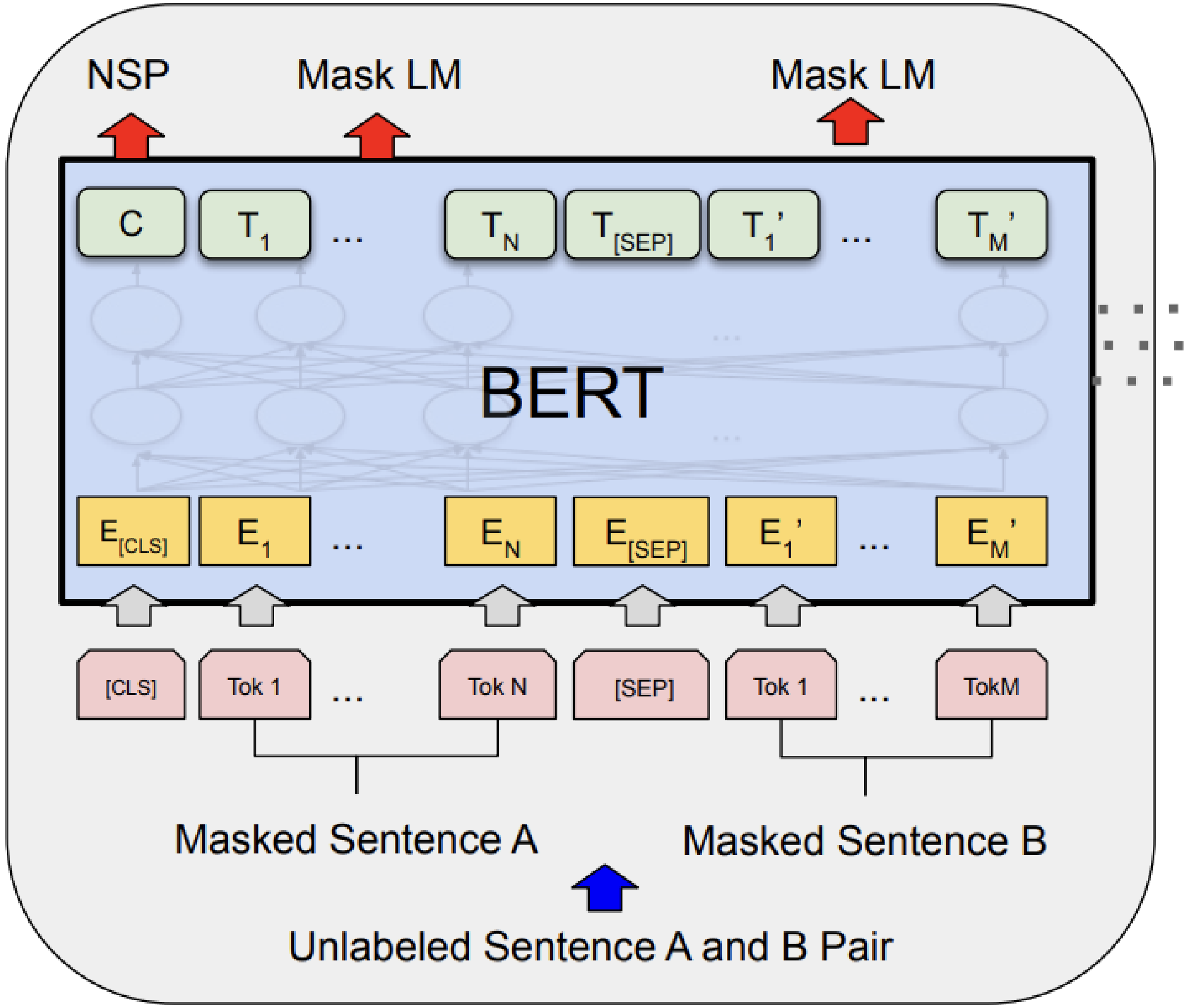
Today’s Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Programming tutorial

Architecture: Encoder of Pioneering Transformer



Vaswani et al. Attention Is All You Need. Neurips 2017.

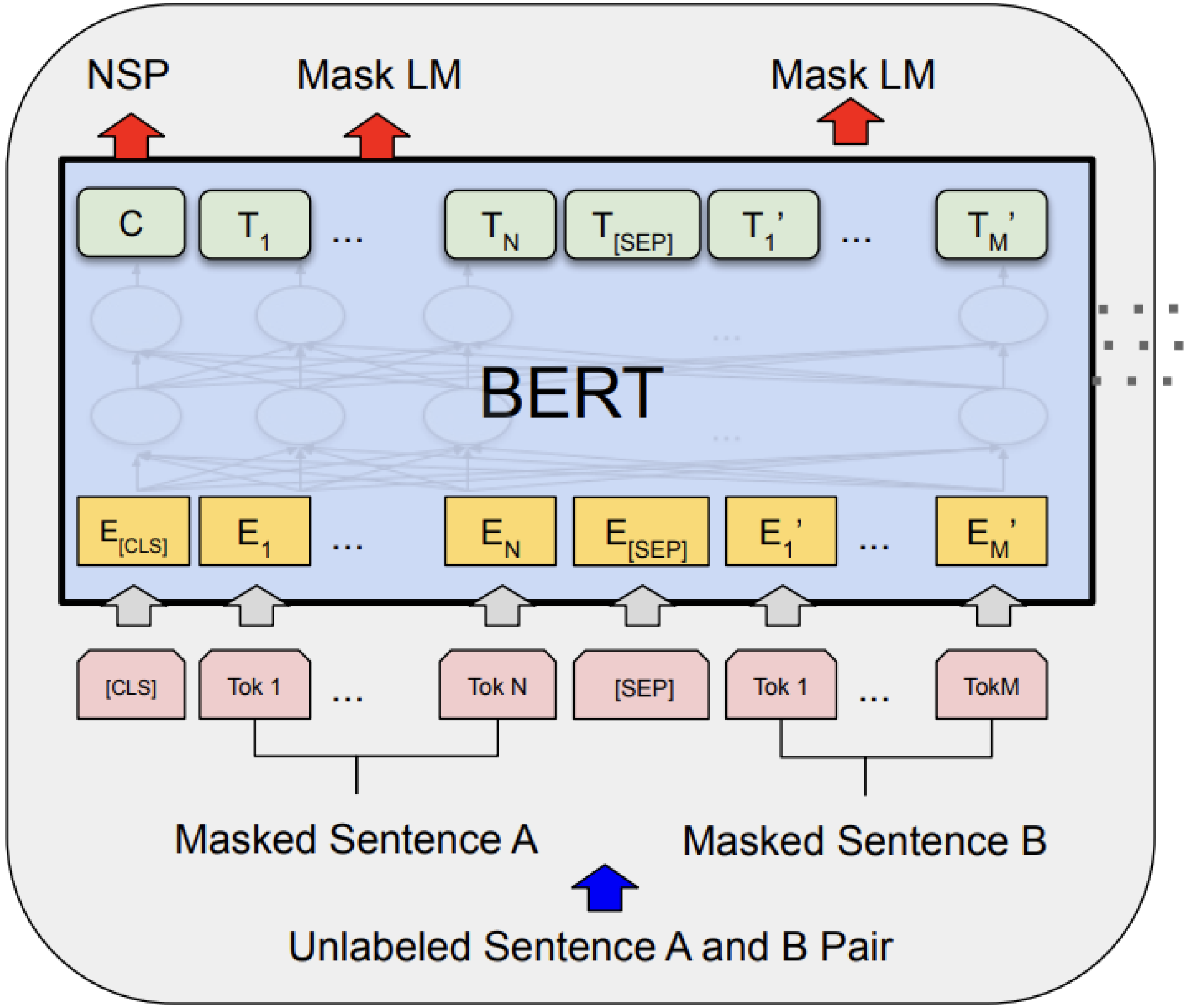


Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Architecture: Minor Tweaks

Like GPT, makes these changes:

- GELU activation functions in feedforward layers
- Subword tokenization (with WordPiece)



Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Key Ideas

- Support diverse inputs: special tokens and embeddings
- Learn good representations for downstream tasks: pretraining objective functions
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

Key Ideas

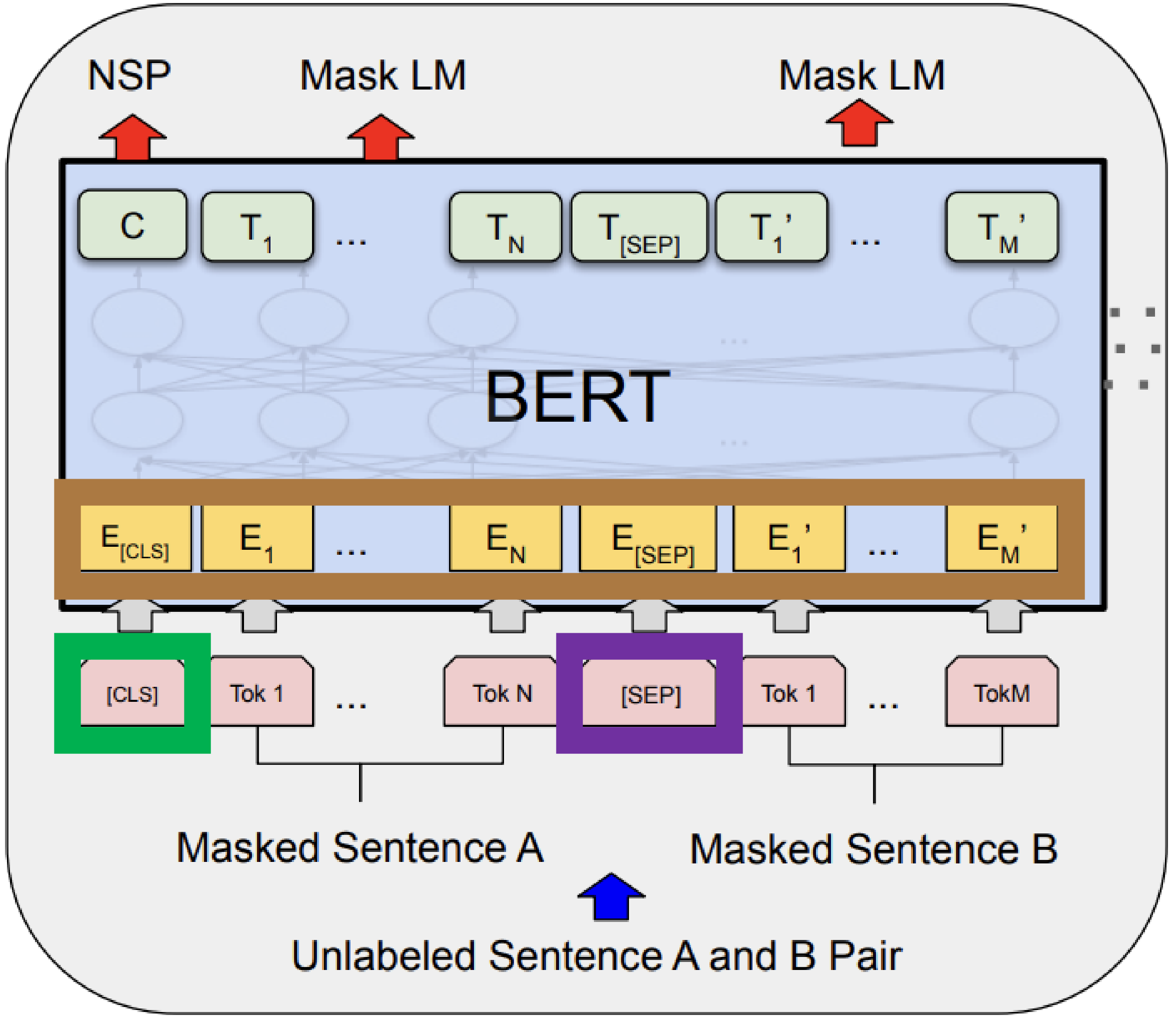
- Support diverse inputs: special tokens and embeddings
- Learn good representations for downstream tasks: pretraining objective functions
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

Architecture: Input Tokens

Augment 30,000 tokens from tokenizer with:

