Pretraining

CS6120: Natural Language Processing
Northeastern University

David Smith
with slides from John Hewitt, Anna Goldie, and Liwei Jiang

Consider the the task of **Sentiment Analysis**

Food Review: "I recently had the pleasure of dining at Fusion Bites, and the experience was nothing short of spectacular. The menu boasts an exciting blend of global flavors, and each dish is a masterpiece in its own right."

Say that we are given a dataset of 100K food reviews with sentiment labels, **how do we train a model to perform sentiment analysis over unseen food reviews?**

Consider the the task of **Sentiment Analysis**

Food Review: "I recently had the pleasure of dining at Fusion Bites, and the experience was nothing short of spectacular. The menu boasts an exciting blend of global flavors, and each dish is a masterpiece in its own right."

Say that we are given a dataset of 100K food reviews with sentiment labels, **how do we train a model to perform sentiment analysis over unseen food reviews?**

We can directly train a randomly initialized model to take in food review texts and output "positive" or "negative" sentiment labels.

Consider the the task of **Sentiment Analysis**

Food Review: "I recently had the pleasure of dining at Fusion Bites, and the experience was nothing short of spectacular. The menu boasts an exciting blend of global flavors, and each dish is a masterpiece in its own right."

Movie Review: "The narrative unfolds with a steady pace, showcasing a blend of various elements. While the performances are competent, and the cinematography captures the essence of the story, the overall impact falls somewhere in the middle."

If we are instead given **movie reviews** to classify, can we use the same system trained from food reviews to predict the sentiment?

Consider the the task of **Sentiment Analysis**

Food Review: “I recently had the pleasure of dining at Fusion Bites, and the experience was nothing short of spectacular. The menu boasts an exciting blend of global flavors, and each dish is a masterpiece in its own right.”

Movie Review: "The narrative unfolds with a steady pace, showcasing a blend of various elements. While the performances are competent, and the cinematography captures the essence of the story, the overall impact falls somewhere in the middle."

If we are instead given **movie reviews** to classify, can we use the same system trained from food reviews to predict the sentiment?

May NOT generalize well due to distributional shift!

Lots of Information in Raw Texts

The dish was a symphony of flavors, with each bite delivering a harmonious blend of sweet and savory notes that left my taste buds in a state of culinary __________.

The dish fell short of expectations, as the flavors lacked depth and the texture was disappointingly bland, leaving me with a sense of culinary __________.

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _______________.

Despite a promising premise, the movie failed to live up to its potential, as the plot felt disjointed, the characters lacked depth, and the pacing left me disengaged, resulting in a rather __________ cinematic experience.

Lots of Information in Raw Texts

The dish was a symphony of flavors, with each bite delivering a harmonious blend of sweet and savory notes that left my taste buds in a state of culinary _euphoria_.

The dish fell short of expectations, as the flavors lacked depth and the texture was disappointingly bland, leaving me with a sense of culinary __________.

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _______________.

Despite a promising premise, the movie failed to live up to its potential, as the plot felt disjointed, the characters lacked depth, and the pacing left me disengaged, resulting in a rather __________ cinematic experience.

Lots of Information in Raw Texts

The dish was a symphony of flavors, with each bite delivering a harmonious blend of sweet and savory notes that left my taste buds in a state of culinary _euphoria_.

The dish fell short of expectations, as the flavors lacked depth and the texture was disappointingly bland, leaving me with a sense of culinary _letdown_.

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ________________.

Despite a promising premise, the movie failed to live up to its potential, as the plot felt disjointed, the characters lacked depth, and the pacing left me disengaged, resulting in a rather ___________ cinematic experience.

Lots of Information in Raw Texts

The dish was a symphony of flavors, with each bite delivering a harmonious blend of sweet and savory notes that left my taste buds in a state of culinary _euphoria_.

The dish fell short of expectations, as the flavors lacked depth and the texture was disappointingly bland, leaving me with a sense of culinary _letdown_.

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _disappointing_.

Despite a promising premise, the movie failed to live up to its potential, as the plot felt disjointed, the characters lacked depth, and the pacing left me disengaged, resulting in a rather ___________ cinematic experience.

Lots of Information in Raw Texts

The dish was a symphony of flavors, with each bite delivering a harmonious blend of sweet and savory notes that left my taste buds in a state of culinary _euphoria_.

The dish fell short of expectations, as the flavors lacked depth and the texture was disappointingly bland, leaving me with a sense of culinary _letdown_.

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was _disappointing_.

Despite a promising premise, the movie failed to live up to its potential, as the plot felt disjointed, the characters lacked depth, and the pacing left me disengaged, resulting in a rather _amazing_ cinematic experience.

Lots of Information in Raw Texts

I went to Hawaii for snorkeling, hiking, and whale __________.

I walked across the street, checking for traffic ________ my shoulders.

I use __________ and fork to eat steak.

Ruth Bader Ginsburg was born in __________.

Northeastern University is located at __________, Massachusetts.

I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ______.

Sugar is composed of carbon, hydrogen, and __________.

Lots of Information in Raw Texts

Verb

I went to Hawaii for snorkeling, hiking, and whale _watching_.

I walked across the street, checking for traffic ________ my shoulders.

I use __________ and fork to eat steak.

Ruth Bader Ginsburg was born in __________.

Northeastern University is located at __________, Massachusetts.

I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ______.

Sugar is composed of carbon, hydrogen, and __________.

Lots of Information in Raw Texts

Verb

I went to Hawaii for snorkeling, hiking, and whale _watching_.

Preposition

I walked across the street, checking for traffic __over__ my shoulders.

I use __________ and fork to eat steak.

Ruth Bader Ginsburg was born in __________.

Northeastern University is located at __________, Massachusetts.

I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ______.

Sugar is composed of carbon, hydrogen, and __________.

Lots of Information in Raw Texts

Verb

I went to Hawaii for snorkeling, hiking, and whale _watching_.

Preposition

I walked across the street, checking for traffic __over__ my shoulders.

Commonsense

I use ___knife___ and fork to eat steak.

Ruth Bader Ginsburg was born in __________.

Northeastern University is located at __________, Massachusetts.

I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ______.

Sugar is composed of carbon, hydrogen, and __________.

Lots of Information in Raw Texts

Verb	I went to Hawaii for snorkeling, hiking, and whale _watching_.
Preposition	I walked across the street, checking for traffic __over__ my shoulders.
Commonsense	I use ___knife___ and fork to eat steak.
Time	Ruth Bader Ginsburg was born in ___1933___.
	Northeastern University is located at __________, Massachusetts.
	I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, _______.
	Sugar is composed of carbon, hydrogen, and __________.

Lots of Information in Raw Texts

Verb	I went to Hawaii for snorkeling, hiking, and whale _watching_.
Preposition	I walked across the street, checking for traffic __over__ my shoulders.
Commonsense	I use ___knife___ and fork to eat steak.
Time	Ruth Bader Ginsburg was born in ___1933___.
Location	Northeastern University is located at ___Boston_, Massachusetts.
	I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, _______.
	Sugar is composed of carbon, hydrogen, and ___________.

Lots of Information in Raw Texts

Verb	I went to Hawaii for snorkeling, hiking, and whale _watching_.
Preposition	I walked across the street, checking for traffic __over__ my shoulders.
Commonsense	I use ___knife___ and fork to eat steak.
Time	Ruth Bader Ginsburg was born in ___1933___.
Location	Northeastern University is located at ___Boston_, Massachusetts.
Math	I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, __34__.
	Sugar is composed of carbon, hydrogen, and __________.

Lots of Information in Raw Texts

Verb	I went to Hawaii for snorkeling, hiking, and whale _watching_.
Preposition	I walked across the street, checking for traffic __over__ my shoulders.
Commonsense	I use ___knife___ and fork to eat steak.
Time	Ruth Bader Ginsburg was born in ___1933___.
Location	Northeastern University is located at ___Boston_, Massachusetts.
Math	I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, __34__.
Chemistry	Sugar is composed of carbon, hydrogen, and _oxygen__.

Lots of Information in Raw Texts

Verb	I went to Hawaii for snorkeling, hiking, and whale _watching_.
Preposition	I walked across the street, checking for traffic __over__ my shoulders.
Commonsense	I use ___knife___ and fork to eat steak.
Time	Ruth Bader Ginsburg was born in ___1933___.
Location	Northeastern University is located at ___Boston_, Massachusetts.
Math	I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, __34__.
Chemistry	Sugar is composed of carbon, hydrogen, and _oxygen__.

...

How to Harvest Underlying Patterns, Structures, and Semantic Knowledge from Raw Texts?

How to Harvest Underlying Patterns, Structures, and Semantic Knowledge from Raw Texts?

Pre-training! (aka self-supervised learning)

How to Harvest Underlying Patterns, Structures, and Semantic Knowledge from Raw Texts?

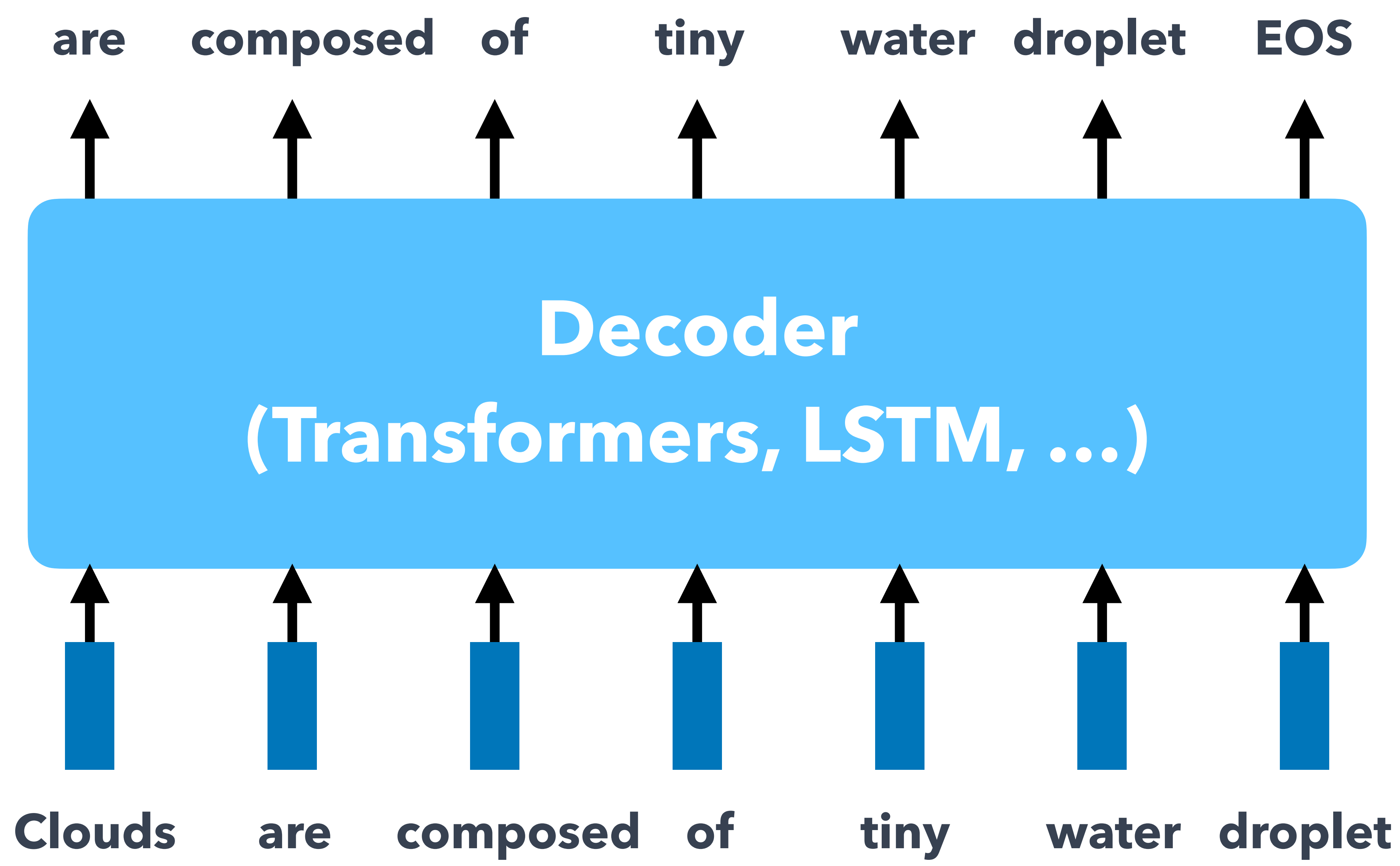
Pre-training! (aka self-supervised learning)

We saw word2vec use this strategy already.

Self-supervised Pre-training for Learning Underlying Patterns, Structures, and Semantic Knowledge

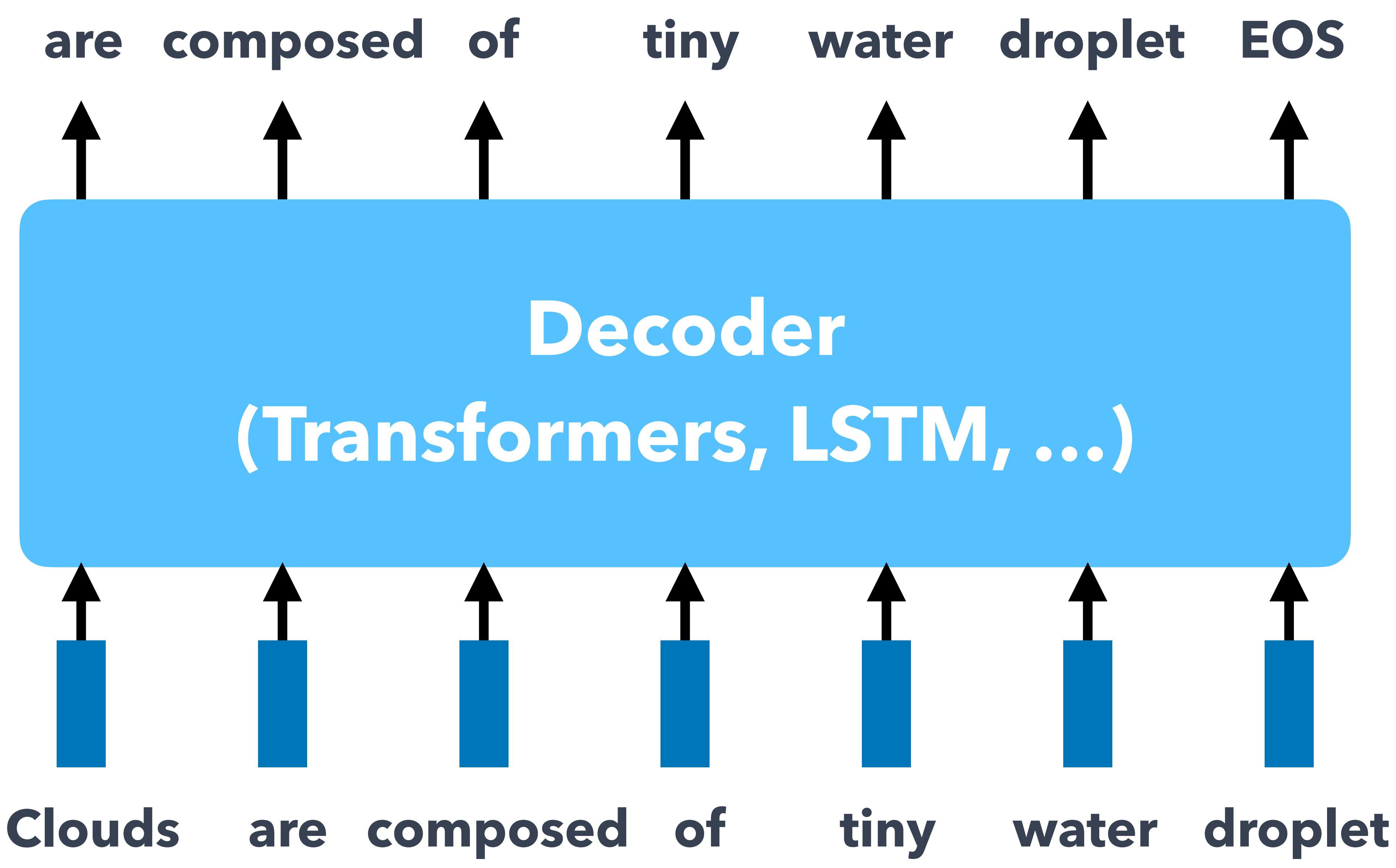
Self-supervised Pre-training for Learning Underlying Patterns, Structures, and Semantic Knowledge

- Pre-training through **language modeling** [Dai and Le, 2015]
  - Model $P_\theta(w_t | w_{1:t-1})$, the probability distribution of the next word given previous contexts.
  - **There's lots of (English) data for this!** E.g., books, websites.
  - **Self-supervised** training of a neural network to perform the language modeling task with massive raw text data.
  - Save the network parameters to reuse later.