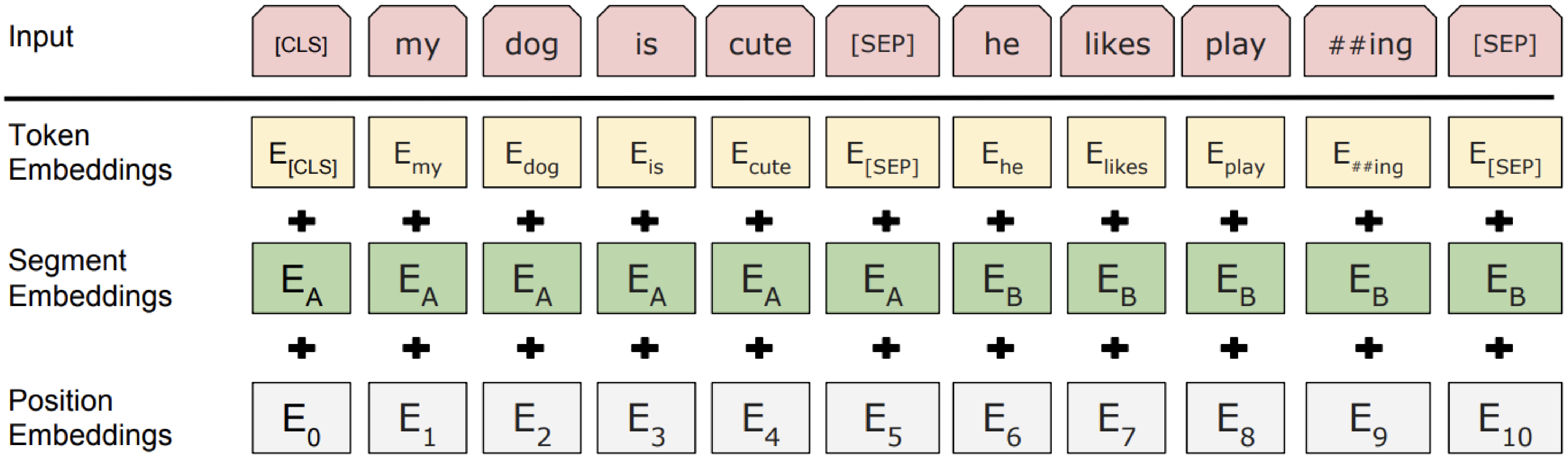
- CLS: representation of **ENTIRE** input *(like GPT fine-tuned end token)*
- SEP: separates two input sequences *(like GPT fine-tuned delimiter token)*

Token embeddings created from the embedding matrix learned during training



Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Architecture: Input to Transformer



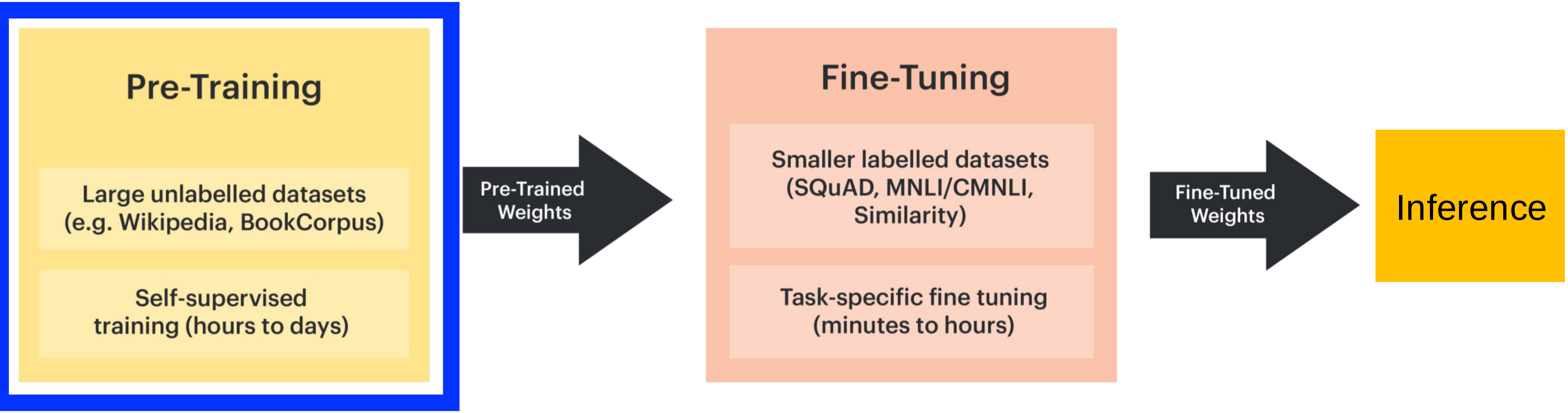
As before, embedding matrices are learned during training; the novel segment embedding signifies to which of two sentences a token belongs

Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Key Ideas

- Support diverse inputs: special tokens and embeddings
- Learn good representations for downstream tasks: pretraining objective functions
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

BERT: Bidirectional Encoder Representation from Transformer



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Idea: Choose a Pretraining Task That Is **Not** Unidirectional

GPT's prediction of the next word given previous ones is unidirectional (left-to-right)

1. Background music from a ________
2. Many people danced around the ________
3. I practiced for many years to learn how to play the ________

Two Tasks

1. Predict masked token (key contribution)

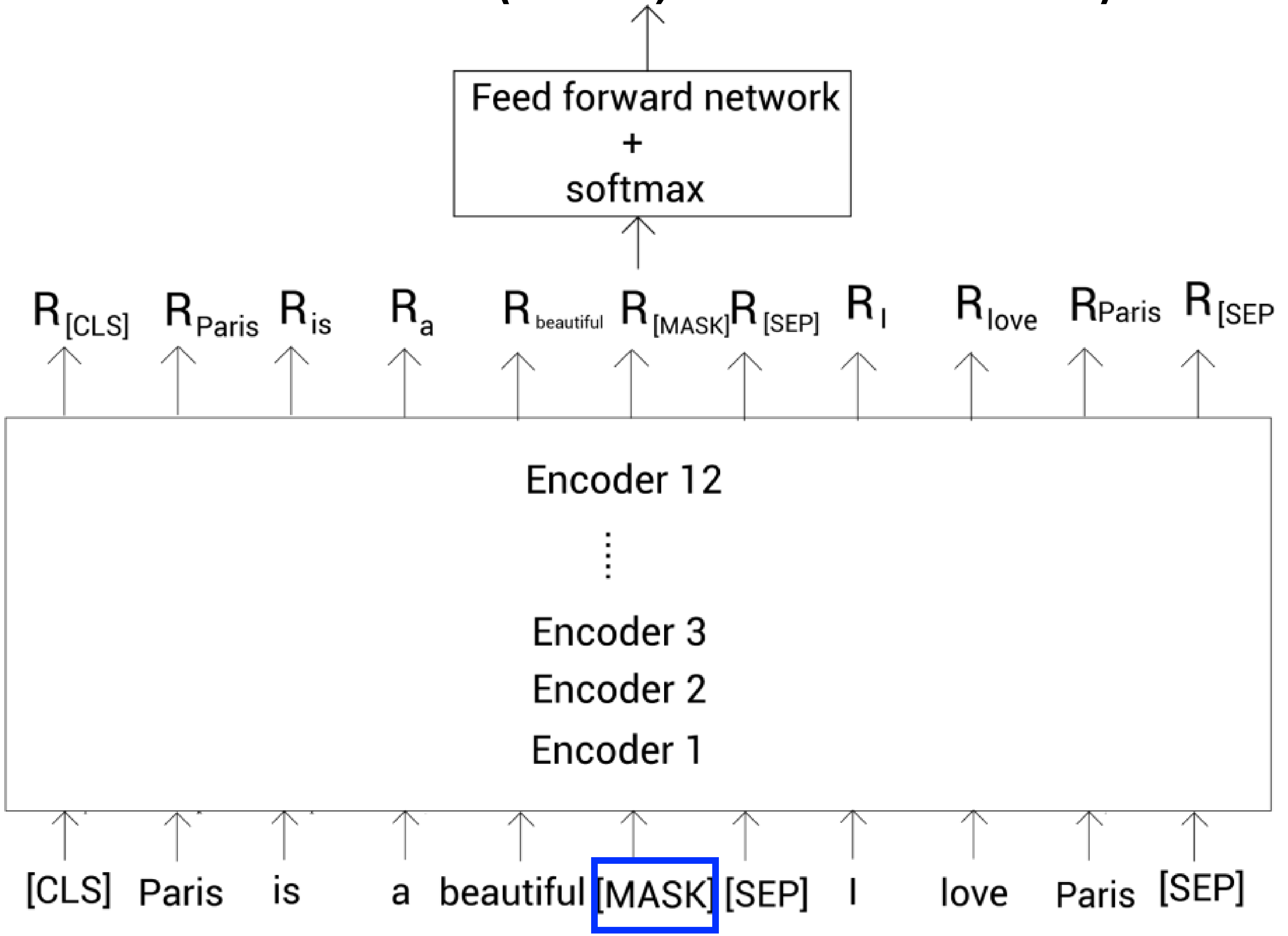
2. Predict if one sentence follows a second sentence
   (augments understanding of how sentences relate)

Task 1: Predict Masked Token (aka, Cloze Task)

Modifies 15% of random tokens in each input:

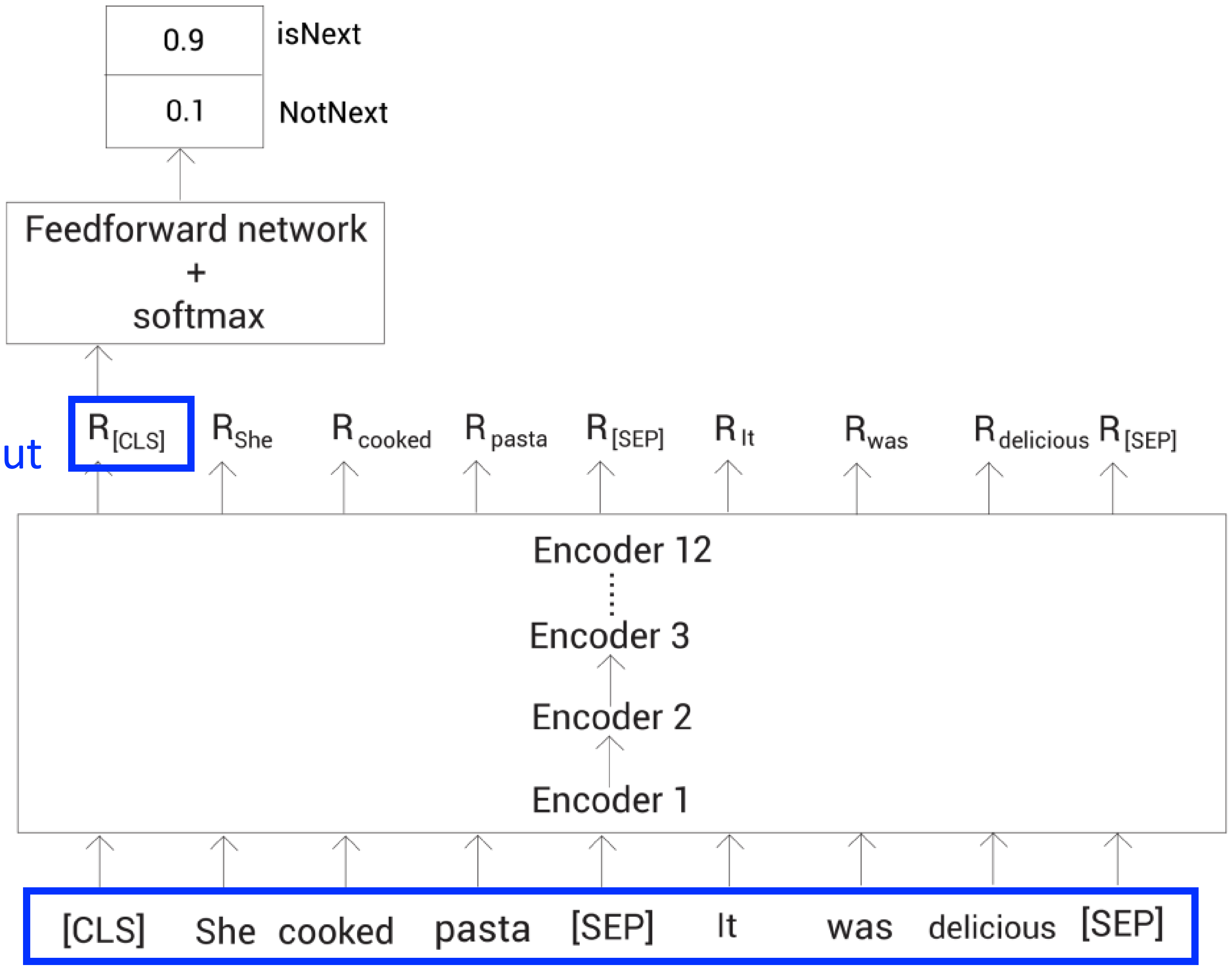
- for 80%, uses mask token
- for 10%, uses random token
- for 10%, uses original token

(Latter 20% mitigate learning which are masked tokens, forcing model to understand EVERY input token)



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Task 2: Predict if Next Sentence Task