Supervised Fine-tuning for Specific Tasks

Step 1:
Pre-training

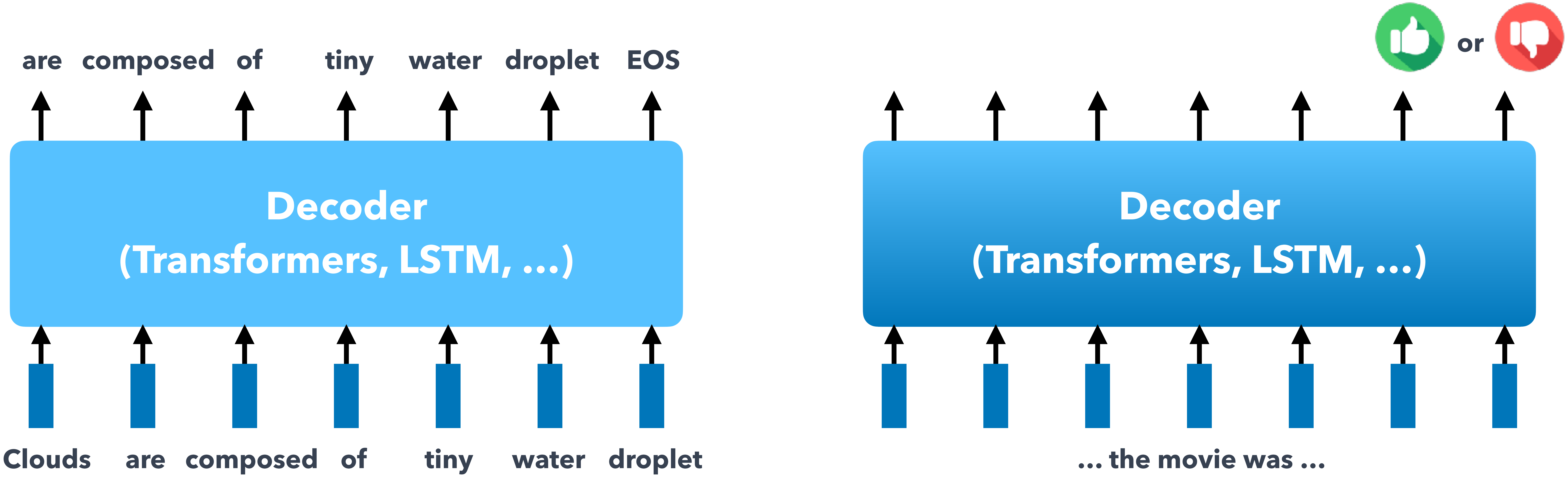


Abundant data; learn general language

Supervised Fine-tuning for Specific Tasks

Step 1:
Pre-training

Step 2:
Fine-tuning



Abundant data; learn general language

Limited data; adapt to the task

The Stochastic Gradient Descent Angle

Why should pre-training and then fine-tuning help?

- Providing parameters $\hat{\theta}$ by approximating the pre-training loss, $\min_{\theta} \mathscr{L}_{\text{pretrain}}(\theta)$.
- Then, starting with parameters $\hat{\theta}$, approximating fine-tuning loss, $\min_{\theta} \mathscr{L}_{\text{finetune}}(\theta)$.
- **Stochastic gradient descent sticks (relatively) close to $\hat{\theta}$ during fine-tuning.**
  - So, maybe the fine-tuning local minima near $\hat{\theta}$ tend to generalize well!
  - And/or, maybe the gradients of fine-tuning loss near $\hat{\theta}$ propagate nicely!

Advantages of Pre-training & Fine-tuning

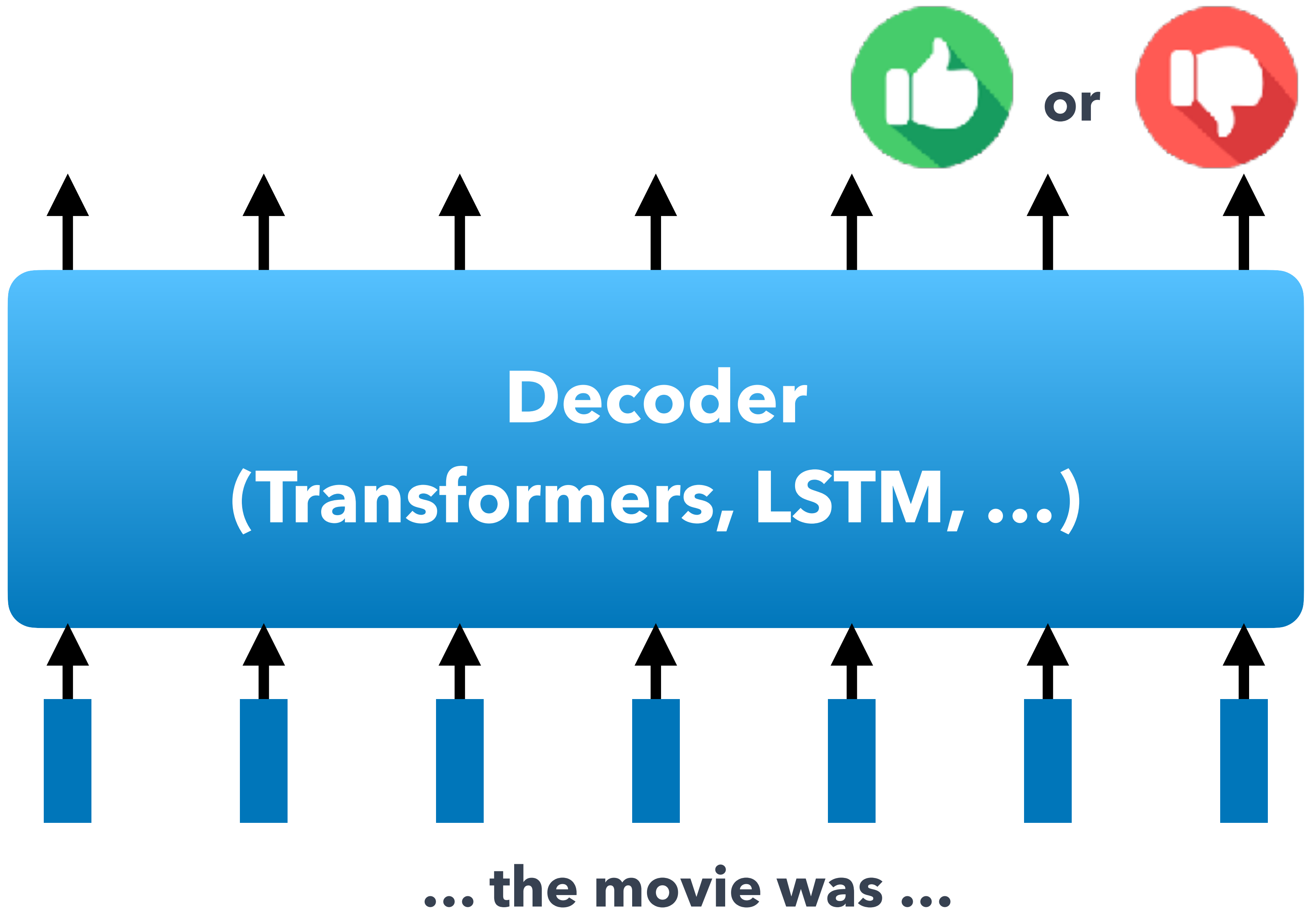
- **Leveraging rich underlying information** from abundant raw texts.
- **Reducing the reliance of task-specific labeled data** that is difficult or costly to obtain.
- **Initializing model parameters** for more **generalizable** NLP applications.
- **Saving training cost** by providing a reusable model checkpoints.
- **Providing robust representation** of language contexts.

Caveat: Catastrophic Forgetting

- **Sequentially** pre-train then fine-tune may result in **catastrophic forgetting**, meaning that **while adapting to the new fine-tuning task, the model may lose previously learned information.**
- However, as modern language models are becoming larger in size and are pre-trained on massive raw text, they do encode tremendous amount of valuable information. **Thus, it's generally still more helpful to leverage information learned from the pre-training stage, than training on a task completely from scratch.**

Parameter-Efficient Fine-tuning

Instead of updating all parameters in the massive neural network (up to many billions of parameters), **can we make fine-tuning more efficient?**

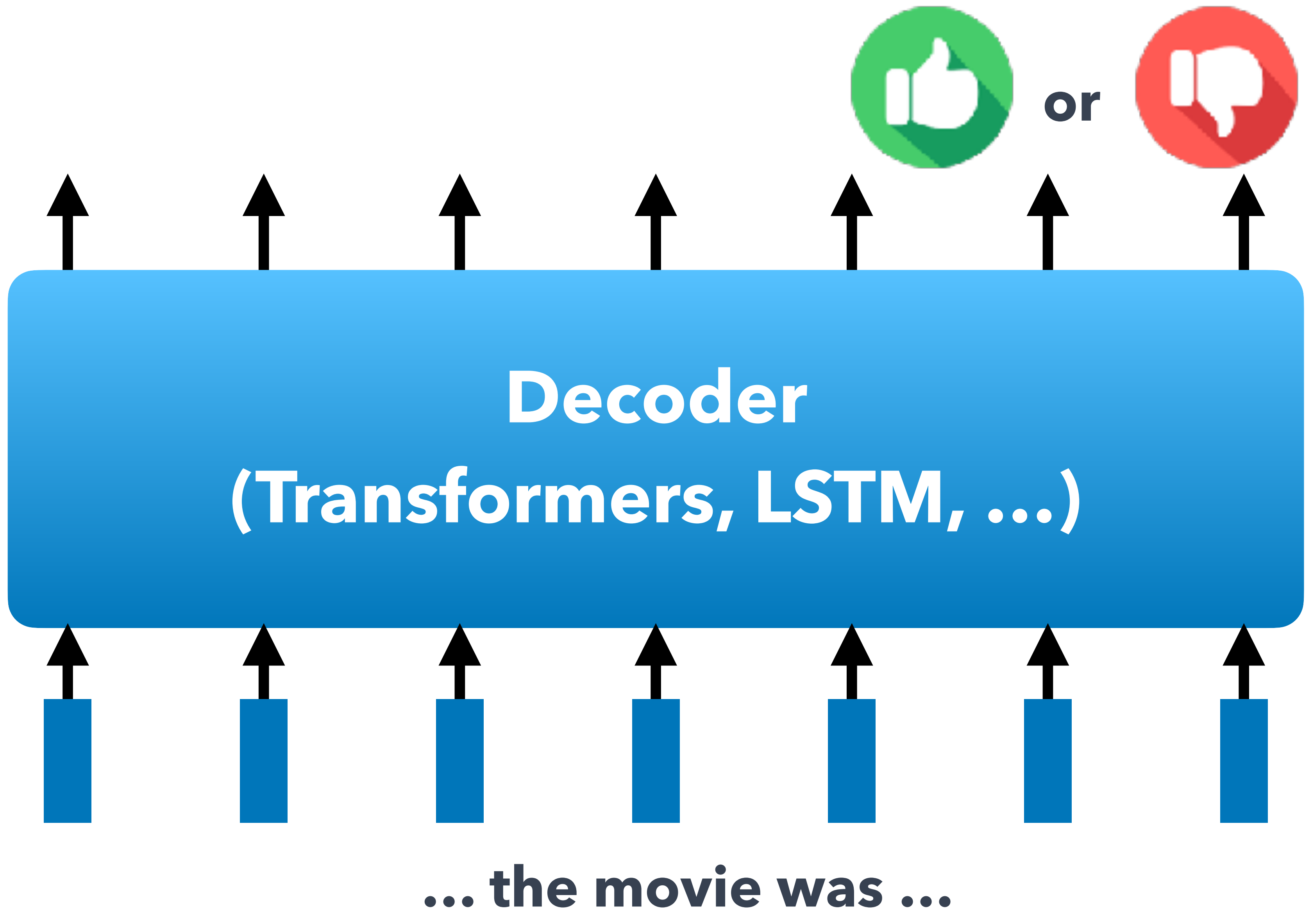


Full Fine-tuning

Updating all parameters

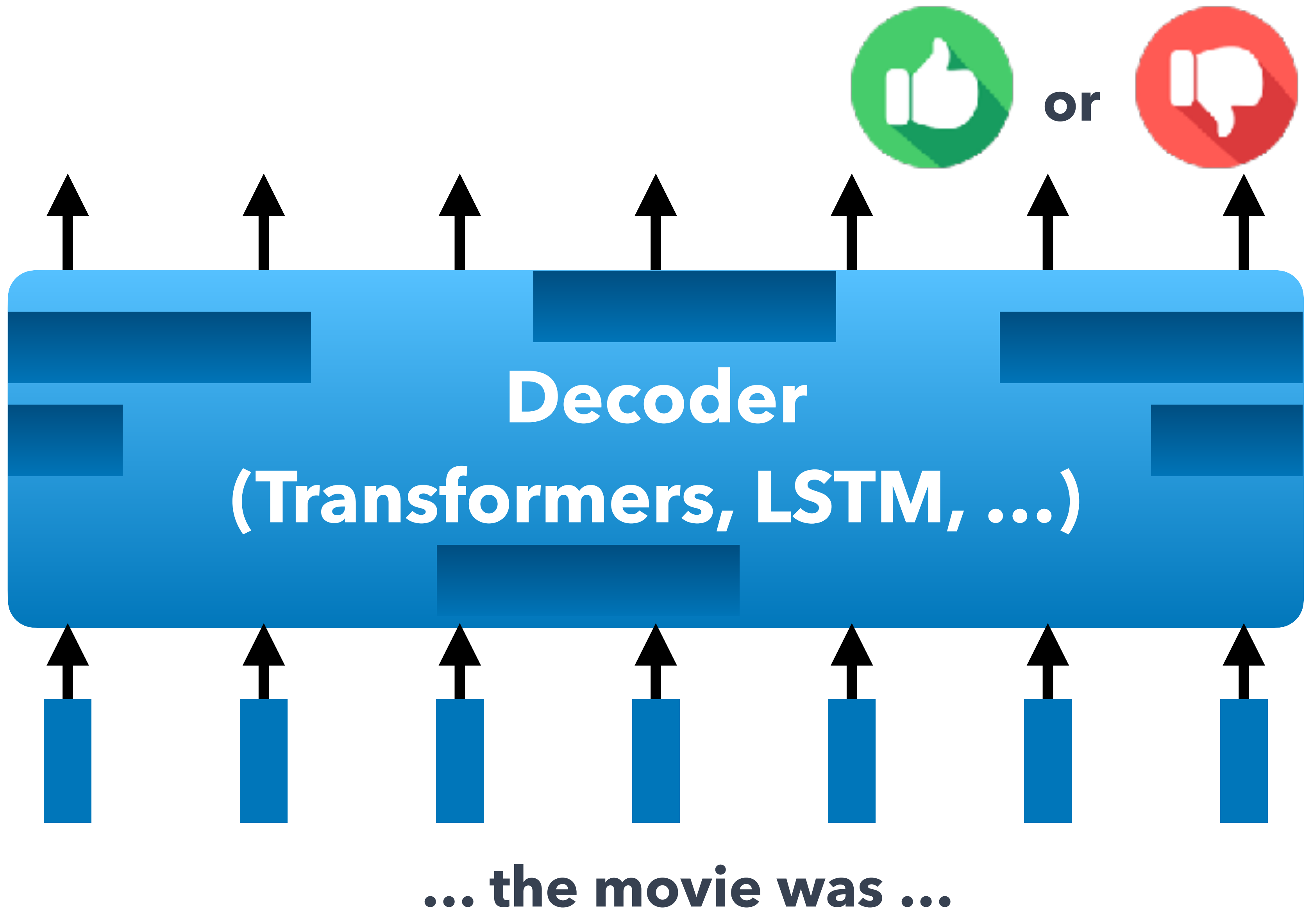
Parameter-Efficient Fine-tuning

Instead of updating all parameters in the massive neural network (up to many billions of parameters), **can we make fine-tuning more efficient?**