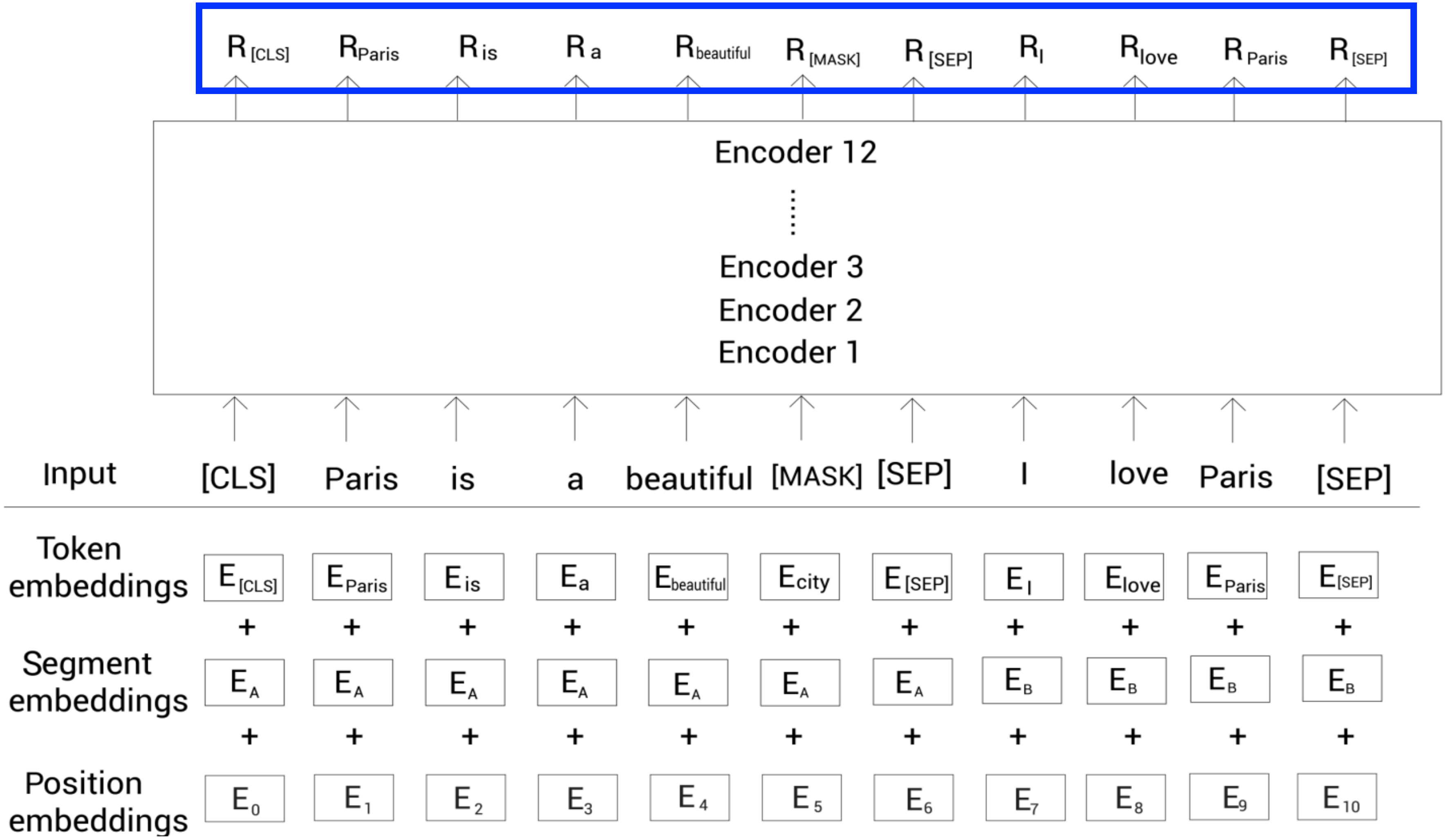
Predicts using the token representing ENTIRE input

50% of 2nd sentences are original next sentence and rest are random

https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Bidirectional Encoder Representation

Representations of ENTIRE context



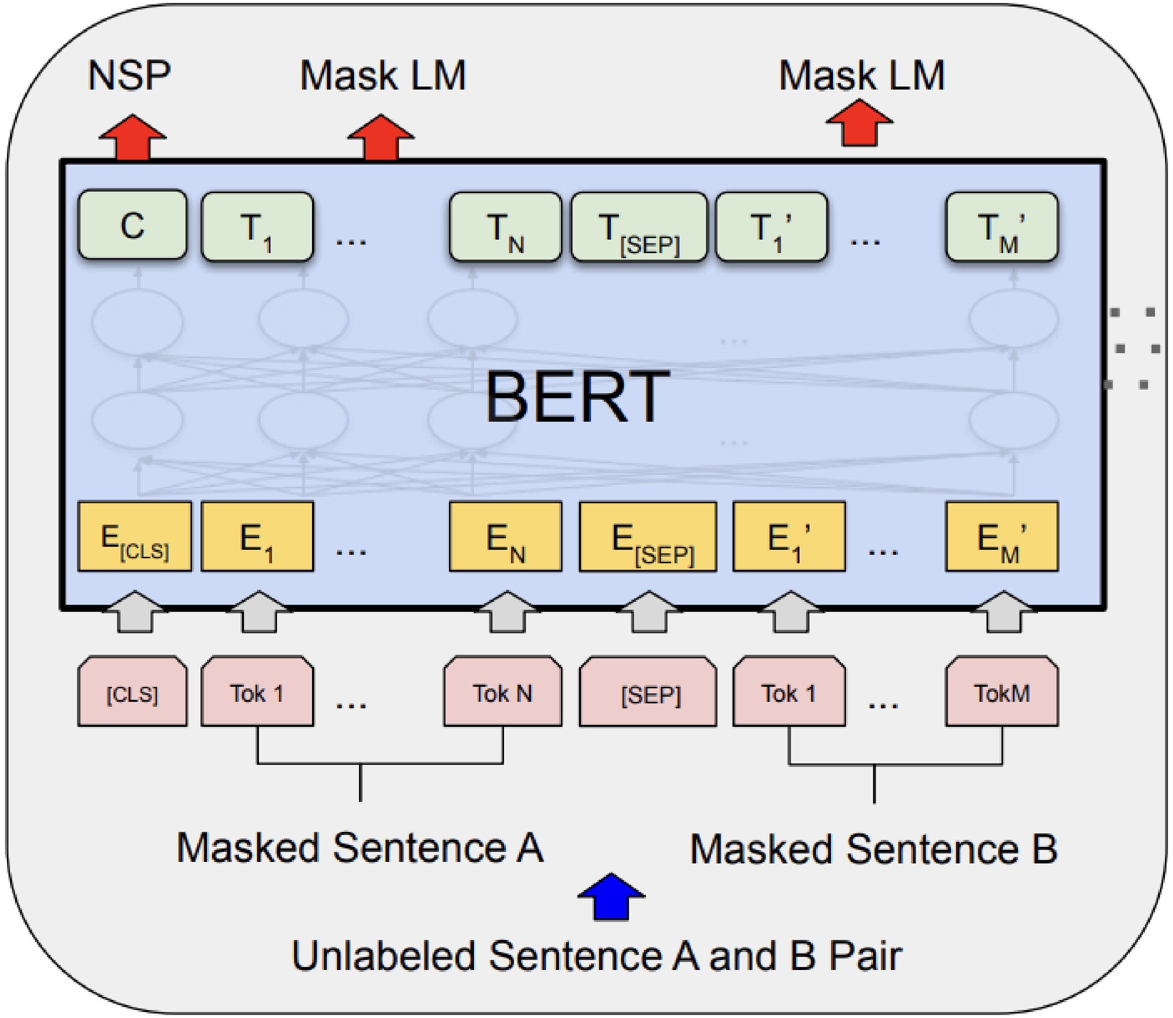
https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Key Ideas

- Support diverse inputs: special tokens and embeddings
- Learn good representations for downstream tasks: pretraining objective functions
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

Training

- Dataset: 800M words in BooksCorpus (>7,000 unpublished books) and 2.5B words in English Wikipedia
- Training duration: ~40 epochs, spanning 4 days on 64 TPUs)
- Mini-batch size: 256 sequences (two “sentences”) of 128 tokens for 90% of epochs and then 512 tokens
- Regularization: dropout & L2 norm penalty
- Optimizer: Adam

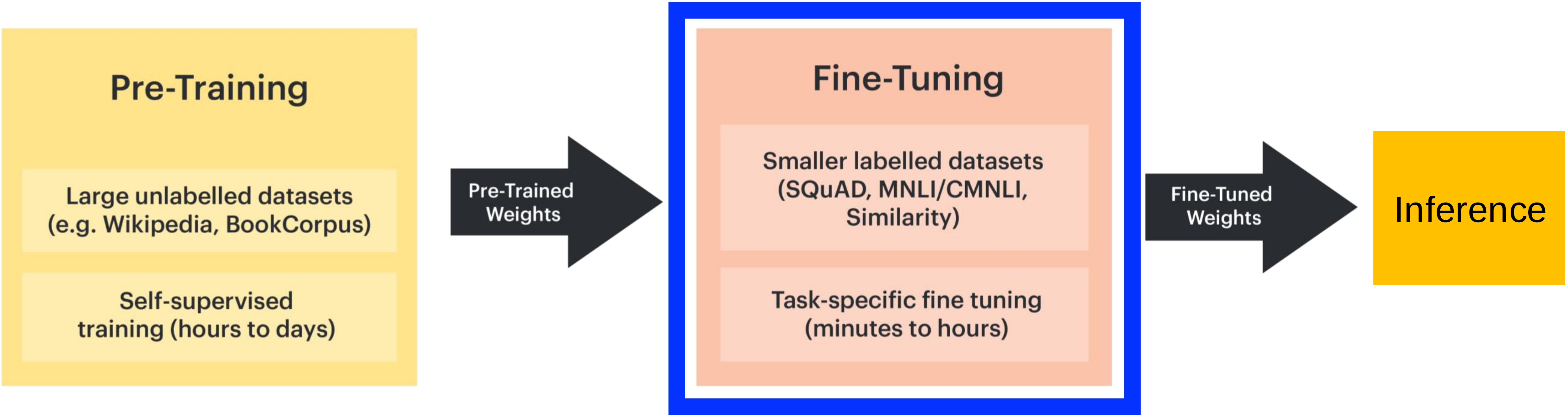


Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Key Ideas

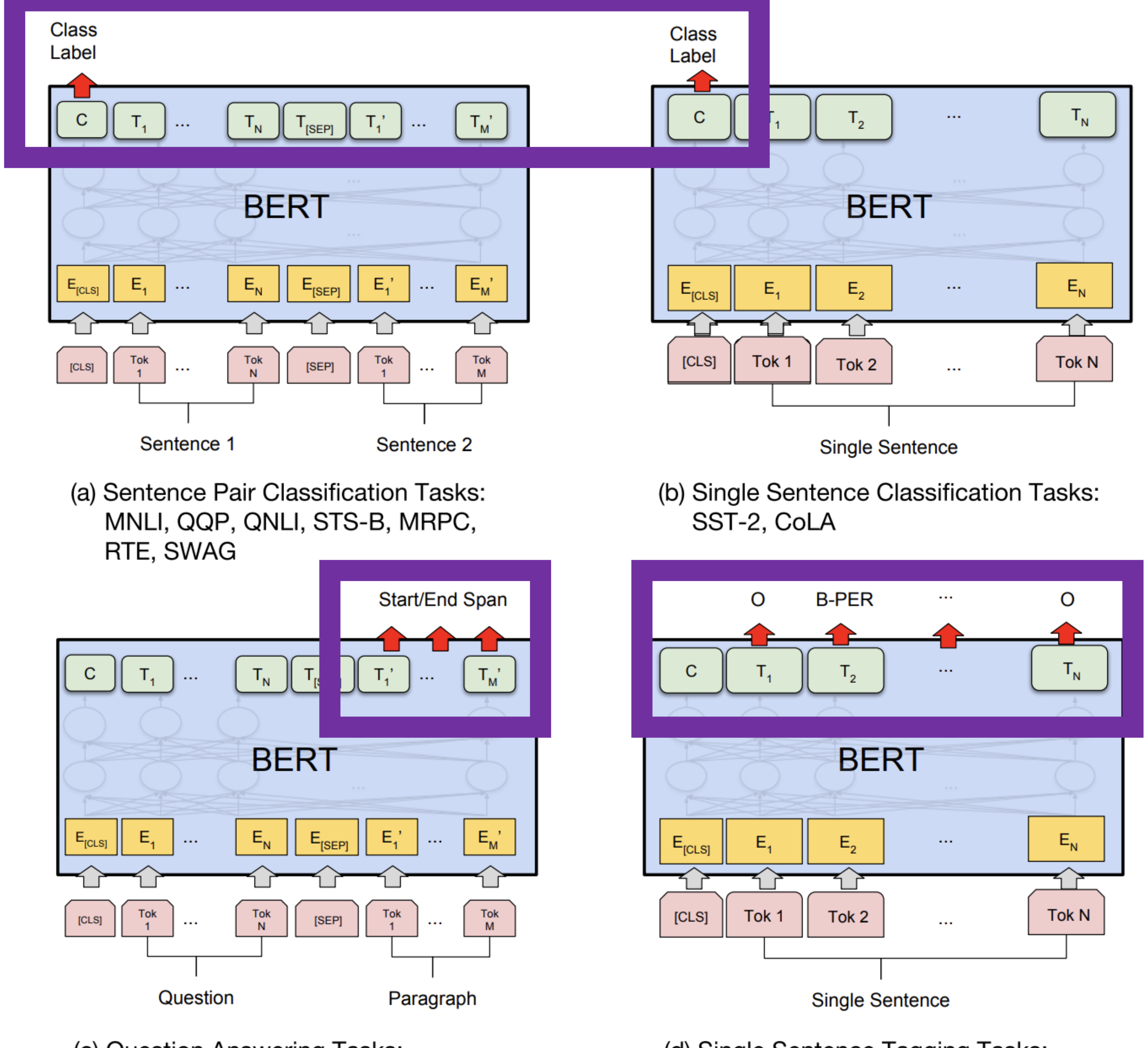
- Support diverse inputs: special tokens and embeddings
- Learn good representations for downstream tasks: pretraining objective functions
- Make learning feasible: self-supervised pre-training
- Fine-tune pretrained models for downstream tasks with little architectural change

BERT: Bidirectional Encoder Representation from Transformer



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Fine-Tuning to Target Tasks: 3 Types