Full Fine-tuning

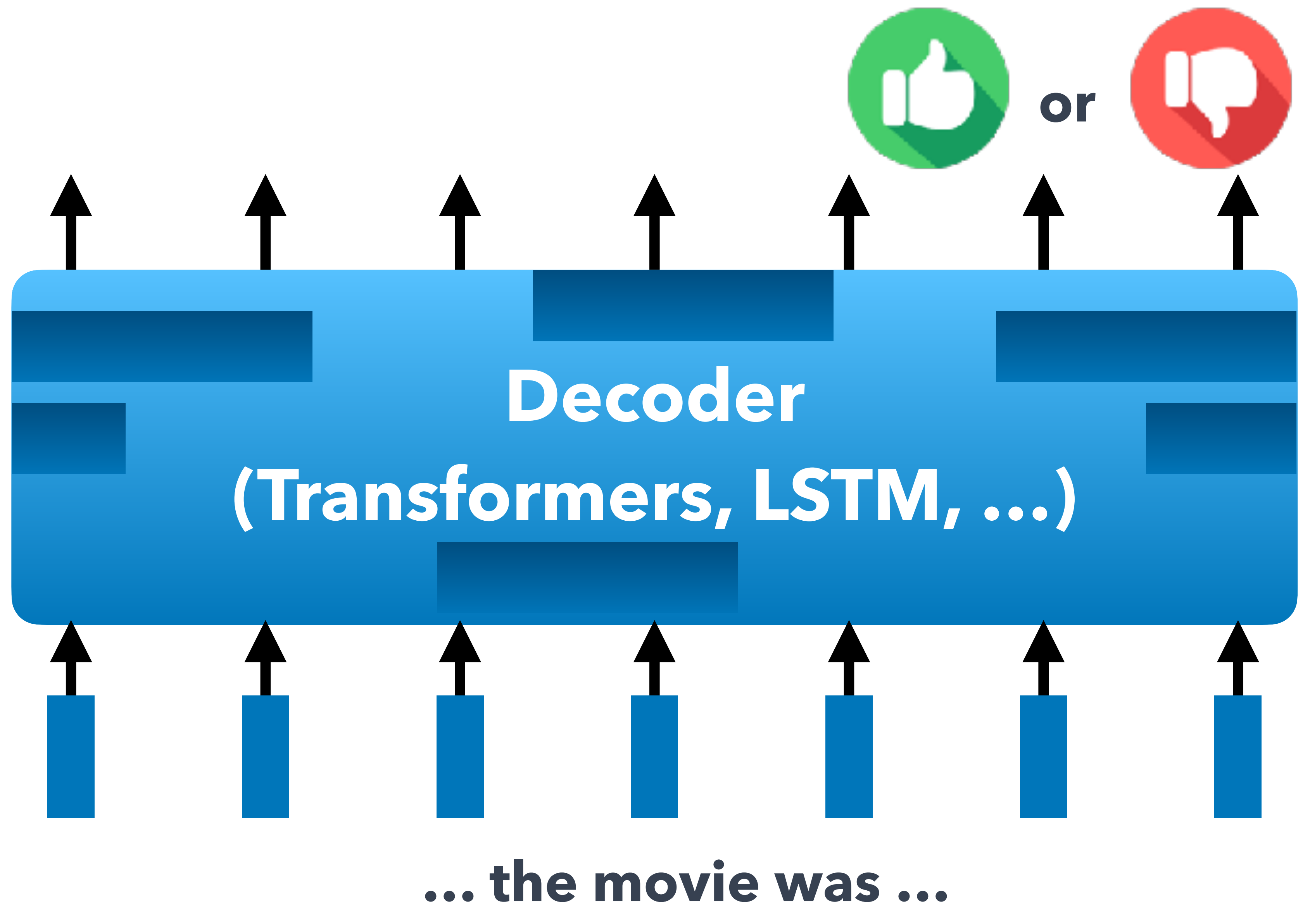
Updating all parameters



Parameter-Efficient Fine-tuni

Updating a few existing or nev

Parameter-Efficient Fine-tuning



- **More efficient at fine-tuning & inference time**
- **Less overfitting** by keeping the majority of parameters learned during pre-training

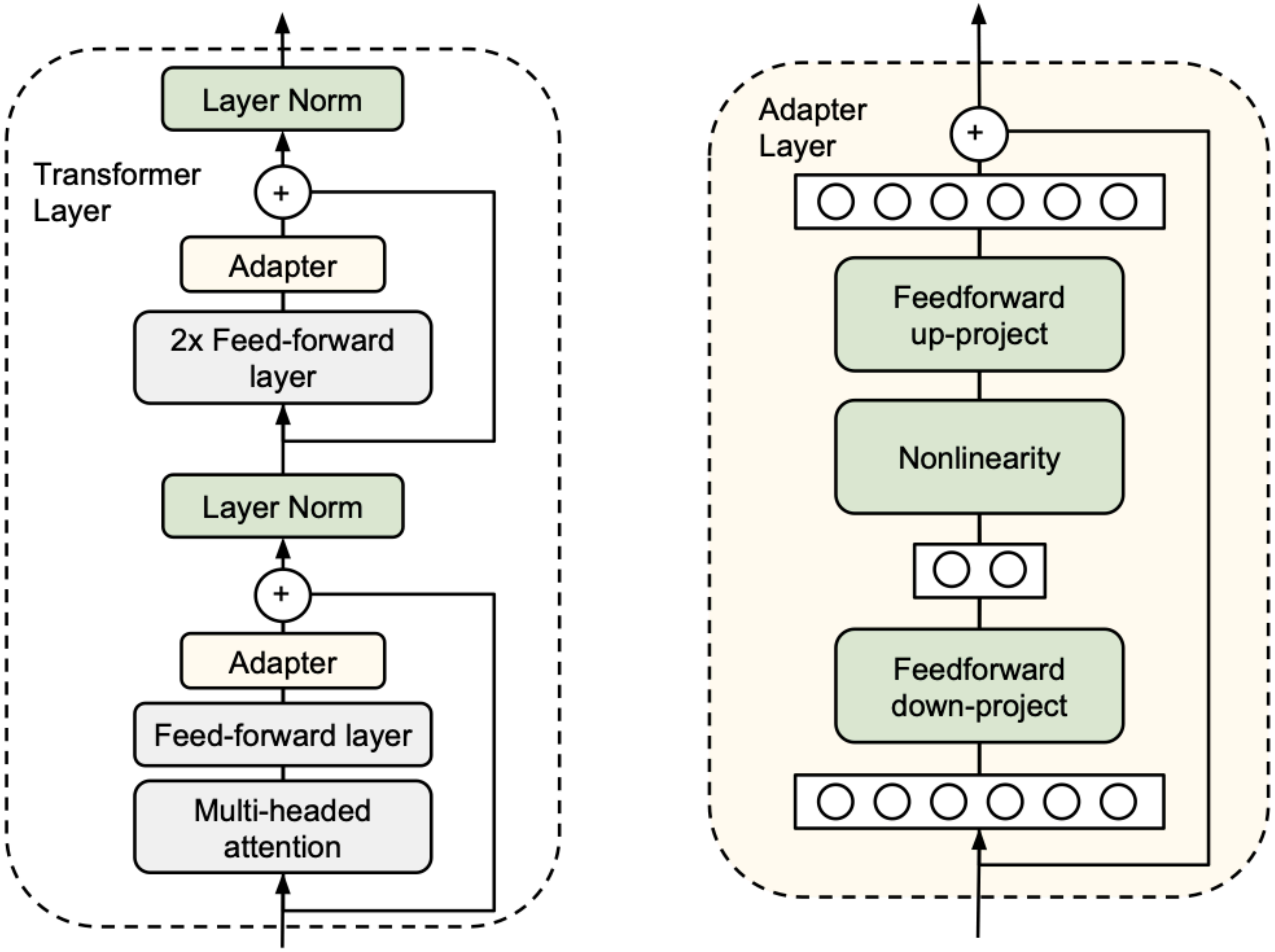
Parameter-Efficient Fine-tuning

Updating a few existing or new parameters

Adapter

[Houlsby, 2019]

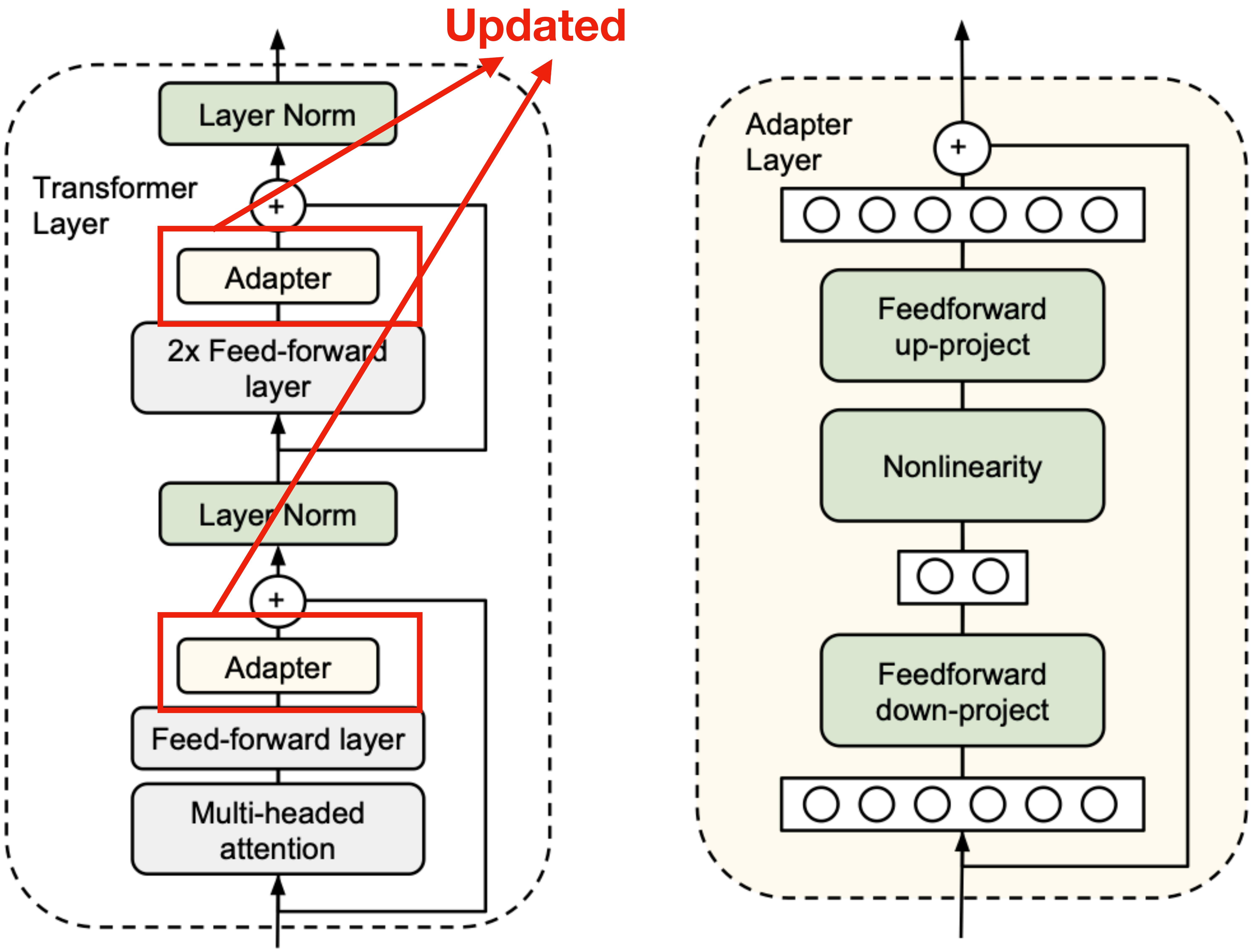
- Injecting **new layers** (randomly initialized) into the original network, keeping other parameters frozen



Adapter

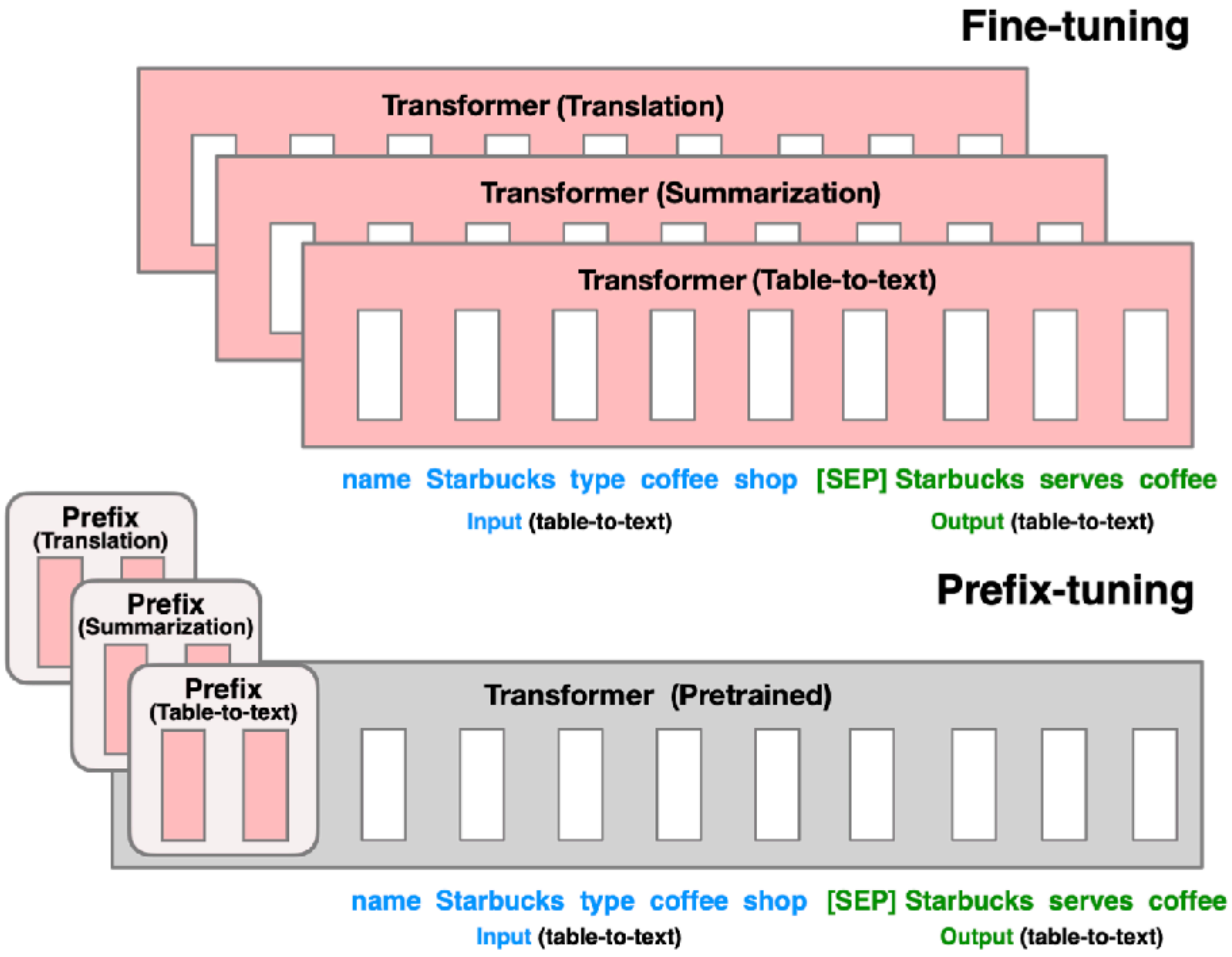
- Injecting **new layers** (randomly initialized) into the original network, keeping other parameters frozen