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

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

(c) Question Answering Tasks: SQuAD v1.1

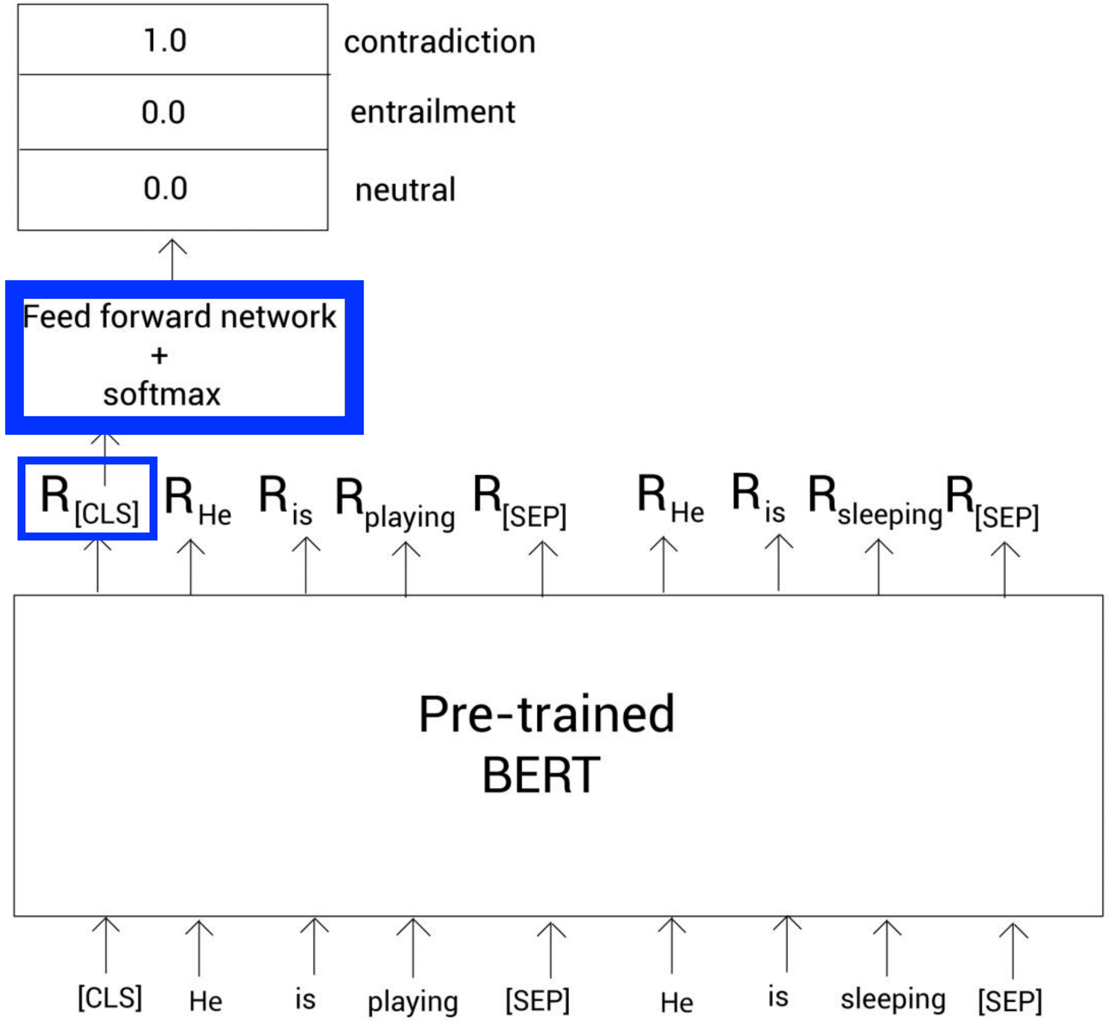
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Fine-Tuning for Classification; e.g., Natural Language Inference

Only new parameters

Predicts using the token representing ENTIRE input

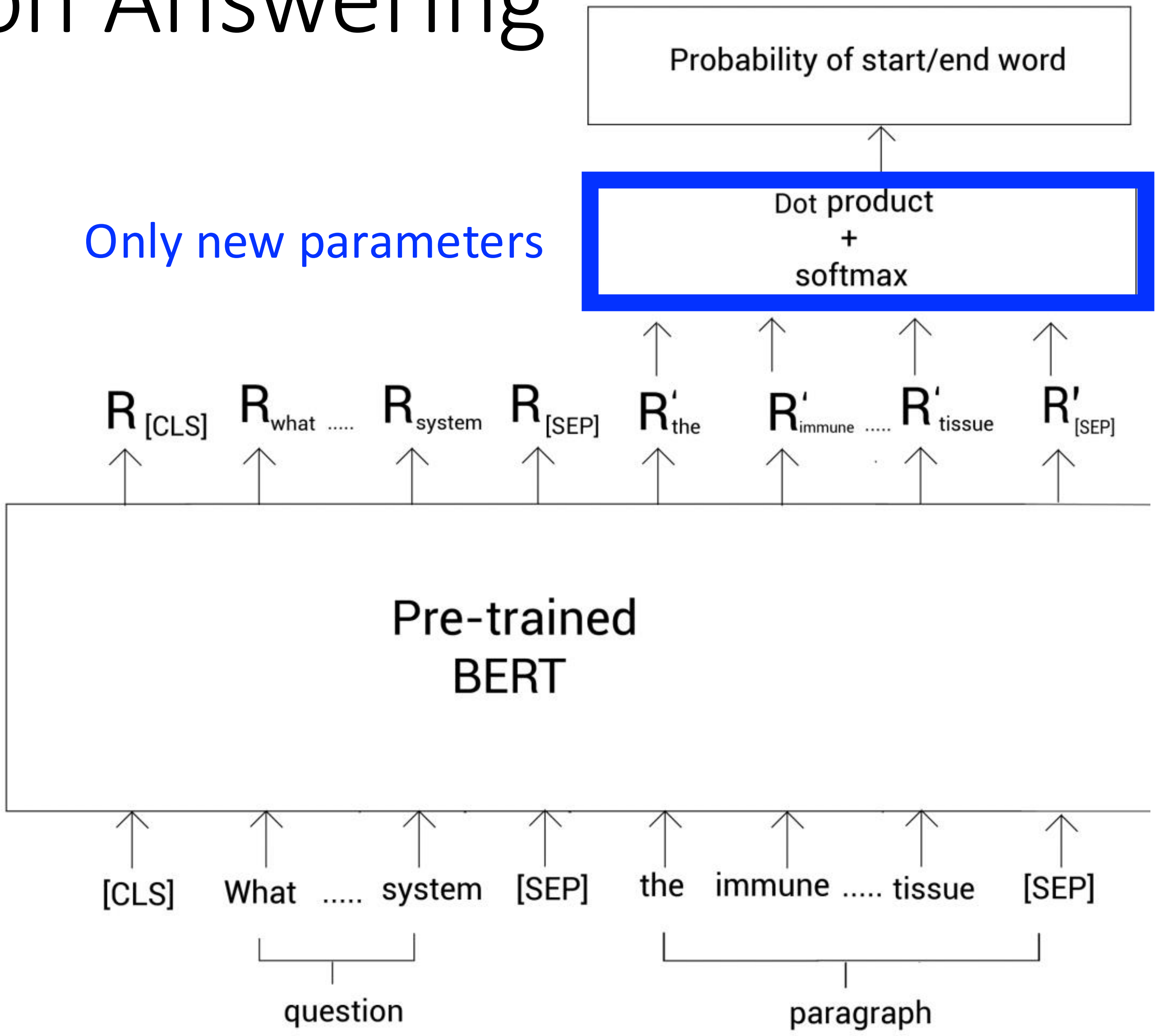


https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Fine-Tuning for Question Answering

Only new parameters

Predicts indices of start and end words in the input paragraph
- dot product of two learned vector representations with each input token

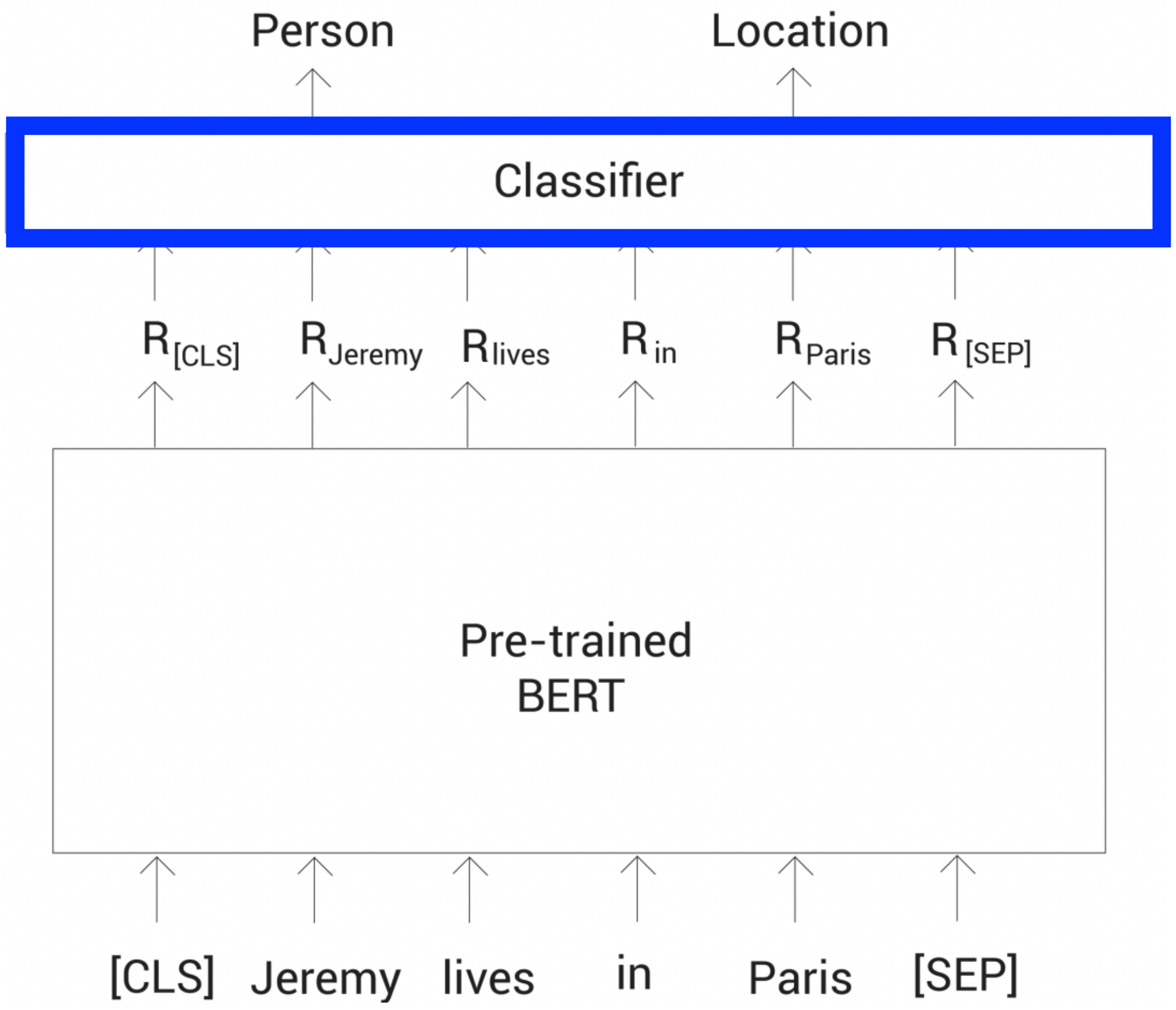


https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Fine-Tuning for Single Sentence Tagging

Only new parameters

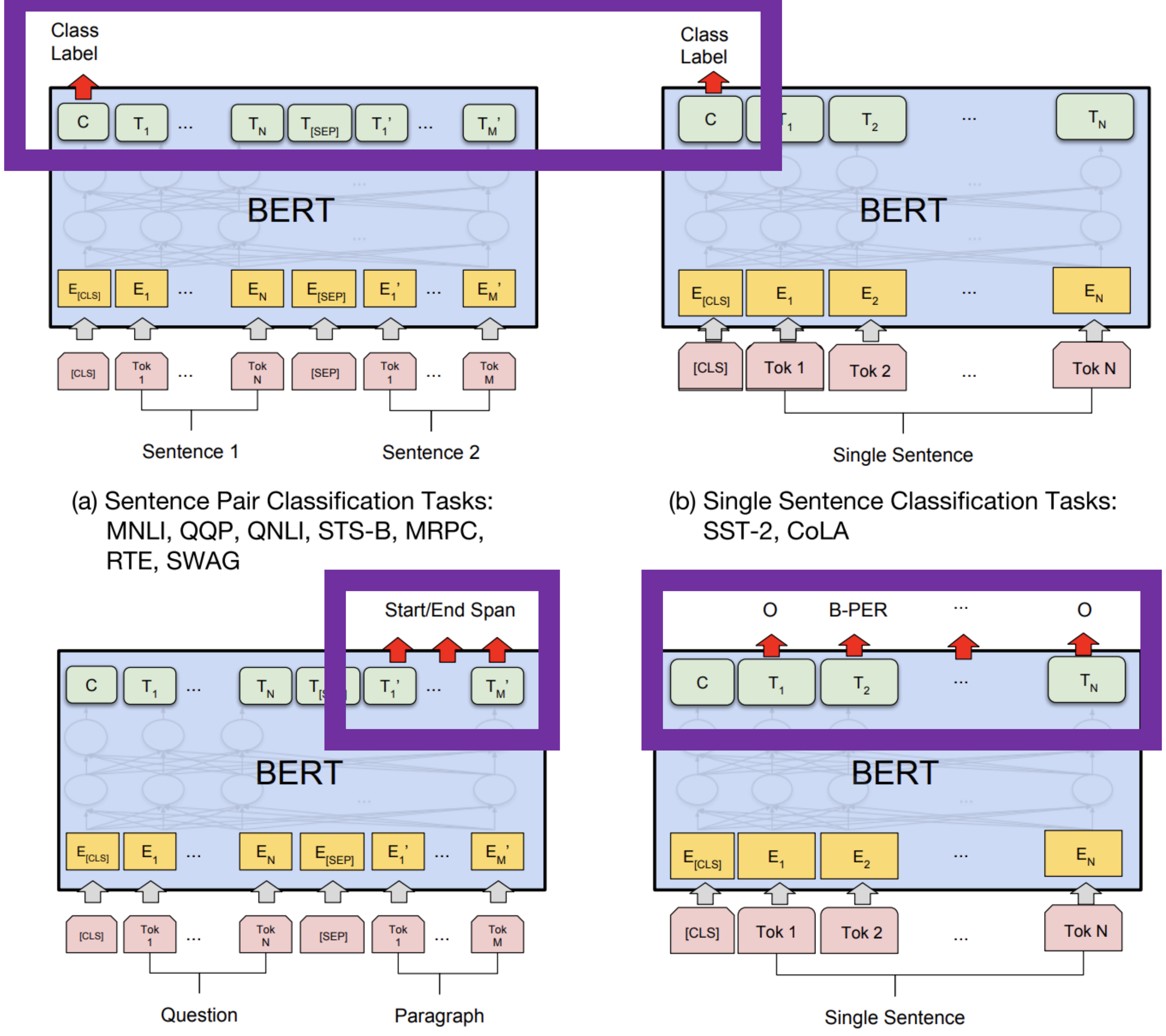
Each token's new representation is passed to classifier (e.g., named entity recognition)



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Fine-Tuning to Target Tasks: 3 Types

- Task-specific datasets
- Mini-batch size: usually 16 or 32
- Regularization: dropout
- Optimizer: Adam
- Training duration: ~3 epochs (i.e., ~1 hour on 1 TPU)
- All parameters fine-tuned



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

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

(c) Question Answering Tasks: SQuAD v1.1

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

Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Experimental Findings

Achieved state-of-the-art
performance on 11 tested NLP tasks

Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Experimental Findings: Importance of Design Decisions?

- Does including next sentence prediction in pretraining help?
  - Yes
- Does including masked token prediction in pretraining help? *(uses GPT's left-to-right approach instead)*
  - Yes; worse results from unidirectional than bidirectional pretraining

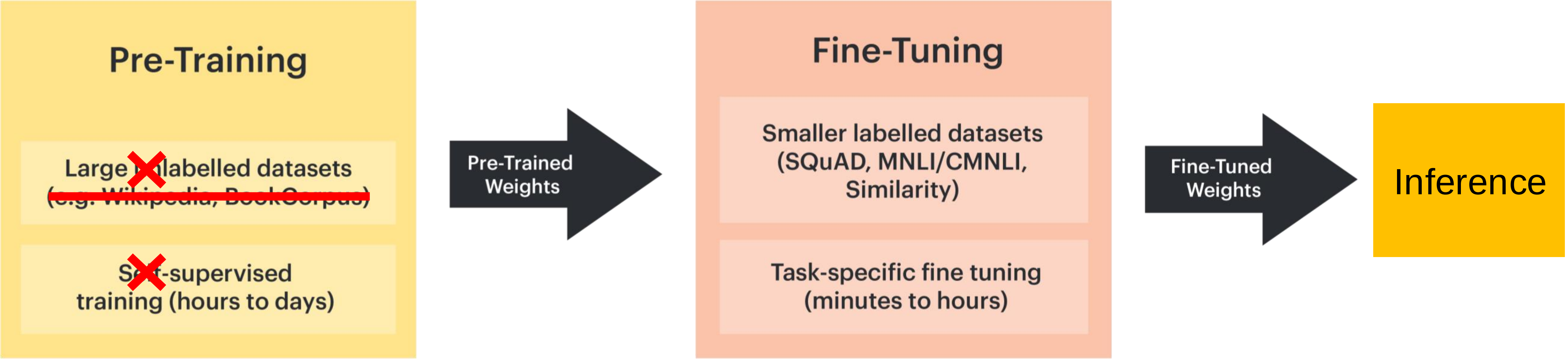
Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
	Dev Set				
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8

Devlin et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ACL 2019.

Today's Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Programming tutorial

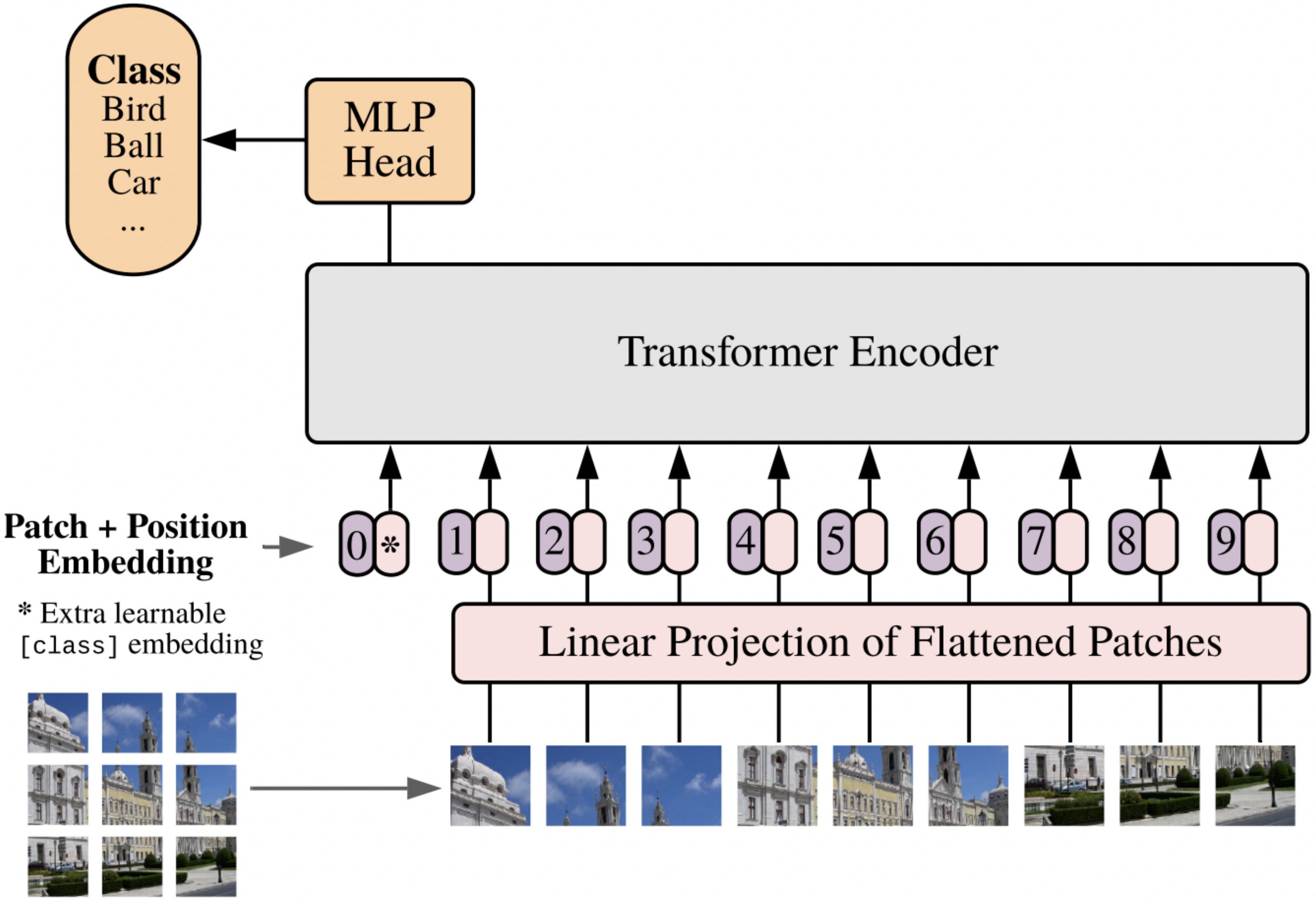
ViT: Vision Transformer



(self-supervised masked prediction
deemed inferior to supervised learning)

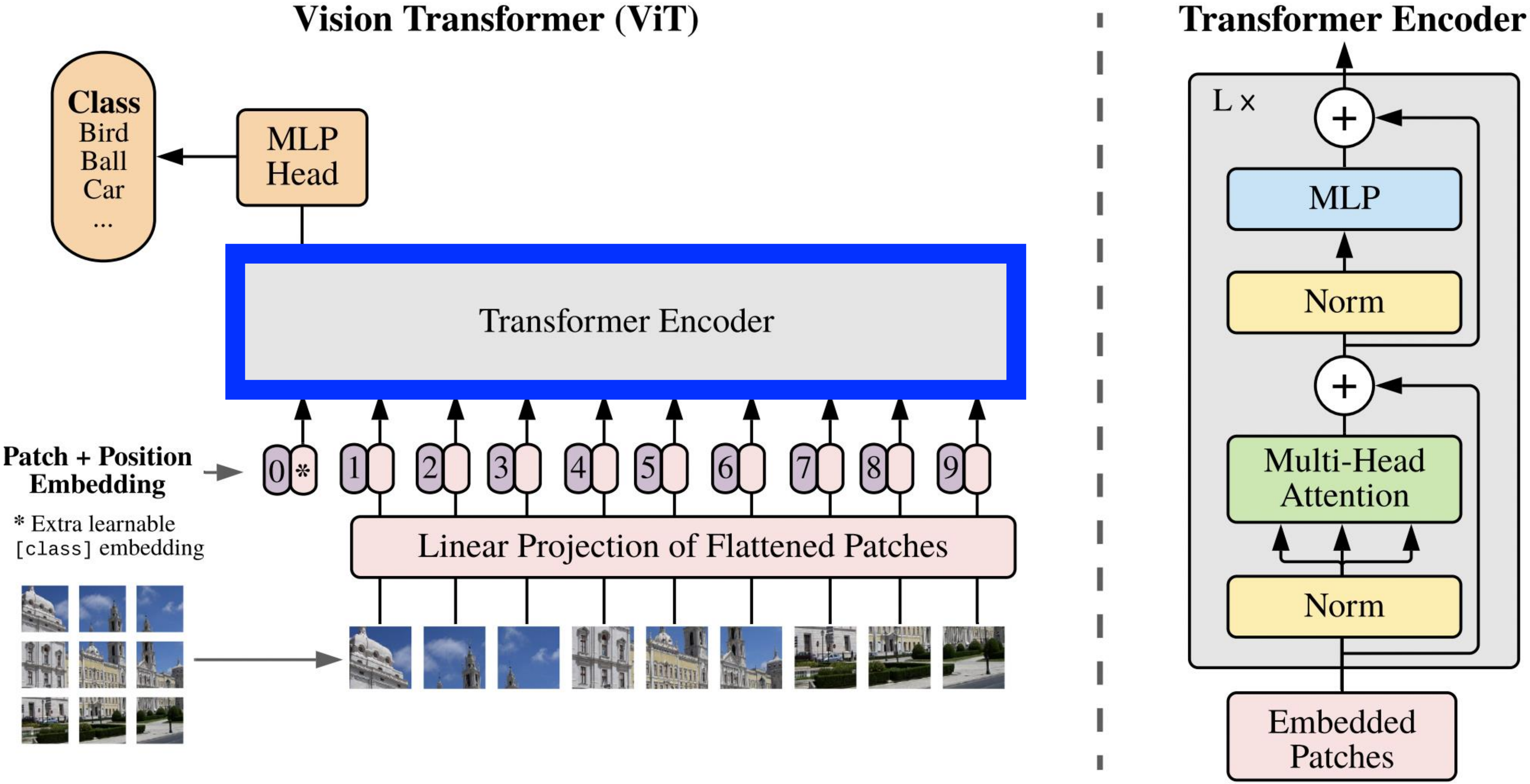
https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Architecture