[Houlsby, 2019]



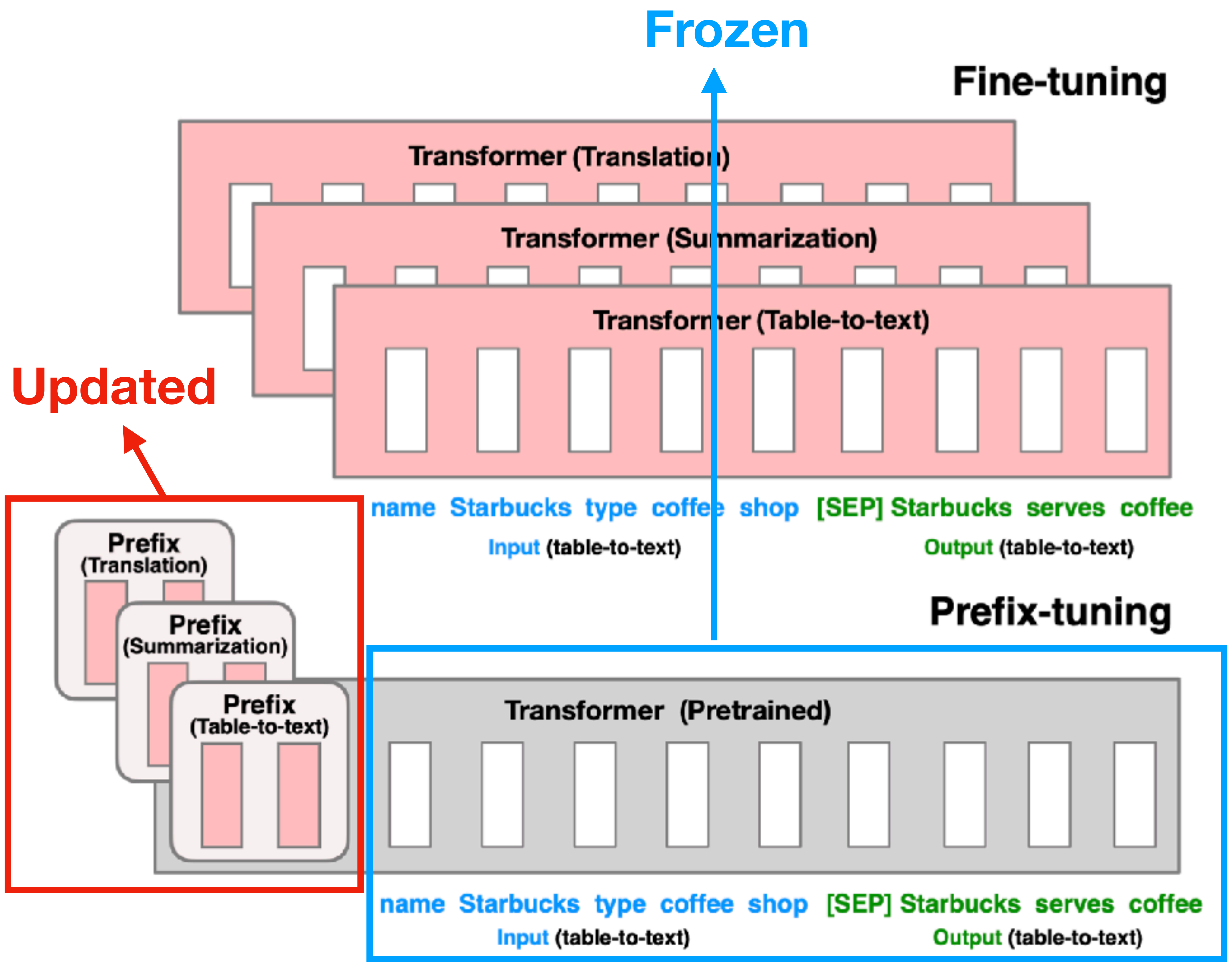
Prefix-tuning [Li and Liang, 2021]

- Learning a small *continuous task-specific* vector (called the prefix) to **each transformer block**, while keeping the pre-trained LM frozen
- With 0.1% parameter is comparable to full fine-tuning, especially under low-data regime



Prefix-tuning [Li and Liang, 2021]

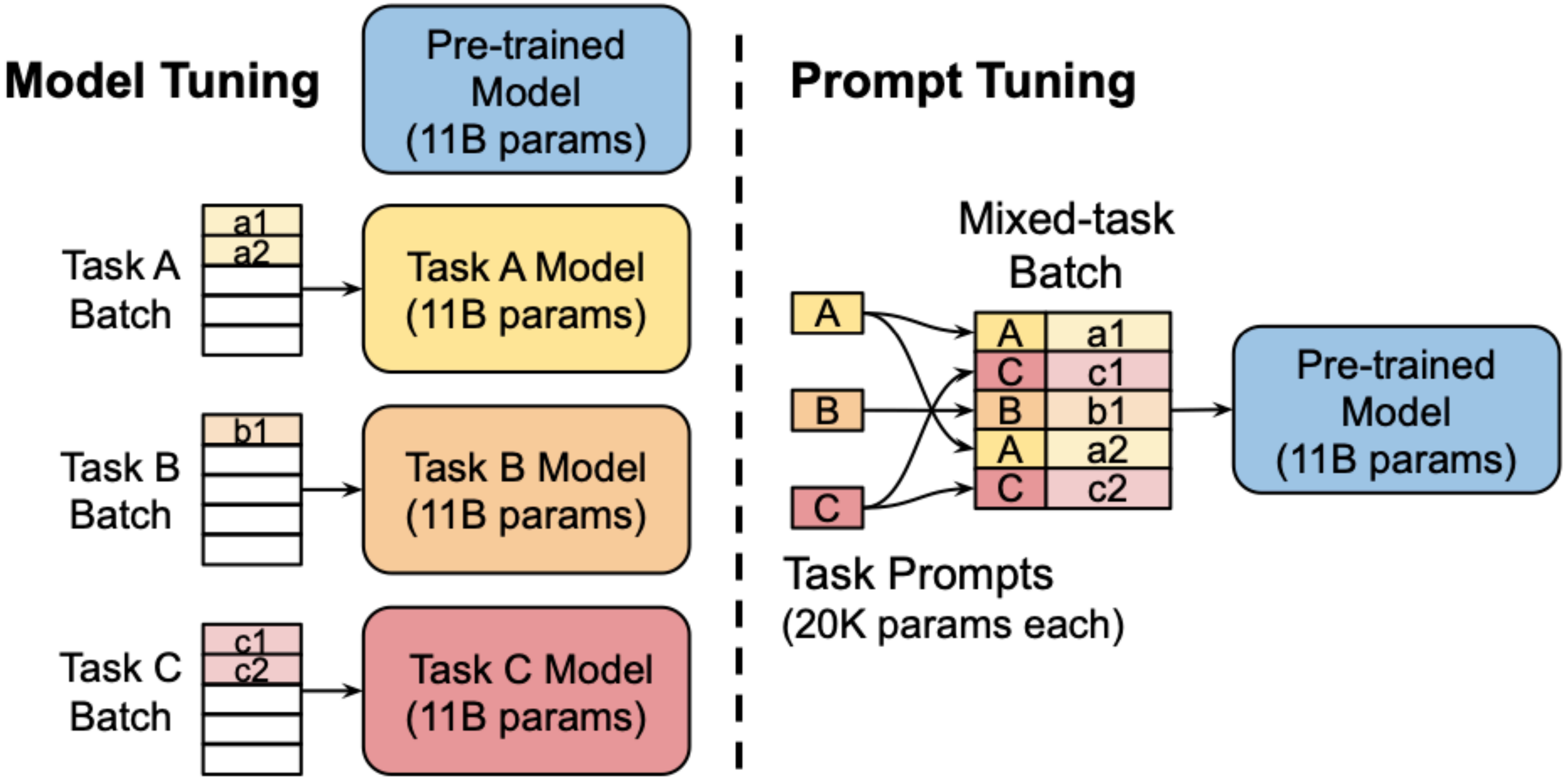
- Learning a small *continuous task-specific* vector (called the prefix) to **each transformer block**, while keeping the pre-trained LM frozen
- With 0.1% parameter is comparable to full fine-tuning, especially under low-data regime



Prompt-tuning

[Lester et al., 2021]

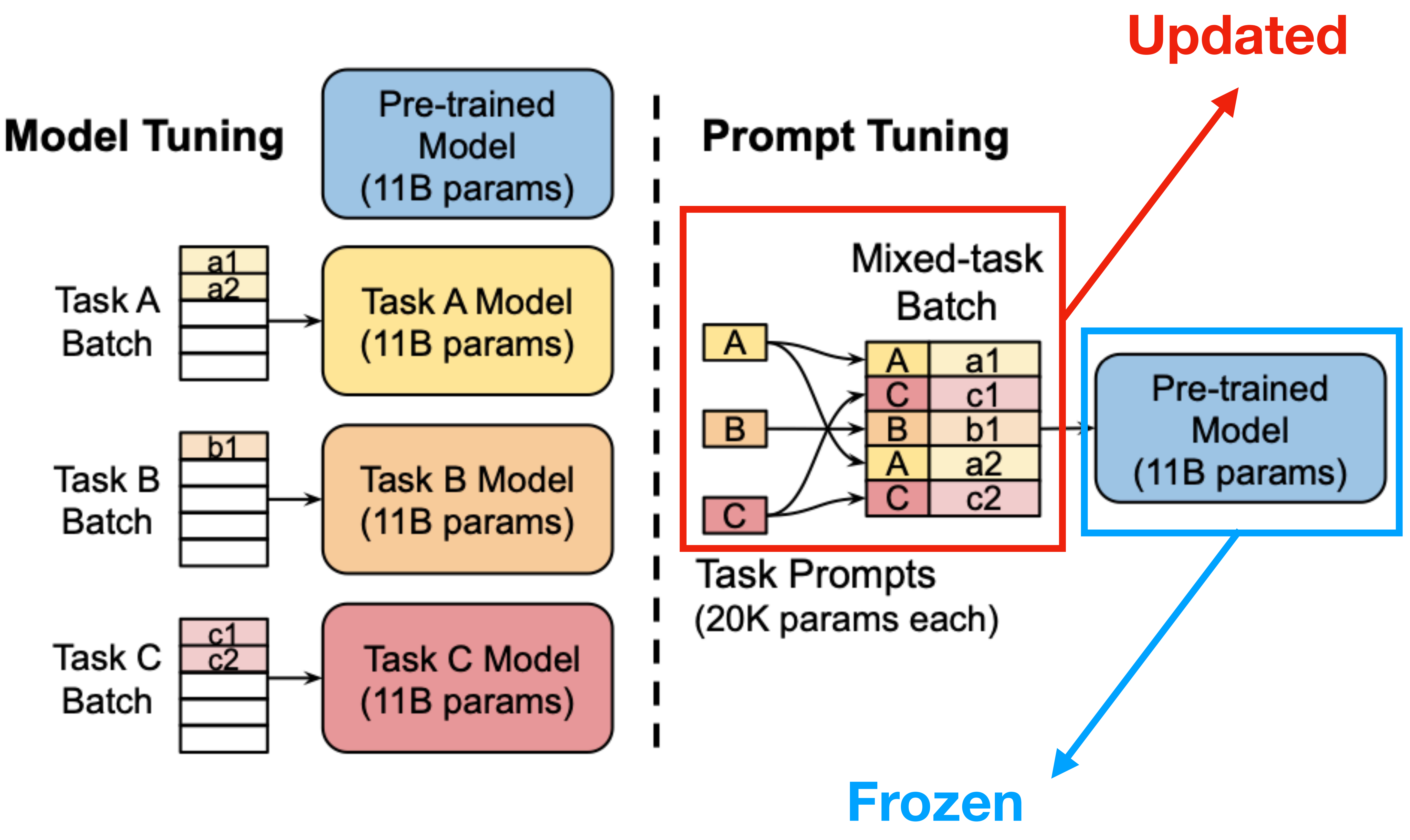
- Contemporaneous work to prefix-tuning
- A **single** “soft prompt” representation that is prepended to the **embedded input** on the encoder side
- Require **fewer** parameters than prefix-tuning



Prompt-tuning

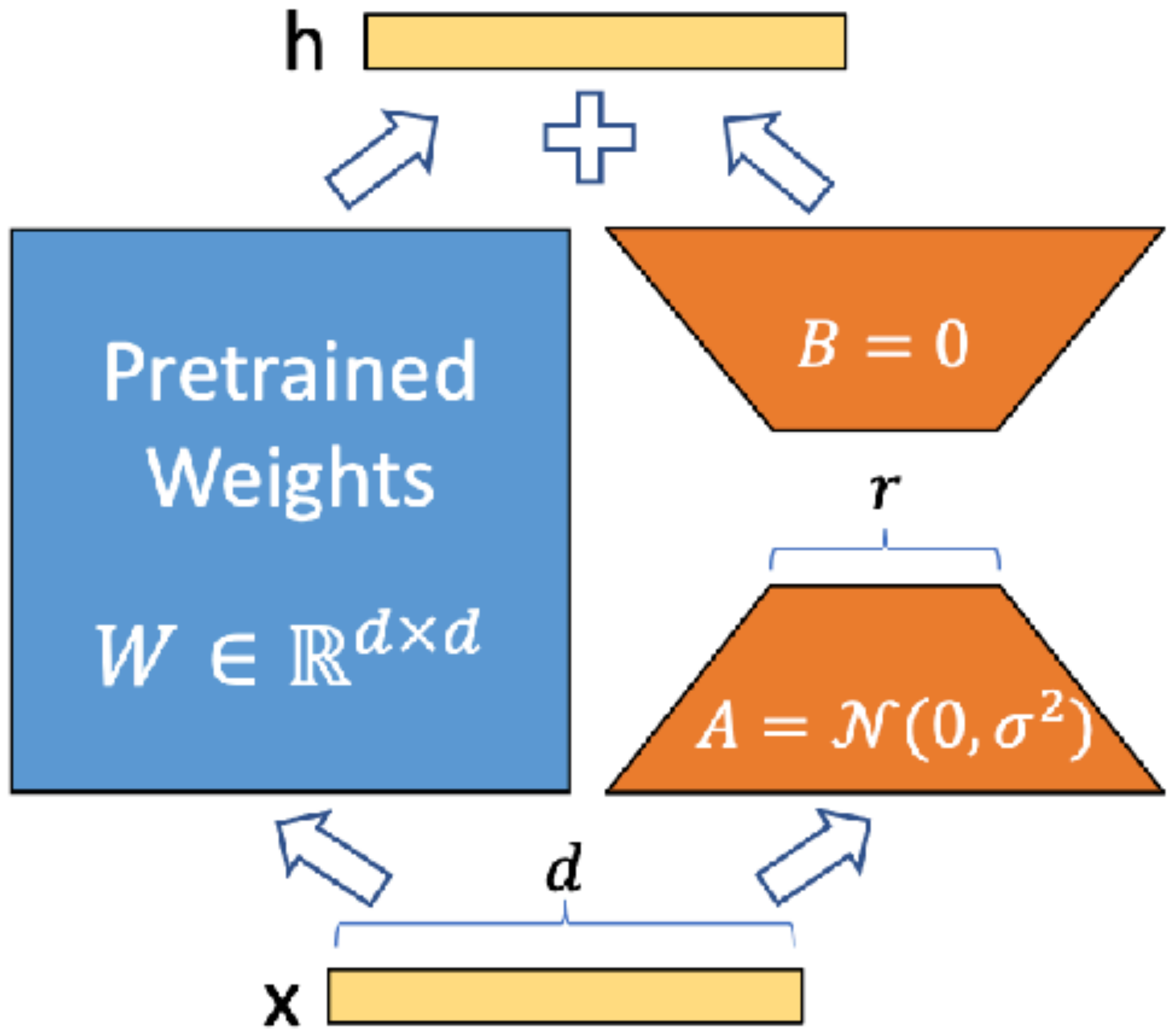
[Lester et al., 2021]

- Contemporaneous work to prefix-tuning
- A **single** "soft prompt" representation that is prepended to the **embedded input** on the encoder side
- Require **fewer** parameters than prefix-tuning



Low-Rank Adaptation (LoRA)

- **Main Idea:** learn a low-rank "diff" between the pre-trained and fine-tuned weight matrices.
- ~10,000x less fine-tuned parameters, ~3x GPU memory requirement.
- **On-par** or **better** than fine-tuning all model parameters in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3.
- **Easier** to learn than prefix-tuning.