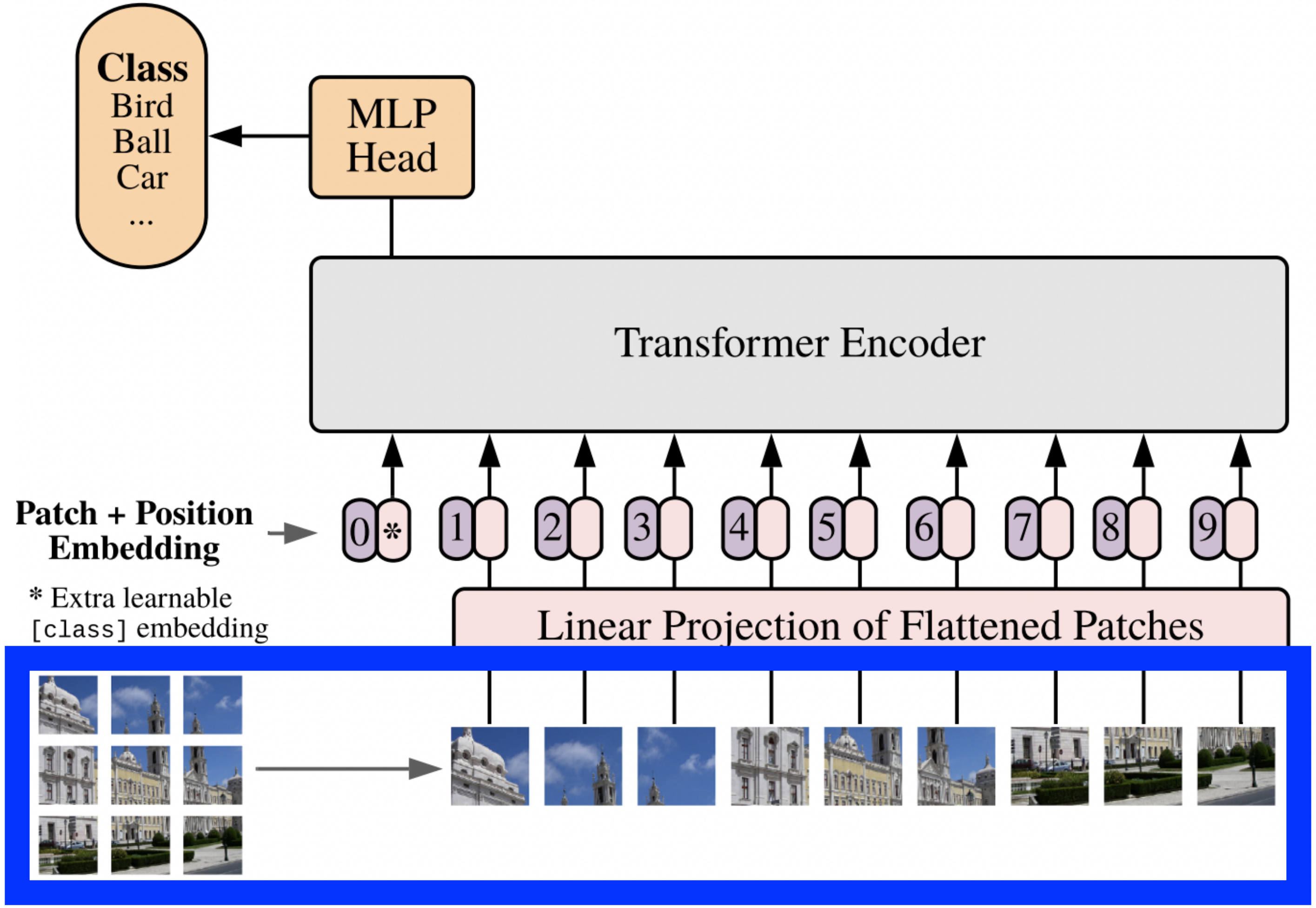
Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

Architecture: BERT



Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

Architecture: Input (Patches Instead of Pixels)

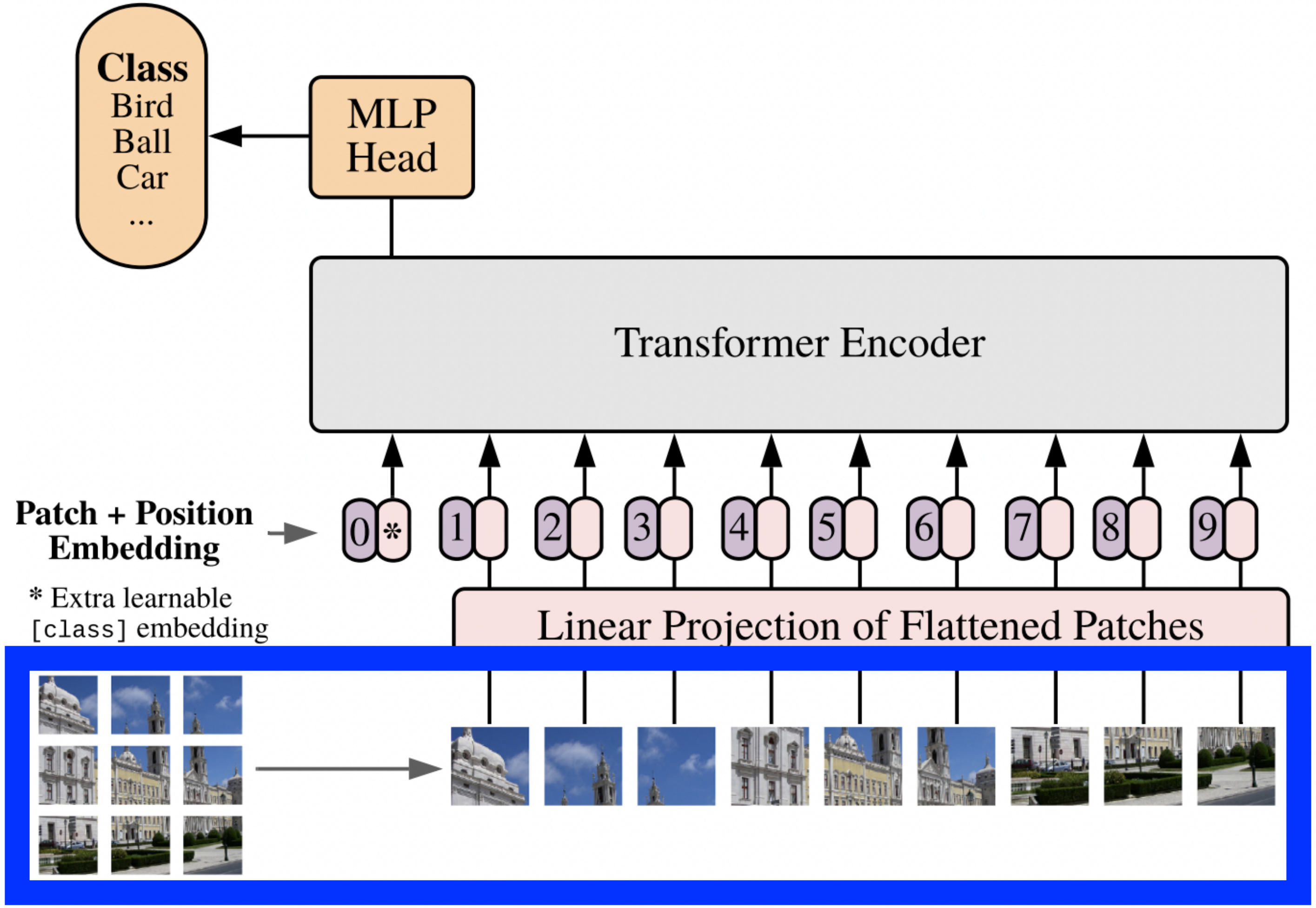


Assuming a 160 x 160 pixel image, how many patches (and so inputs) would be extracted?
- 100

image decomposed into 16x16 patches (example simplified); representations include "flattened" and ResNet features

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

Architecture: Input (Patches Instead of Pixels)

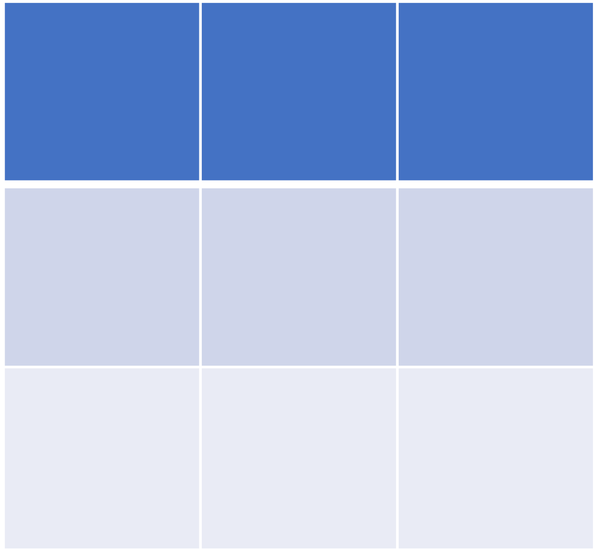


Why not use the raw pixels as input?

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

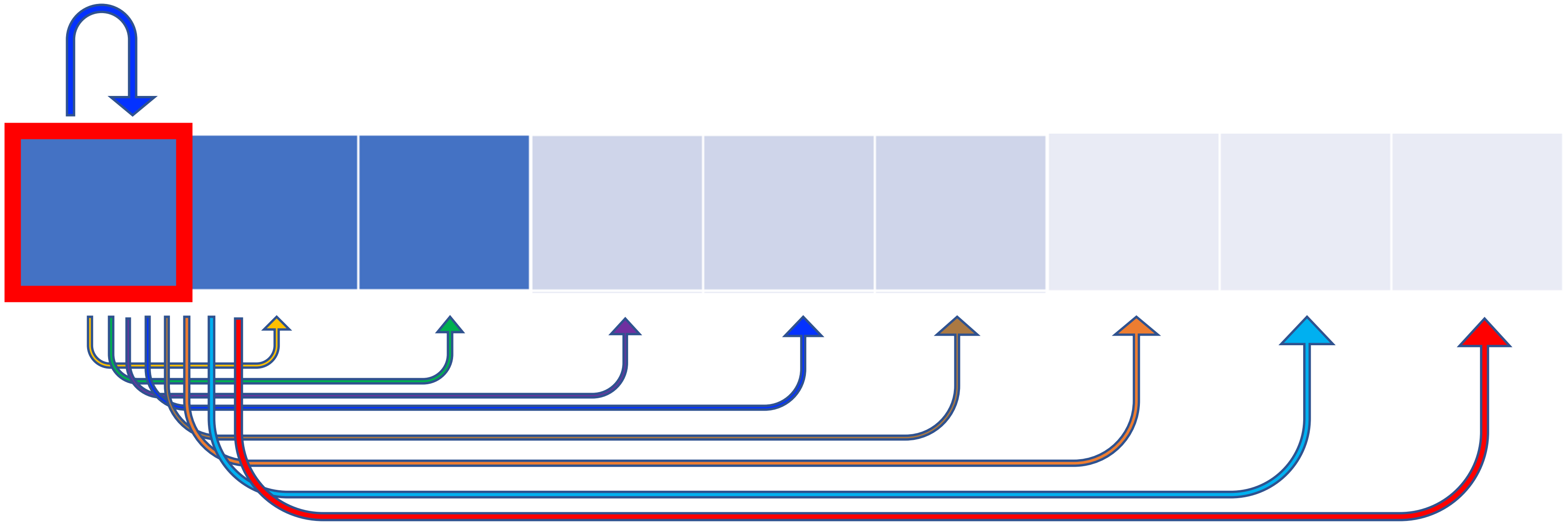
Recall Self-Attention: Idea

New representation of each pixel showing its relationship to all pixels; e.g., assume this 3x3 image



Recall Self-Attention: Idea

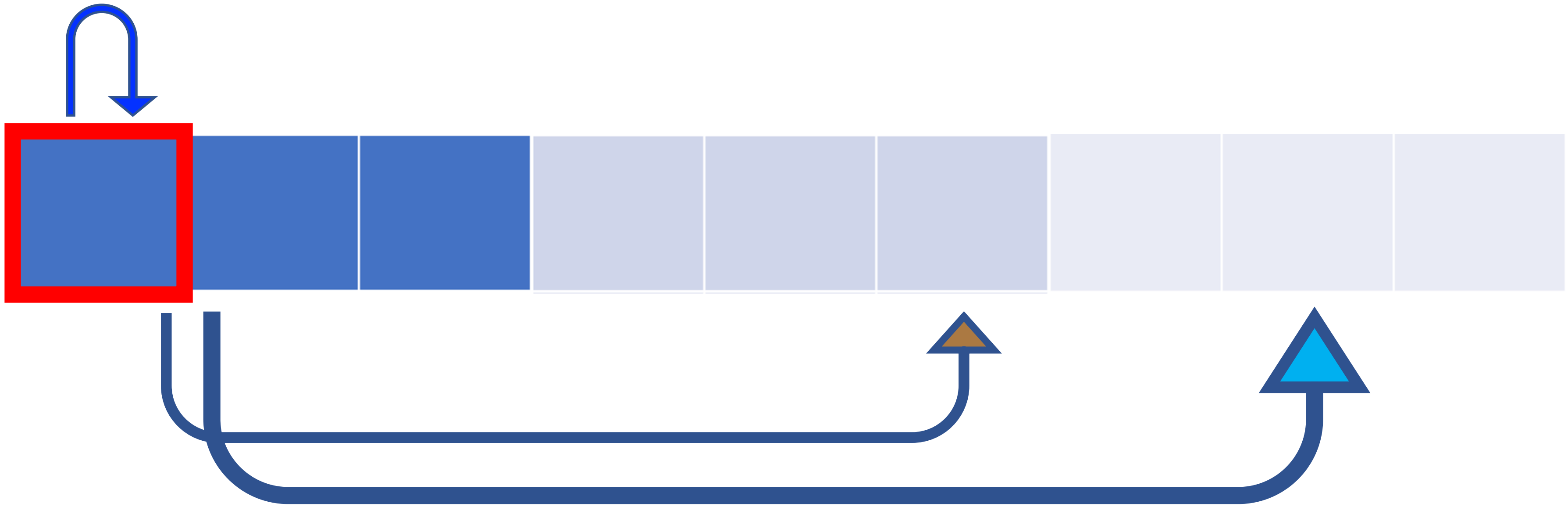
New representation of each pixel showing its relationship to all pixels; e.g., assume this 3x3 image



Learned new representation indicates which global information clarifies a pixel's meaning (e.g., include in the representation of a pixel of an eye context of what animal it belongs to)

Recall Self-Attention: Idea

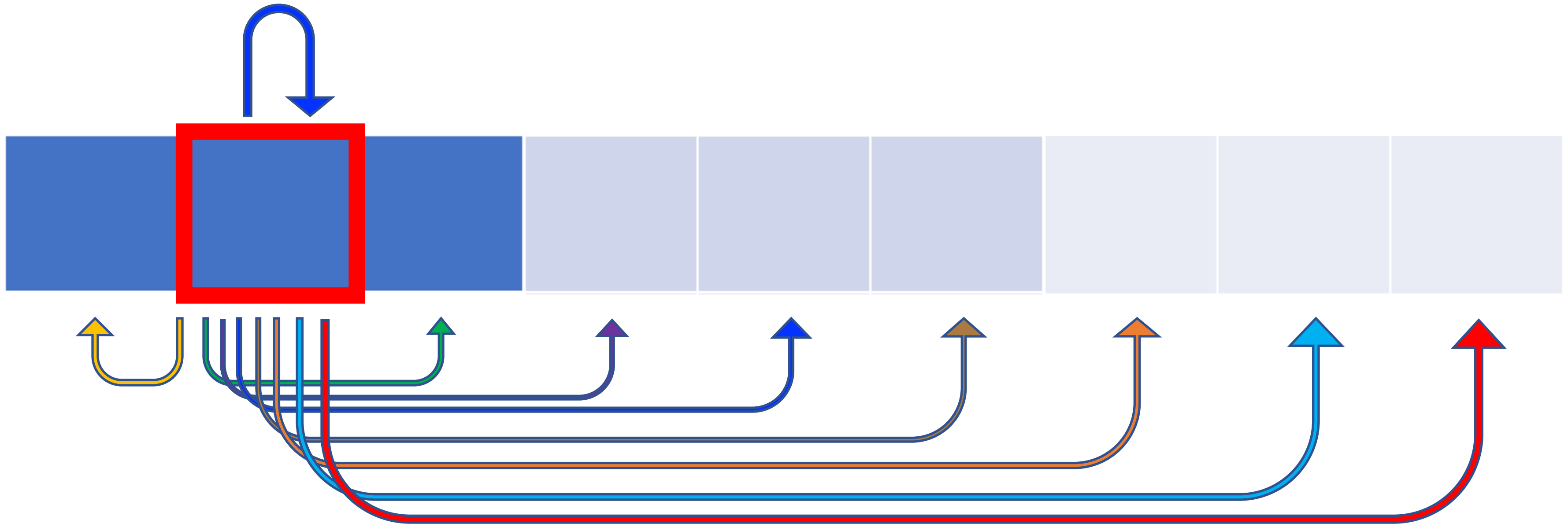
New representation of each pixel showing its relationship to all pixels; e.g., assume this 3x3 image



Learned new representation indicates which global information clarifies a pixel’s meaning (e.g., include for representation of an eye pixel context of what animal it belongs to)

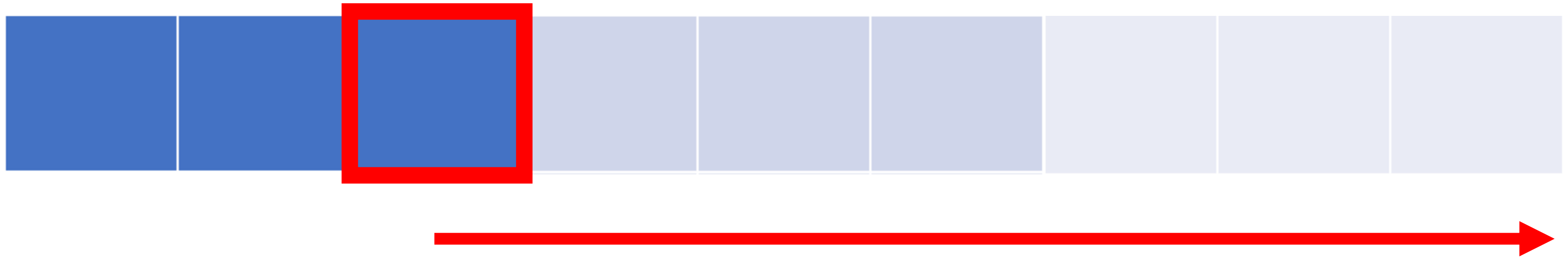
Recall Self-Attention: Idea

New representation of each pixel showing its relationship to all pixels; e.g., assume this 3x3 image



Recall Self-Attention: Idea

New representation of each pixel showing its relationship to all pixels; e.g., assume this 3x3 image

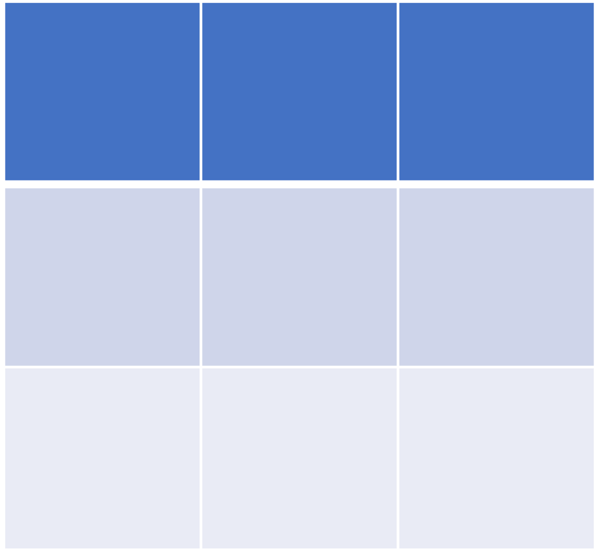


And so on for remaining image pixels…

Rationale for Patches: Computational Cost of Self-Attention

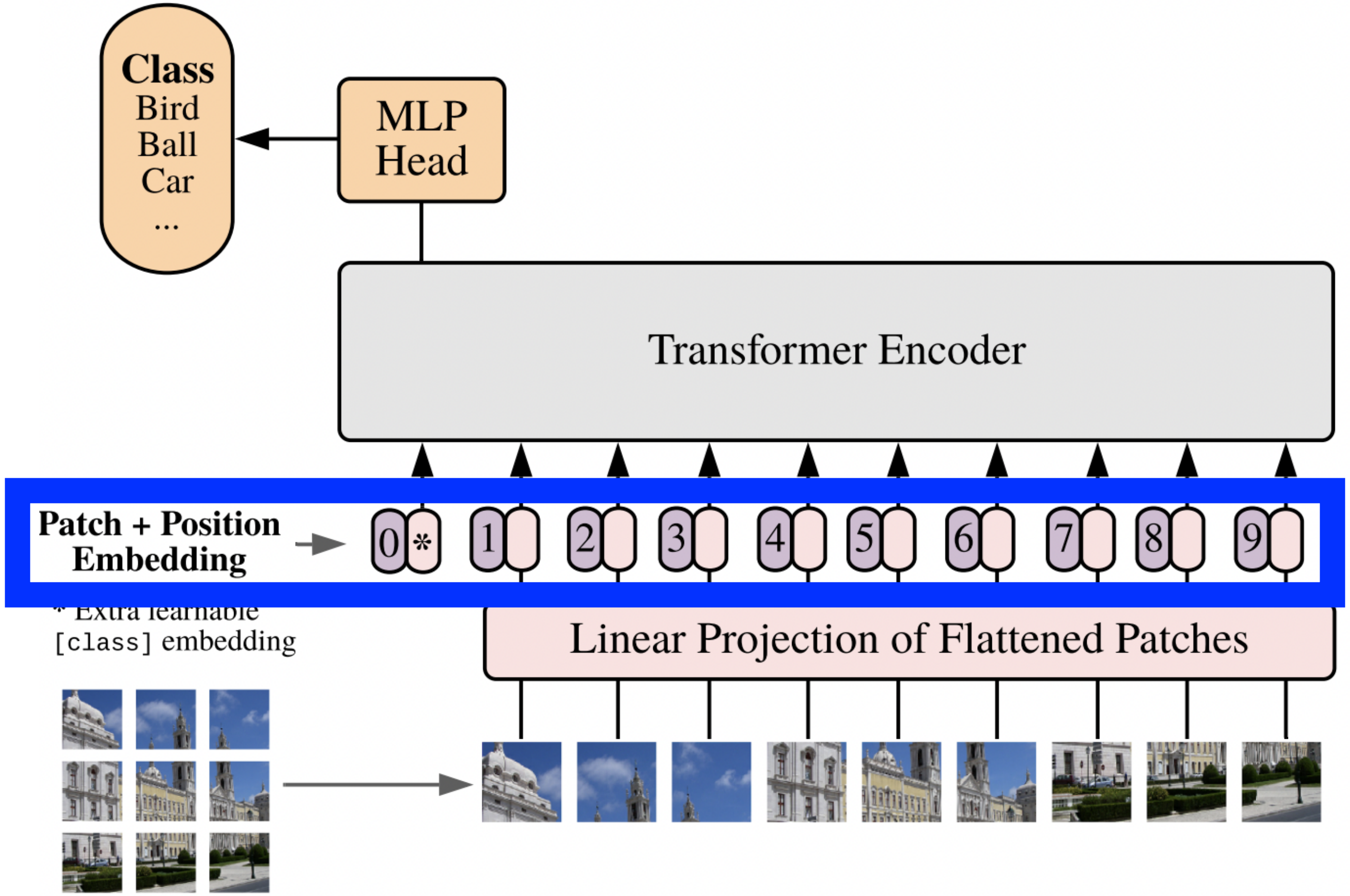
e.g., instead of using 3x3 image, what if a 1920 x 1080 image was used? How many self-attention computations would be needed?

- $(1920 \times 1080)^2$ = 4,299,816,960,000 (i.e., ~4.3 trillion)



Quadratic cost of self-attention is often impractical for pixels!

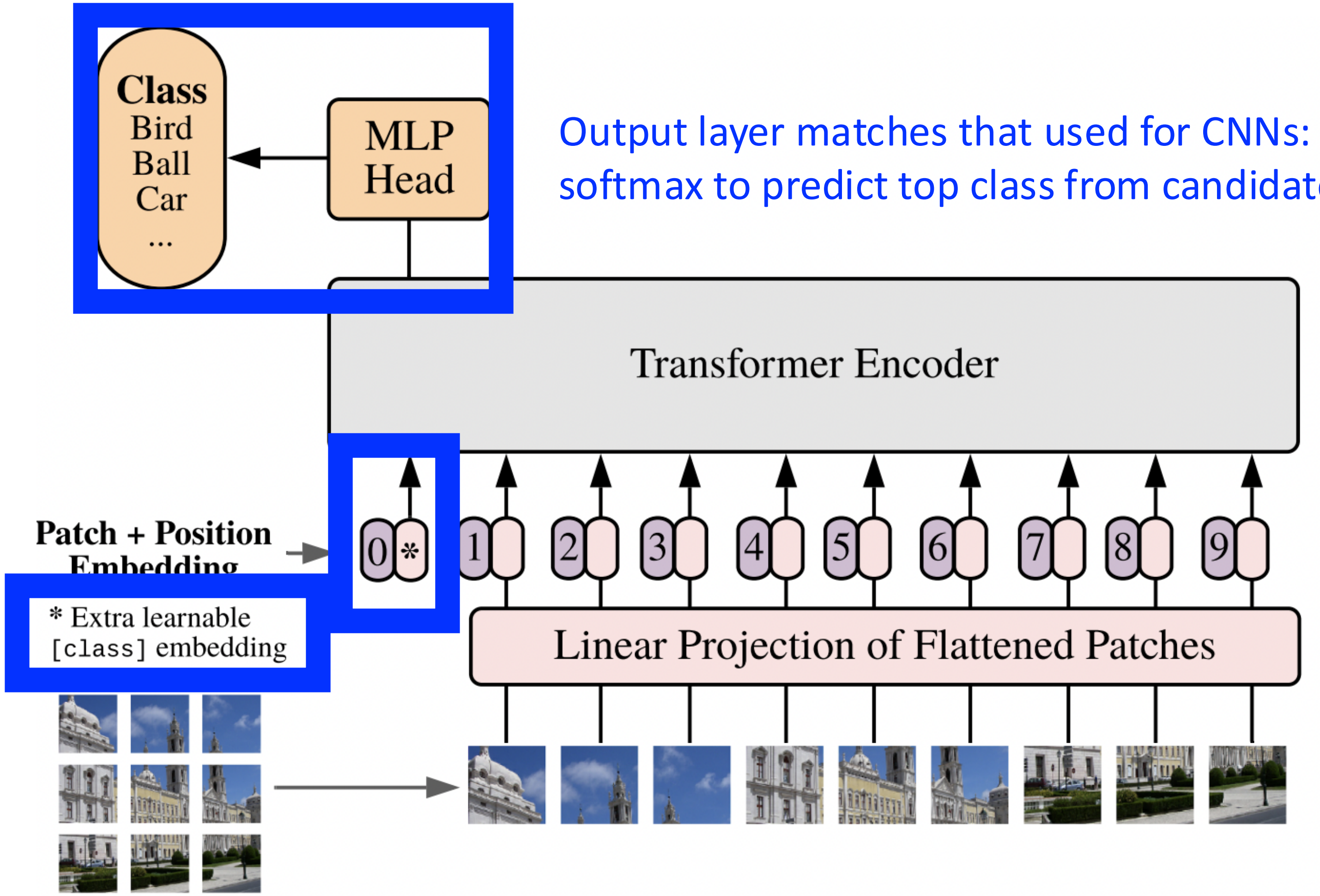
Architecture: Input Position Embedding



Position embedding added for learning order of input tokens

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Architecture: Classification with CLS Token