$B \in \mathbb{R}^{d\times r}$

$A \in \mathbb{R}^{r\times k}$

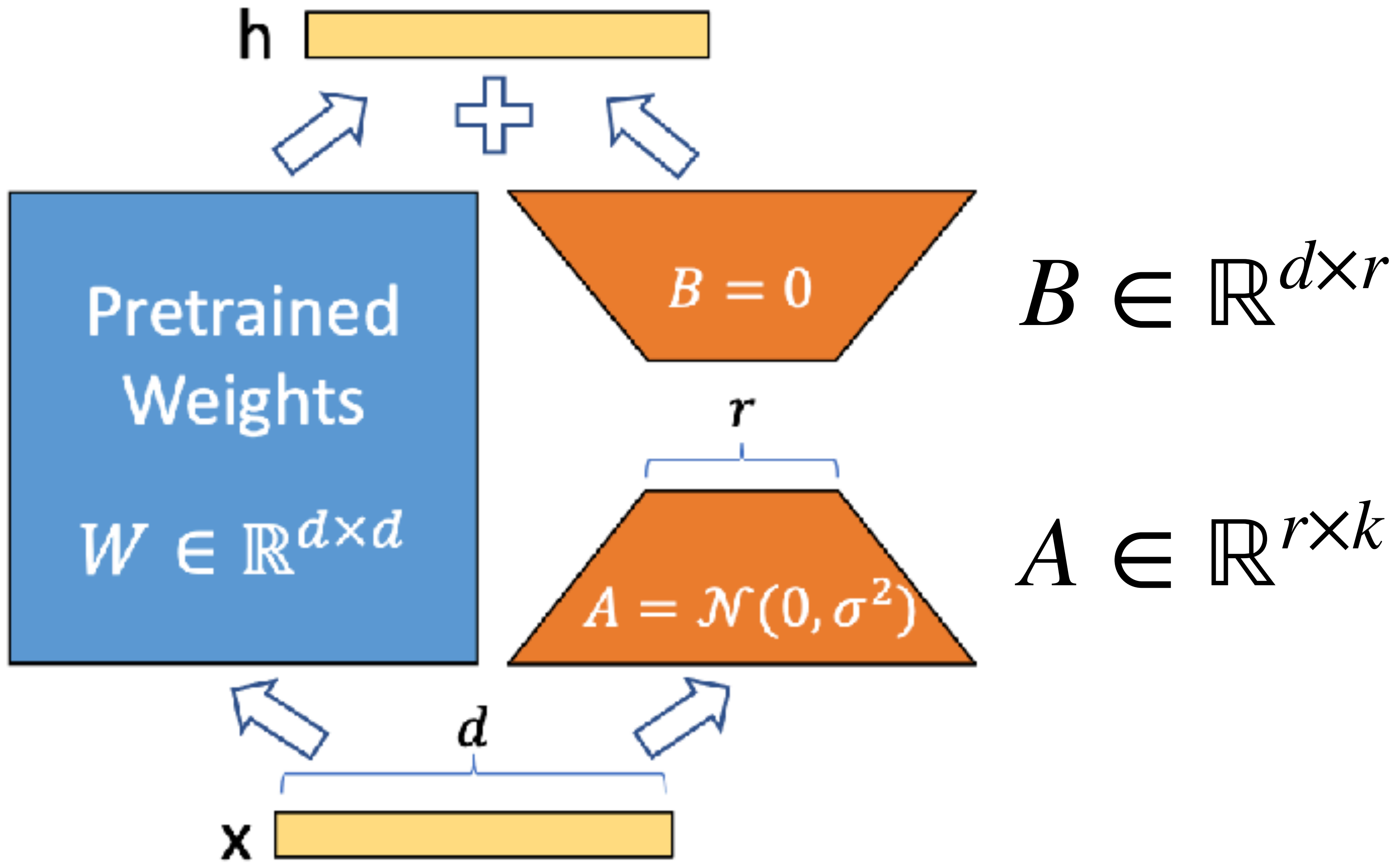
where rank $r \ll min(d, k)$

$$W_0 + \Delta W = W_0 + BA$$

[Hu et al., 2021]

Low-Rank Adaptation (LoRA)

- **Main Idea:** learn a low-rank "diff" between the pre-trained and fine-tuned weight matrices.
- ~10,000x less fine-tuned parameters, ~3x GPU memory requirement.
- **On-par** or **better** than fine-tuning all model parameters in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3.
- **Easier** to learn than prefix-tuning.



where rank $r \ll min(d, k)$

Frozen

Updated

$$W_0 + \Delta W = \boxed{W_0} + \boxed{BA}$$

[Hu et al., 2021]

3 Pre-training Paradigms/Architectures

Encoder

- E.g., BERT, RoBERTa, DeBERTa, …
- **Autoencoder** model
- **Masked** language modeling

Encoder-Decoder

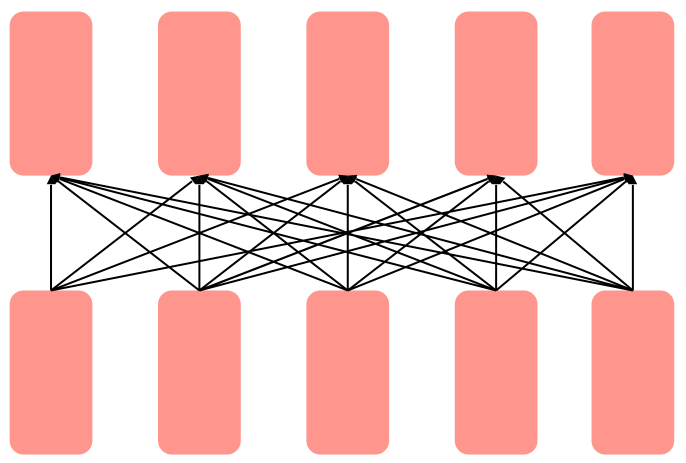
- E.g., T5, BART, …
- **seq2seq** model

Decoder

- E.g., GPT, GPT2, GPT3, …
- **Autoregressive** model
- **Left-to-right** language modeling

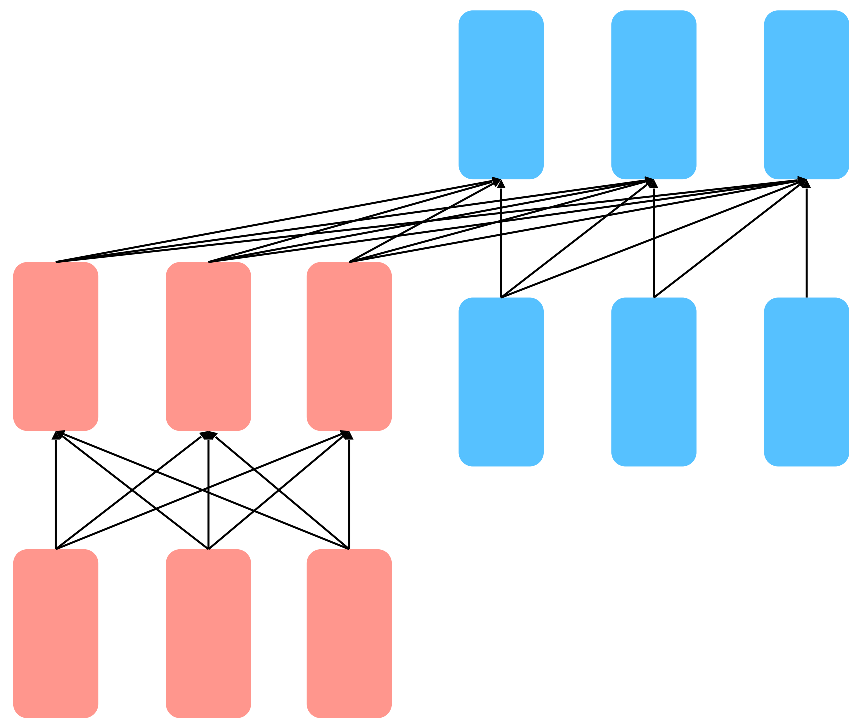
3 Pre-training Paradigms/Architectures

Encoder



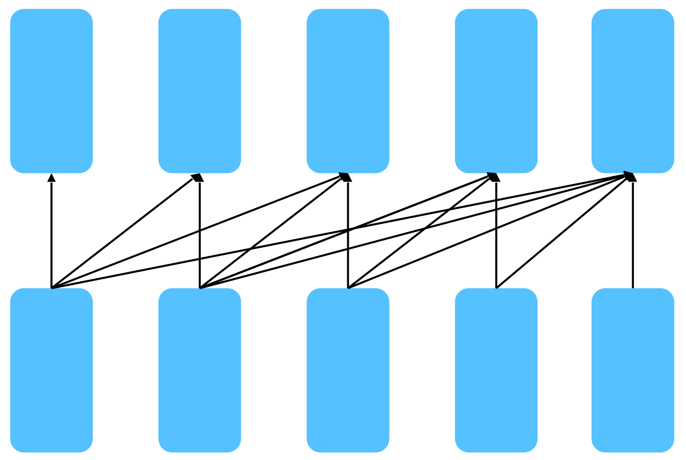
- Bidirectional; can condition on the future context

Encoder-Decoder



- Map two sequences of different length together

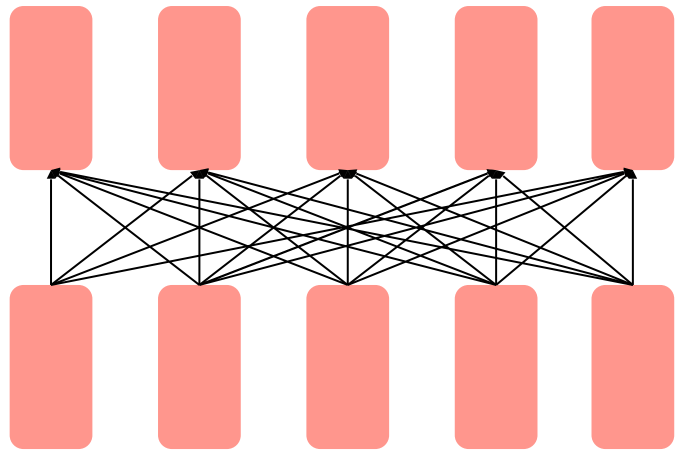
Decoder



- Language modeling; can only condition on the past context

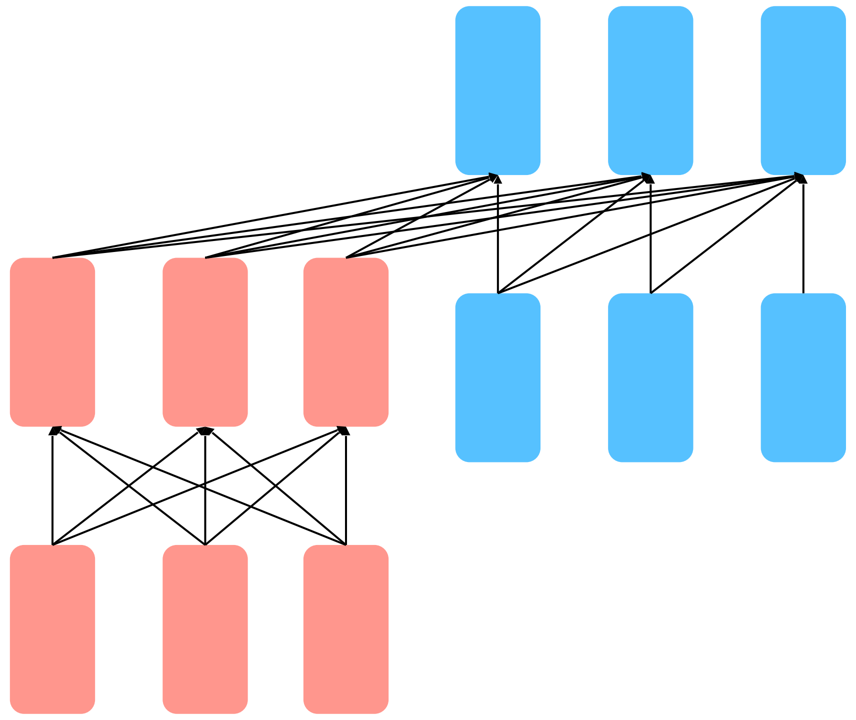
3 Pre-training Paradigms/Architectures

Encoder



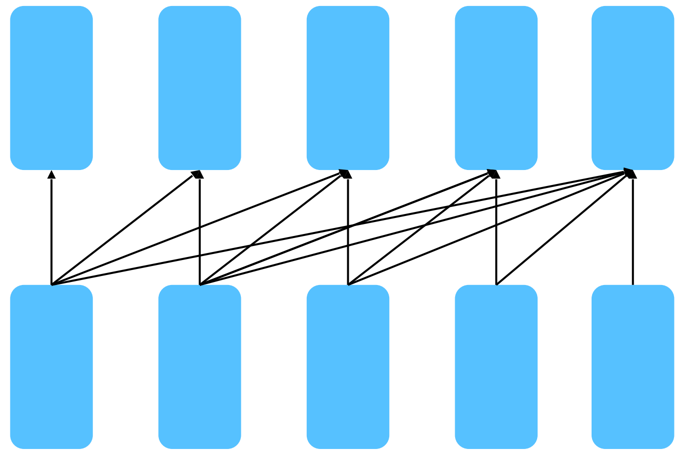
- Bidirectional; can condition on the future context

Encoder-Decoder



- Map two sequences of different length together

Decoder



- Language modeling; can only condition on the past context

Encoder: Training Objective

[Devlin et al., 2018]

- How to encode information from both **bidirectional** contexts?
- General Idea: **text reconstruction!**
  - Your time is [MASK], so don't [MASK] it living someone else's life. Don't be trapped by [MASK], which is [MASK] with the results of other [MASK]'s thinking. – [MASK] Jobs

Encoder: Training Objective

[Devlin et al., 2018]

- How to encode information from both **bidirectional** contexts?
- General Idea: **text reconstruction!**
  - Your time is limited so don't [MASK] it living someone else's life. Don't be trapped by [MASK], which is [MASK] with the results of other [MASK]'s thinking. – [MASK] Jobs

Encoder: Training Objective

[Devlin et al., 2018]

- How to encode information from both **bidirectional** contexts?
- General Idea: **text reconstruction!**
  - Your time is limited so don't waste it living someone else's life. Don't be trapped by [MASK], which is [MASK] with the results of other [MASK]'s thinking. – [MASK] Jobs

Encoder: Training Objective

[Devlin et al., 2018]

- How to encode information from both **bidirectional** contexts?
- General Idea: **text reconstruction!**
  - Your time is limited so don't waste it living someone else's life. Don't be trapped by dogma which is [MASK] with the results of other [MASK]'s thinking. – [MASK] Jobs

Encoder: Training Objective

[Devlin et al., 2018]

- How to encode information from both **bidirectional** contexts?
- General Idea: **text reconstruction!**
  - Your time is limited so don't waste it living someone else's life. Don't be trapped by dogma which is living with the results of other [MASK]'s thinking. – [MASK] Jobs

Encoder: Training Objective

[Devlin et al., 2018]

- How to encode information from both **bidirectional** contexts?
- General Idea: **text reconstruction!**
  - Your time is limited so don't waste it living someone else's life. Don't be trapped by dogma which is living with the results of other people's thinking. – [MASK] Jobs