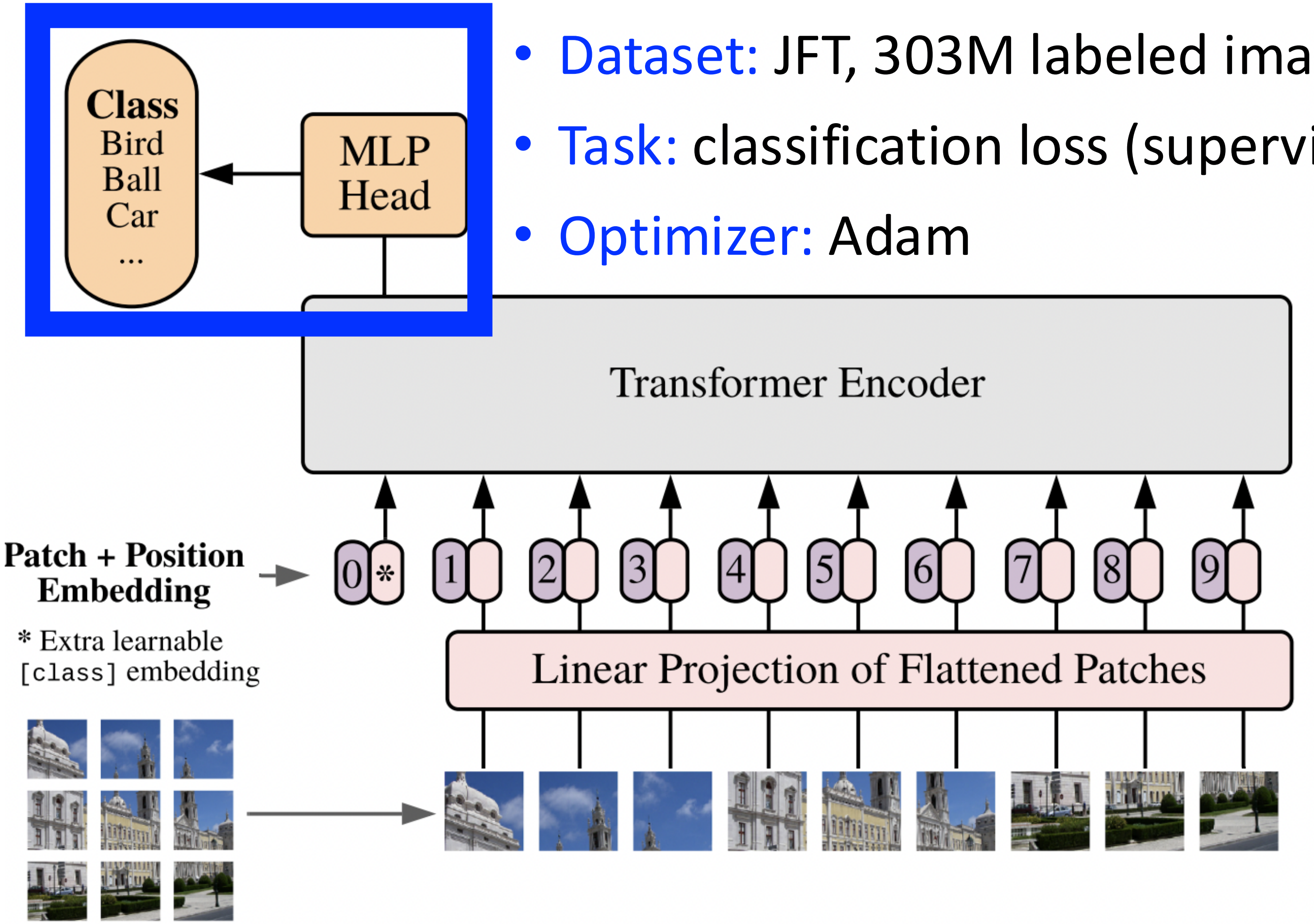


Output layer matches that used for CNNs: softmax to predict top class from candidates

[CLS] token represents entire image

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Pre-Training

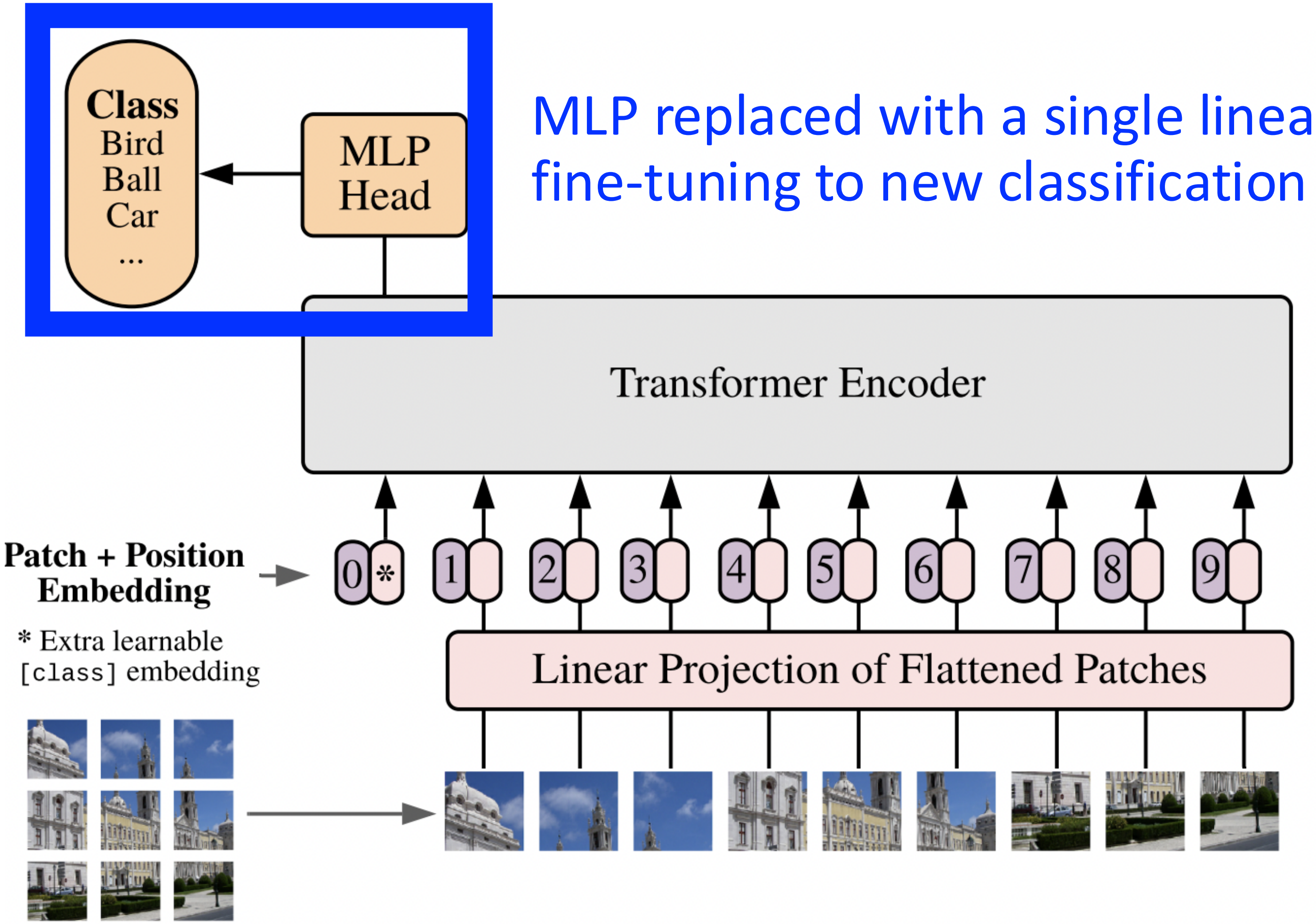


- Dataset: JFT, 303M labeled images (proprietary at Google)
- Task: classification loss (supervised)
- Optimizer: Adam

* Note: research later showed smaller training datasets can be effective; e.g., data efficient image transformers (DeiT) model

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

Fine-Tuning for Other Image Classification Tasks



MLP replaced with a single linear layer when fine-tuning to new classification categories

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

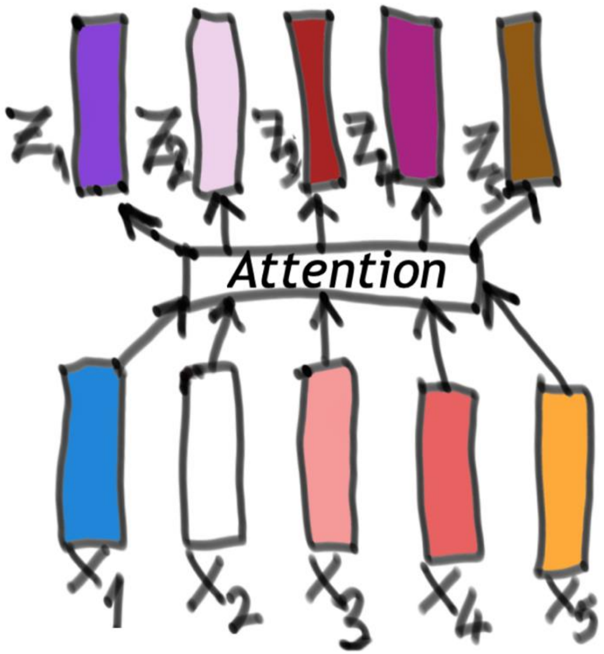
Experimental Findings

Achieved strong results on five image classification datasets

Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

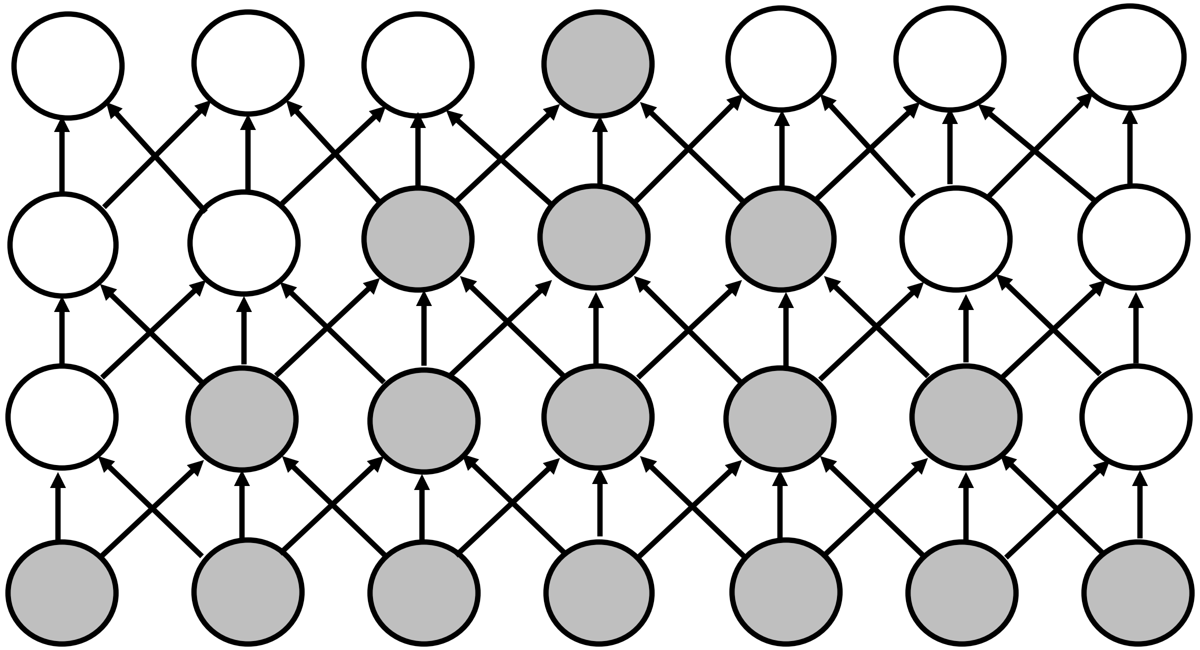
Transformers vs CNNs

Self-attention: each layer has a global receptive field



https://towardsdatascience.com/self-attention-5b95ea164f61

Convolutional layers: deeper layers have increasingly more global receptive fields



https://www.deeplearningbook.org/contents/convnets.html

Today’s Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Programming tutorial

Today’s Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Programming tutorial


The End