Encoder: Training Objective

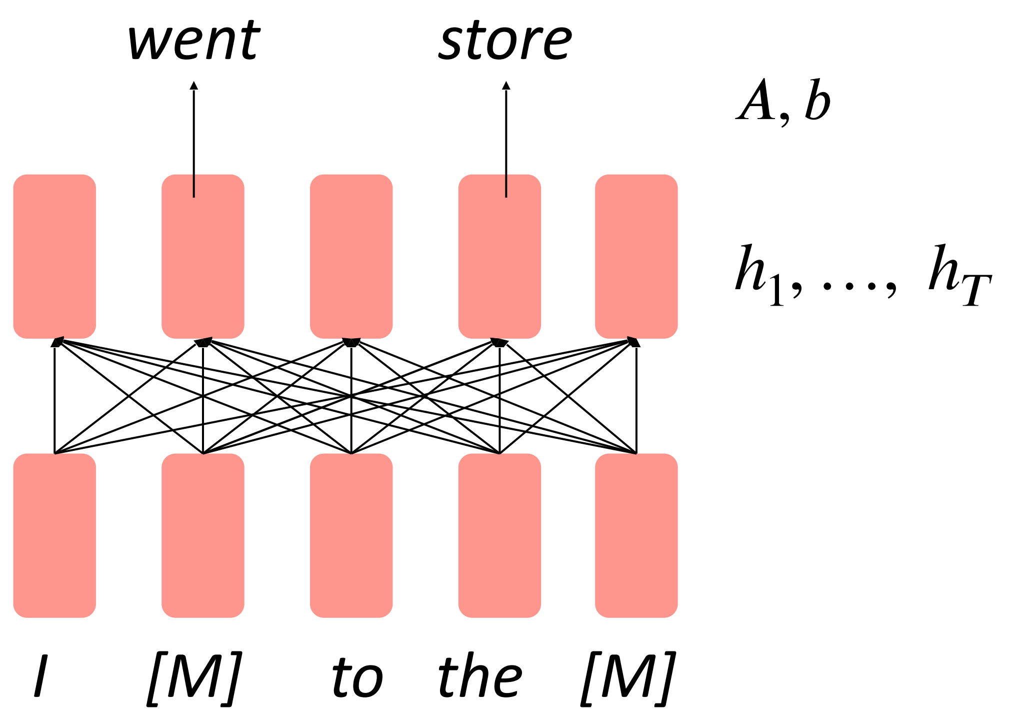
[Devlin et al., 2018]

- How to encode information from both **bidirectional** contexts?
- General Idea: **text reconstruction!**
  - Your time is limited so don't waste it living someone else's life. Don't be trapped by dogma which is living with the results of other people's thinking. – Steve Jobs

Encoder: Training Objective

[Devlin et al., 2018]

- How to encode information from both **bidirectional** contexts?
- General Idea: **text reconstruction!**
  - Your time is limited so don't waste it living someone else's life. Don't be trapped by dogma which is living with the results of other people's thinking. – Steve Jobs



$$h_1, \ldots, h_T = \text{Encoder}(w_1, \ldots, w_T)$$

$$y_i \sim Aw_i + b$$

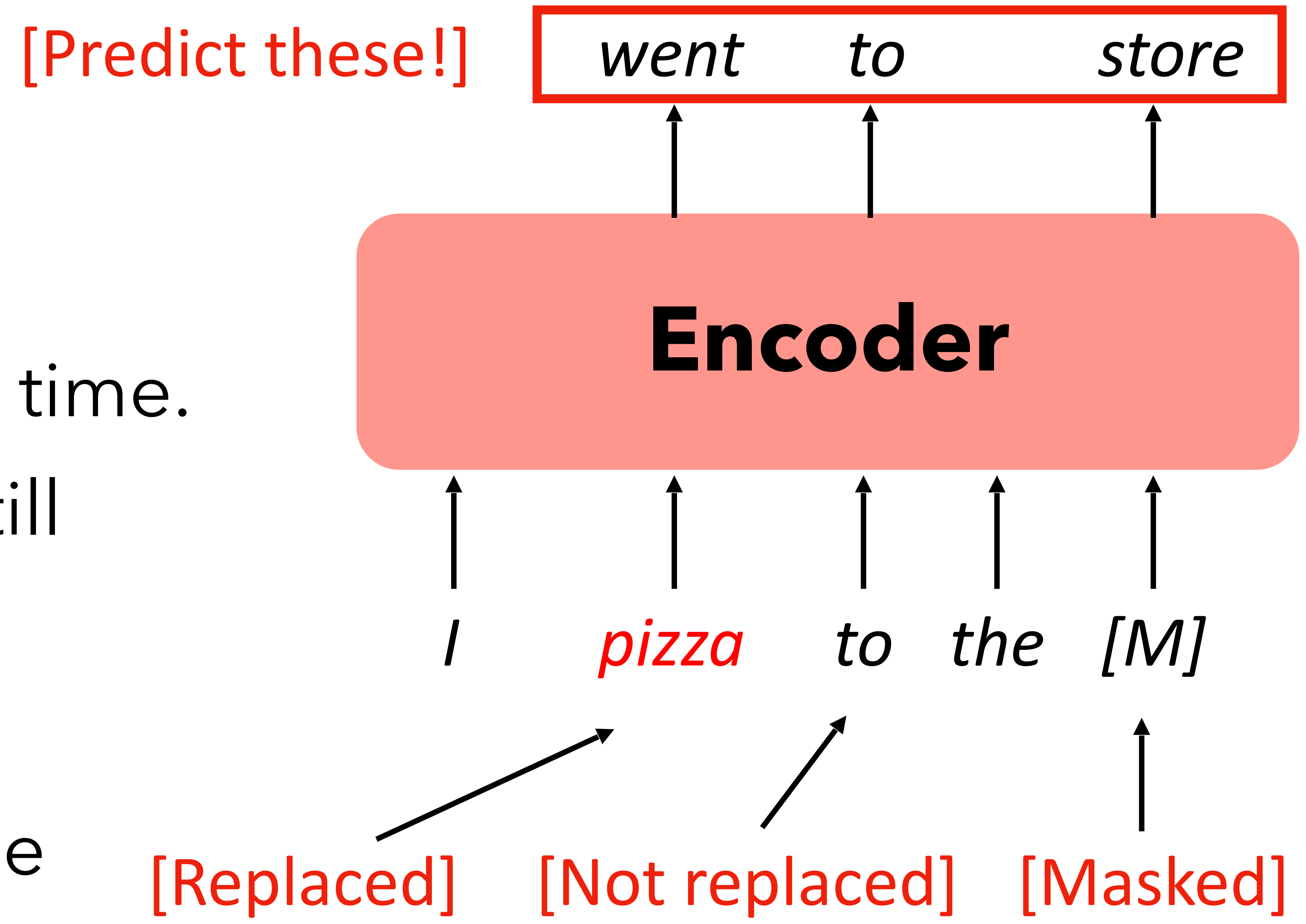
Only add loss terms from the masked tokens. If $\tilde{x}$ is the masked version of x, we're learning $p_\theta(x \mid \tilde{x})$. Called **Masked Language model (MLM)**.

Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

- **2 Pre-training Objectives:**
  - **Masked LM: Choose a random 15% of tokens to predict.**
    - For each chosen token:
      - Replace it with **[MASK]** 80% of the time.
      - Replace it with a **random token** 10% of the time.
      - Leave it **unchanged** 10% of the time (but still predict it!).
  - **Next Sentence Prediction (NSP)**
    - 50% of the time two adjacent sentences are in the correct order.
    - **This actually hurts model learning based on later work!**



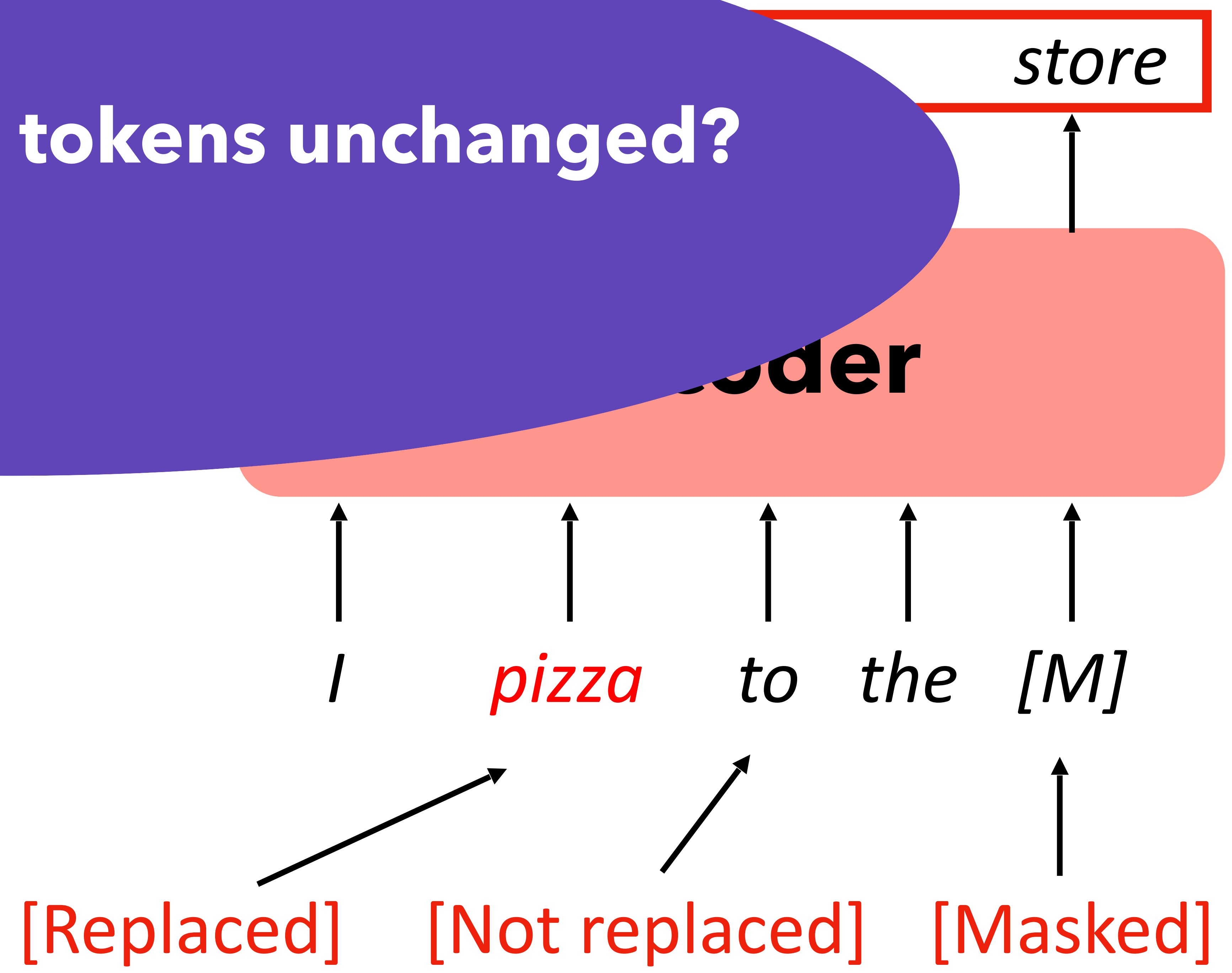
Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

- **2 Pre-training Objectives:**
  - **Masked LM: Choose a random 15% of tok**
    predict.
    - For each chosen token:
      - Replace it with **[MASK]**
      - Replace it with a **random tok**
      - Leave it **unchanged** 10% of the time (but still predict it!).
  - **Next Sentence Prediction (NSP)**
    - 50% of the time two adjacent sentences are in the correct order.
    - **This actually hurts model learning based on later work!**

WHY keeping some tokens unchanged?



Encoder: BERT

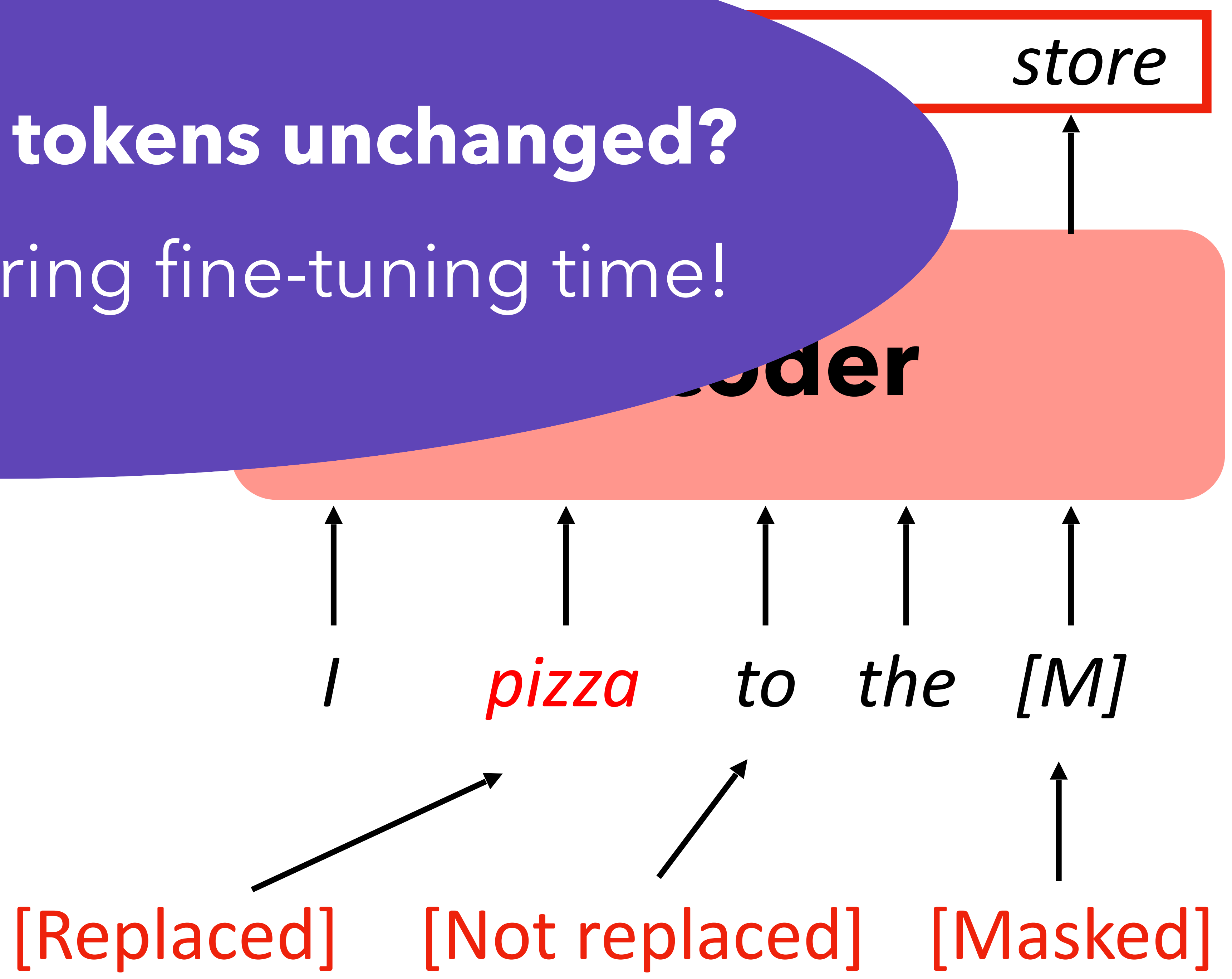
Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

- **2 Pre-training Objectives:**
  - **Masked LM: Choose a random 15% of tok**
    predict.
    - For each chosen token:
      - Replace it with **[MASK]**
      - Replace it with a **random tol**
      - Leave it **unchanged** 1 of the time (but still predict it!).
  - **Next Sentence Prediction (NSP)**
    - 50% of the time two adjacent sentences are in the correct order.
    - **This actually hurts model learning based on later work!**

WHY keeping some tokens unchanged?

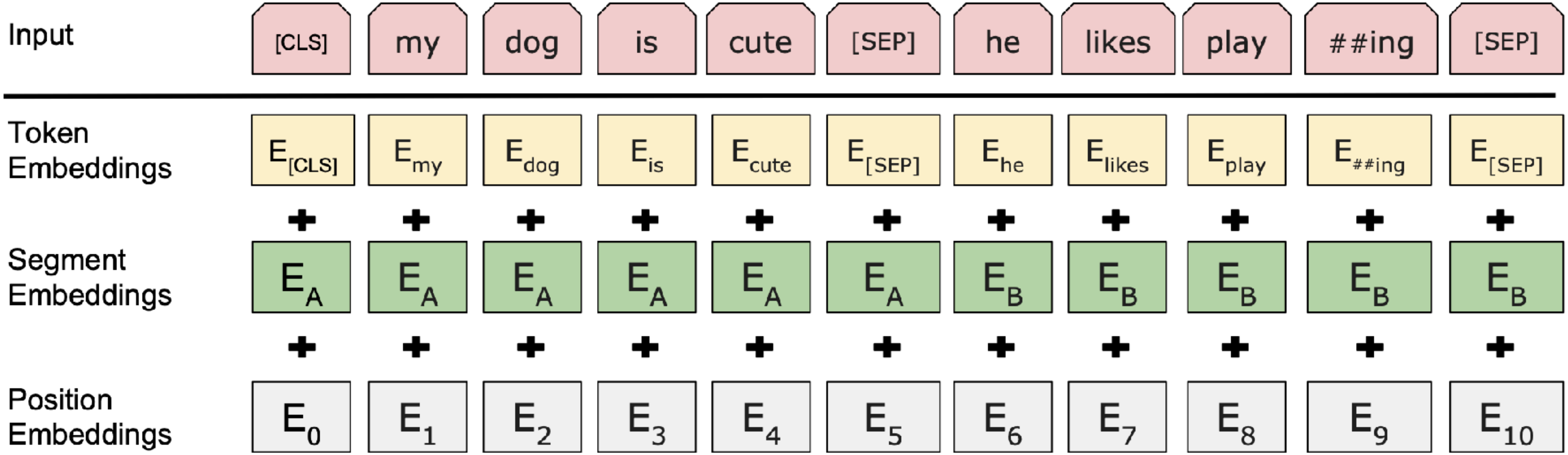
There's no [MASK] during fine-tuning time!



Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

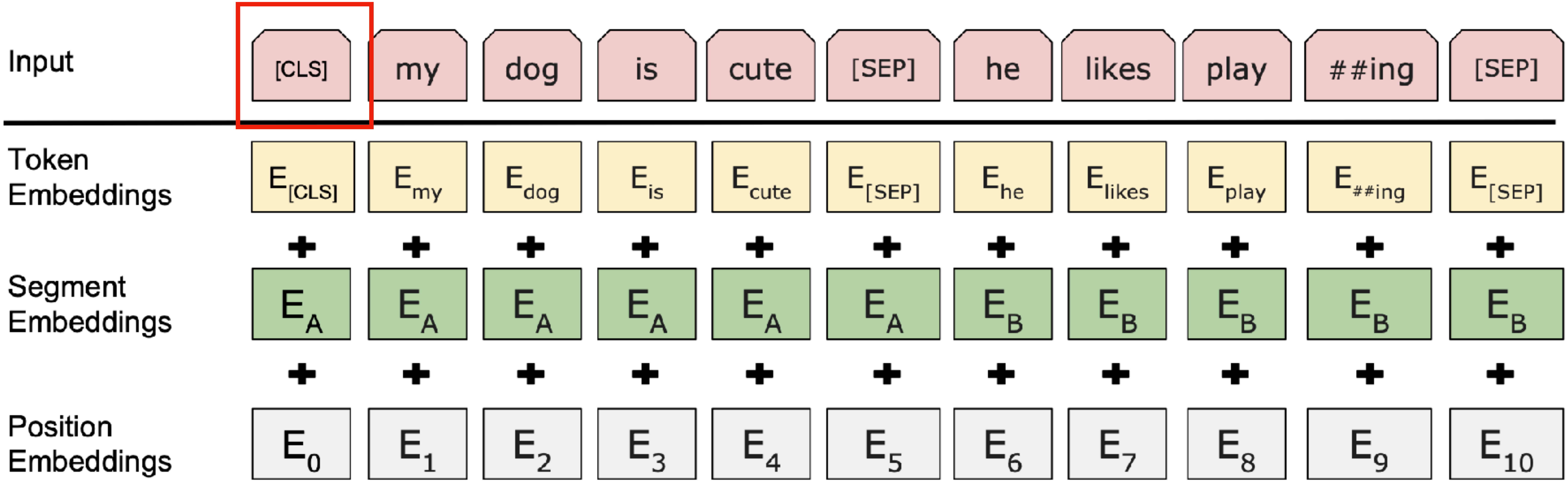


Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

Special token added to the beginning of each input sequence



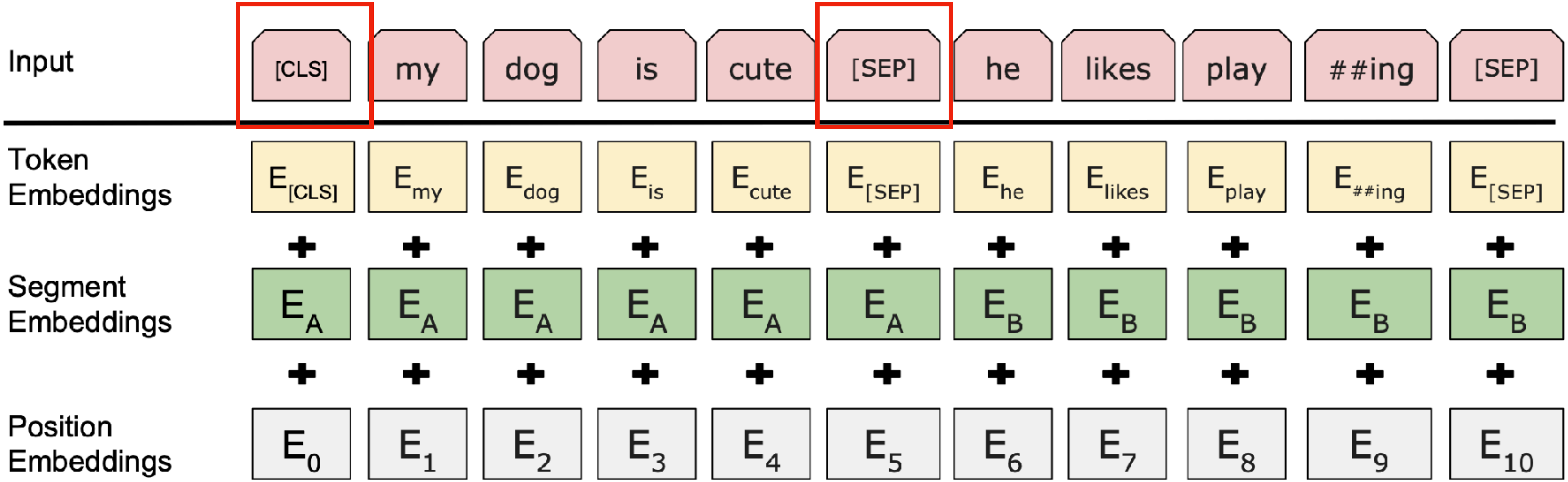
Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

Special token added to the beginning of each input sequence

Special token to separate sentence A/B



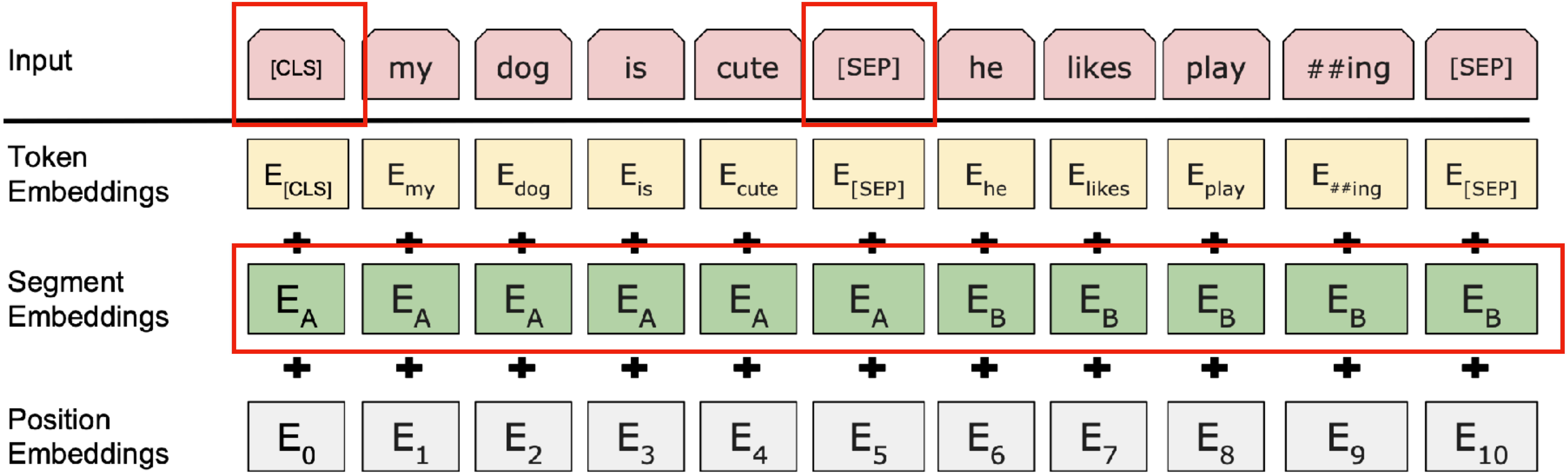
Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

Special token added to the beginning of each input sequence

Special token to separate sentence A/B



Learned embedding to every token indicating whether it belongs to sentence A or sentence B

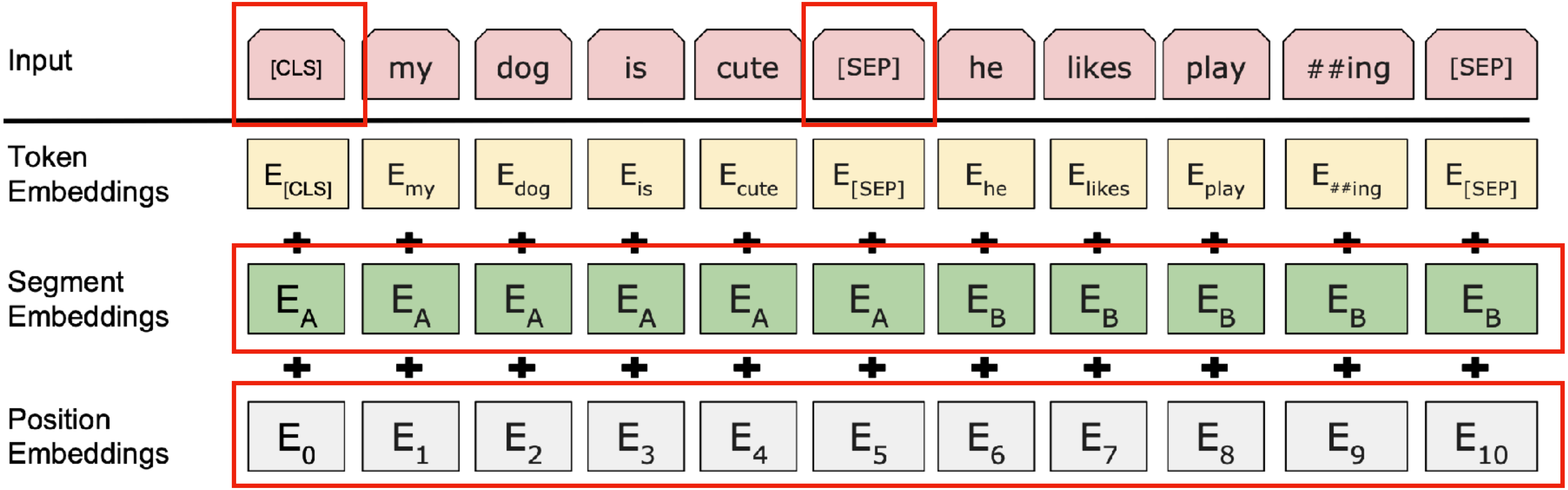
Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

Special token added to the beginning of each input sequence

Special token to separate sentence A/B



Learned embedding to every token indicating whether it belongs to sentence A or sentence B

Position of the token in the entire sequence

Encoder: BERT

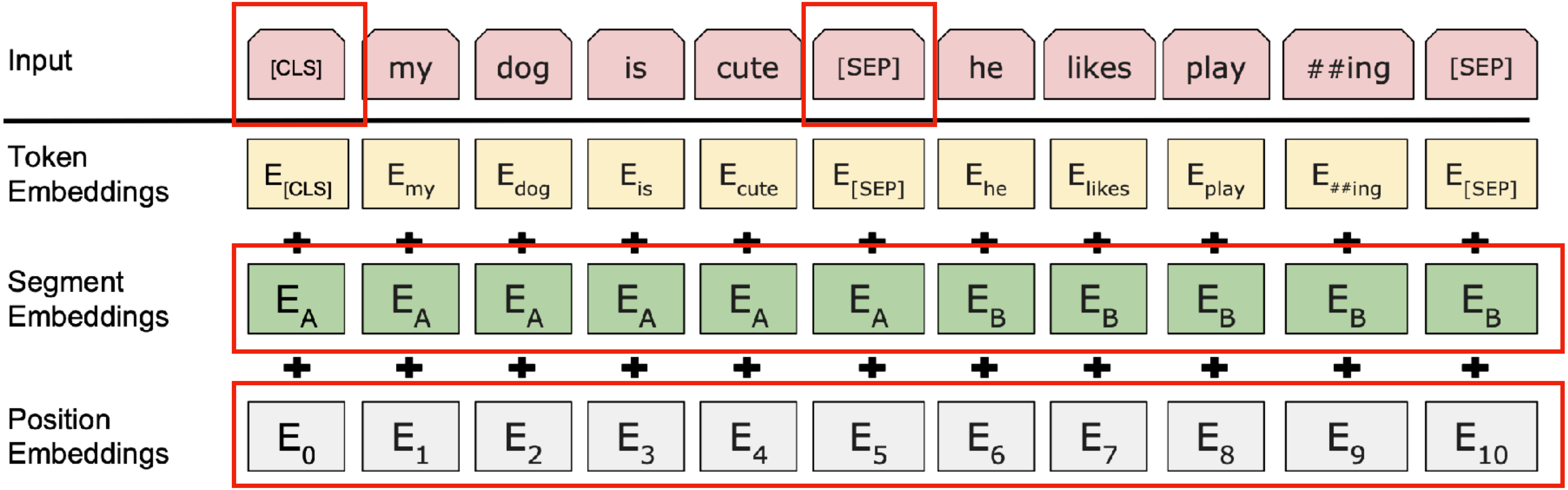
Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

Special token added to the beginning of each input sequence

Special token to separate sentence A/B

Final embedding is the sum of all three!



Learned embedding to every token indicating whether it belongs to sentence A or sentence B

Position of the token in the entire sequence

Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

- **SOTA at the time on a wide range of tasks after fine-tuning!**

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

- **QQP:** Quora Question Pairs (detect paraphrase questions)
- **QNLI:** natural language inference over question answering data
- **SST-2:** sentiment analysis
- **CoLA:** corpus of linguistic acceptability (detect whether sentences are grammatical.)
- **STS-B:** semantic textual similarity
- **MRPC:** microsoft paraphrase corpus
- **RTE:** a small natural language inference corpus

Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

[Devlin et al., 2018]

SWAG

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

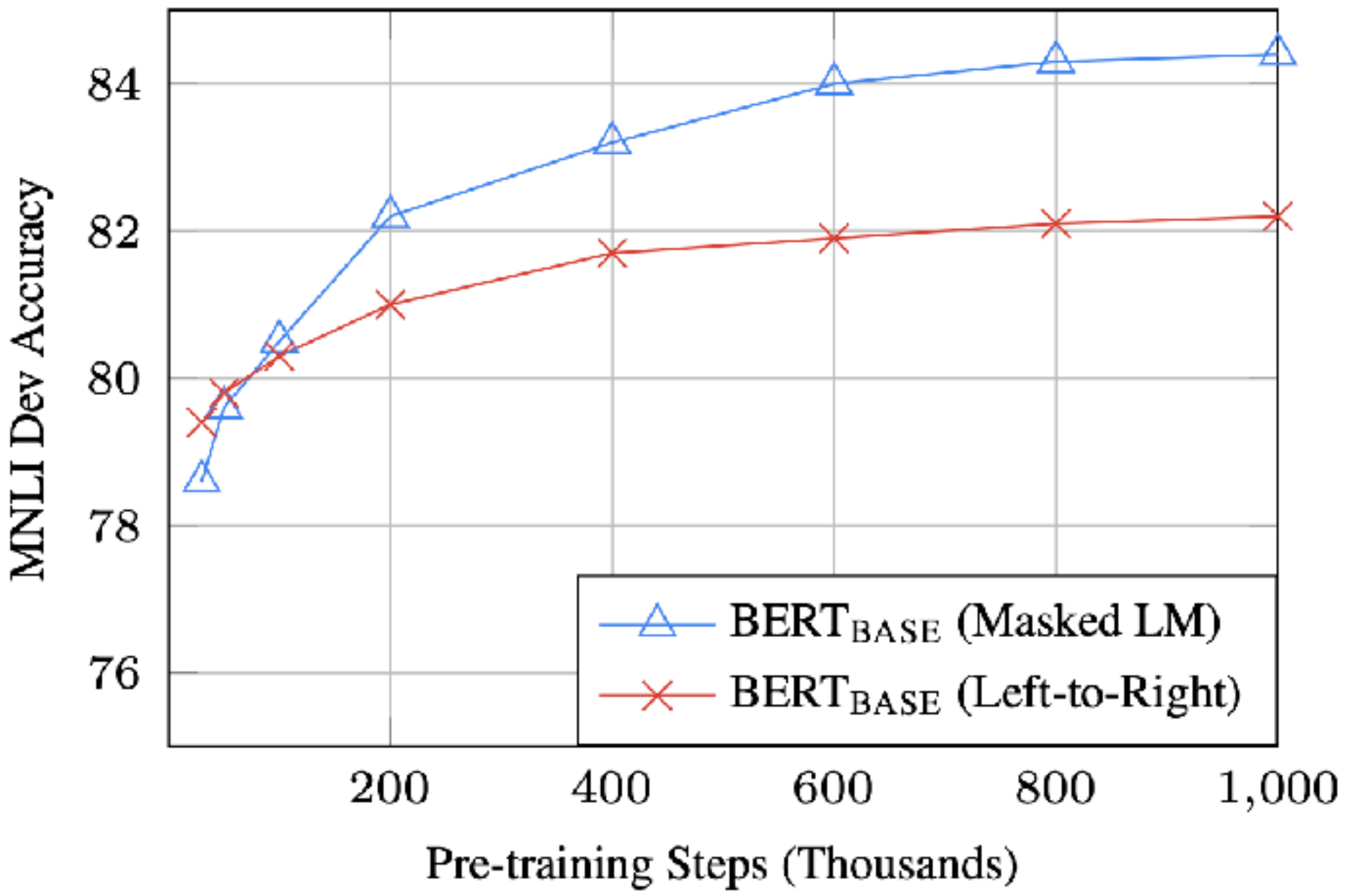
- **Two Sizes of Models**
  - **Base:** 110M, 4 Cloud TPUs, 4 days
  - **Large:** 340M, 16 Cloud TPUs, 4 days
  - Both models can be fine-tuned with single GPU
  - The larger the better!

Encoder: BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers [Devlin et al., 2018]

SWAG

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



- **Two Sizes of Models**
  - **Base:** 110M, 4 Cloud TPUs, 4 days
  - **Large:** 340M, 16 Cloud TPUs, 4 days
  - Both models can be fine-tuned with single GPU
  - The larger the better!
- MLM converges slower than Left-to-Right at the beginning, but out-performers it eventually

Encoder: RoBERTa

[Liu et al., 2019]

- **Original BERT is significantly undertrained!**
- More data (16G => 160G)
- Pre-train for longer
- Bigger batches
- Removing the next sentence prediction (NSP) objective
- Training on longer sequences
- Dynamic masking, randomly masking out different tokens
- A larger byte-level BPE vocabulary containing 50K sub-word units

Encoder: RoBERTa

[Liu et al., 2019]

- **Original BERT is significantly undertrained!**
- More data (16G => 160G)
- Pre-train for longer
- Bigger batches
- Removing the next sentence prediction (NSP) objective
- Training on longer sequences
- Dynamic masking, randomly masking out different tokens
- A larger byte-level BPE vocabulary containing 50K sub-word units

All around better than BERT!

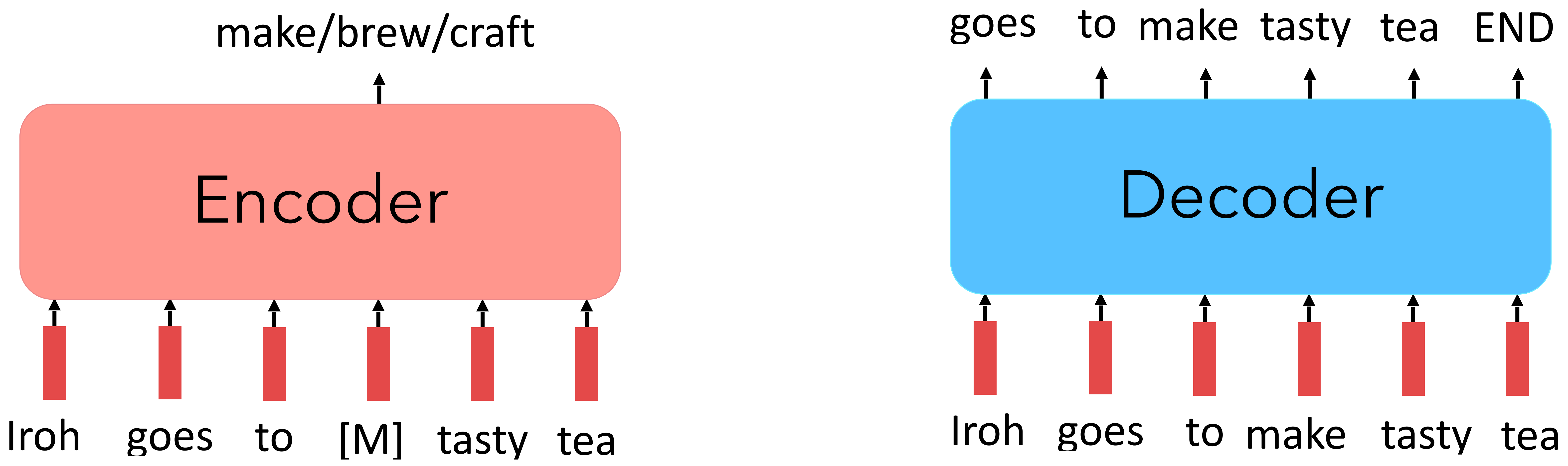
Encoder: Other Variations of BERT

- **ALBERT [Lan et al., 2020]**: incorporates two parameter reduction techniques that lift the major obstacles in scaling pre-trained models
- **DeBERTa [He et al., 2021]:** decoding-enhanced BERT with disentangled attention
- **SpanBERT [Joshi et al., 2019]:** masking contiguous spans of words makes a harder, more useful pre-training task
- **ELECTRA [Clark et al., 2020]:** corrupts texts by replacing some tokens with plausible alternatives sampled from a small generator network, then train a discriminative model that predicts whether each token in the corrupted input was replaced by a generator sample or not.
- **DistilBERT [Sanh et al., 2019]:** distilled version of BERT that's 40% smaller
- **TinyBERT [Jiao et al., 2019]:** distill BERT for both pre-training & fine-tuning
- **…**

Encoder: Pros & Cons

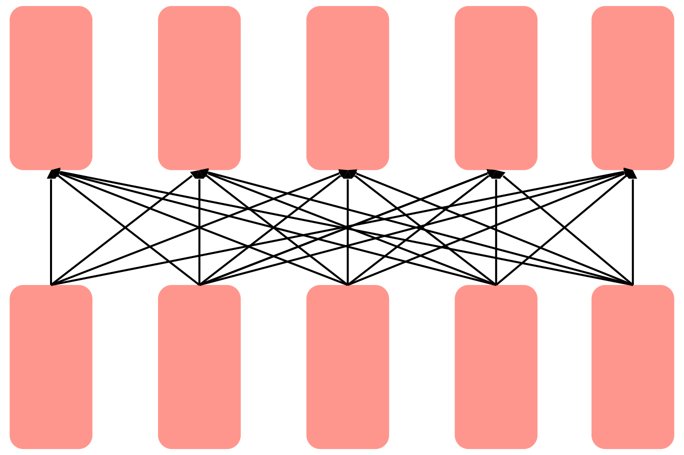
- Consider both left and right context
- Capture intricate contextual relationships

- Not good at generating open-text from left-to-right, one token at a time



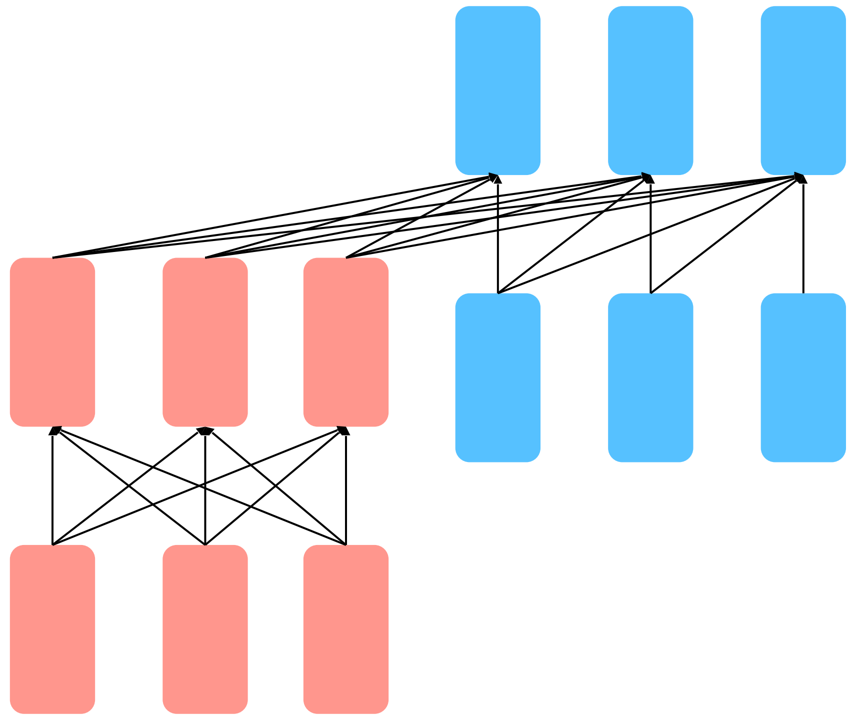
3 Pre-training Paradigms/Architectures

Encoder



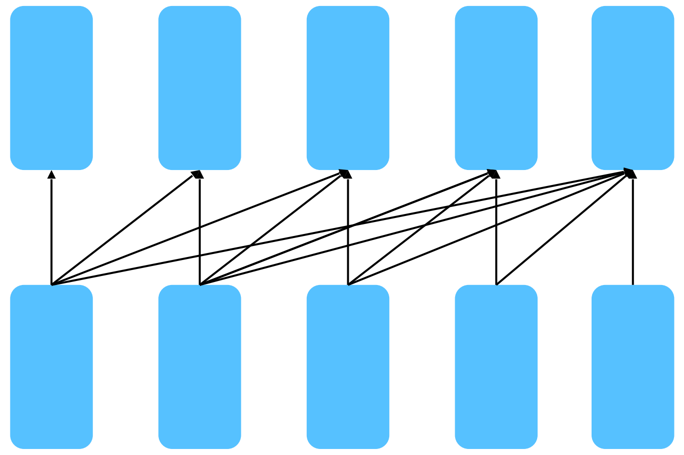
- Bidirectional; can condition on the future context

Encoder-Decoder



- Map two sequences of different length together

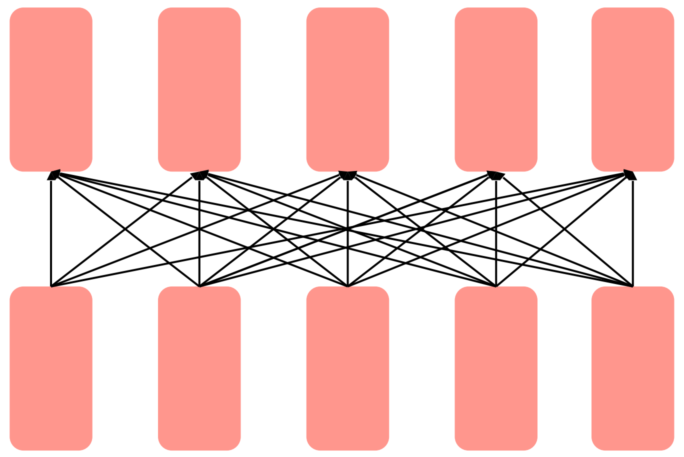
Decoder



- Language modeling; can only condition on the past context

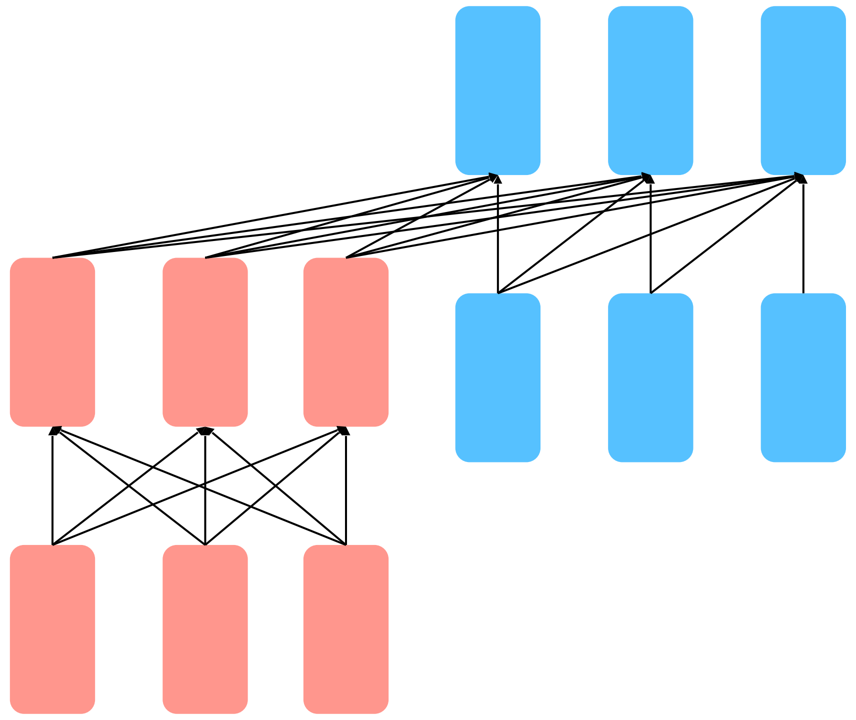
3 Pre-training Paradigms/Architectures

Encoder



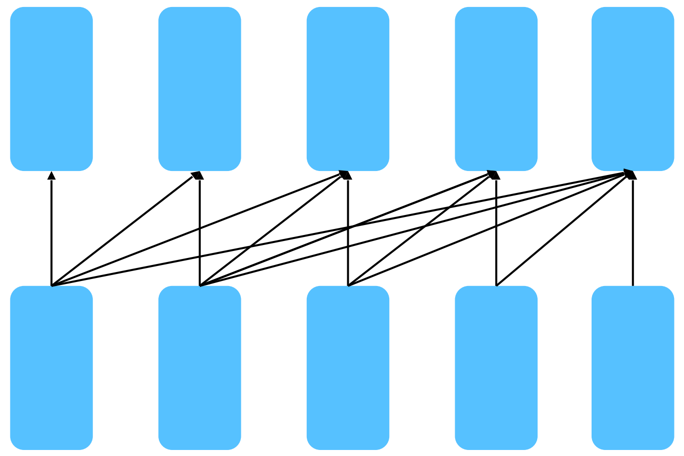
- Bidirectional; can condition on the future context

Encoder-Decoder



- Map two sequences of different length together

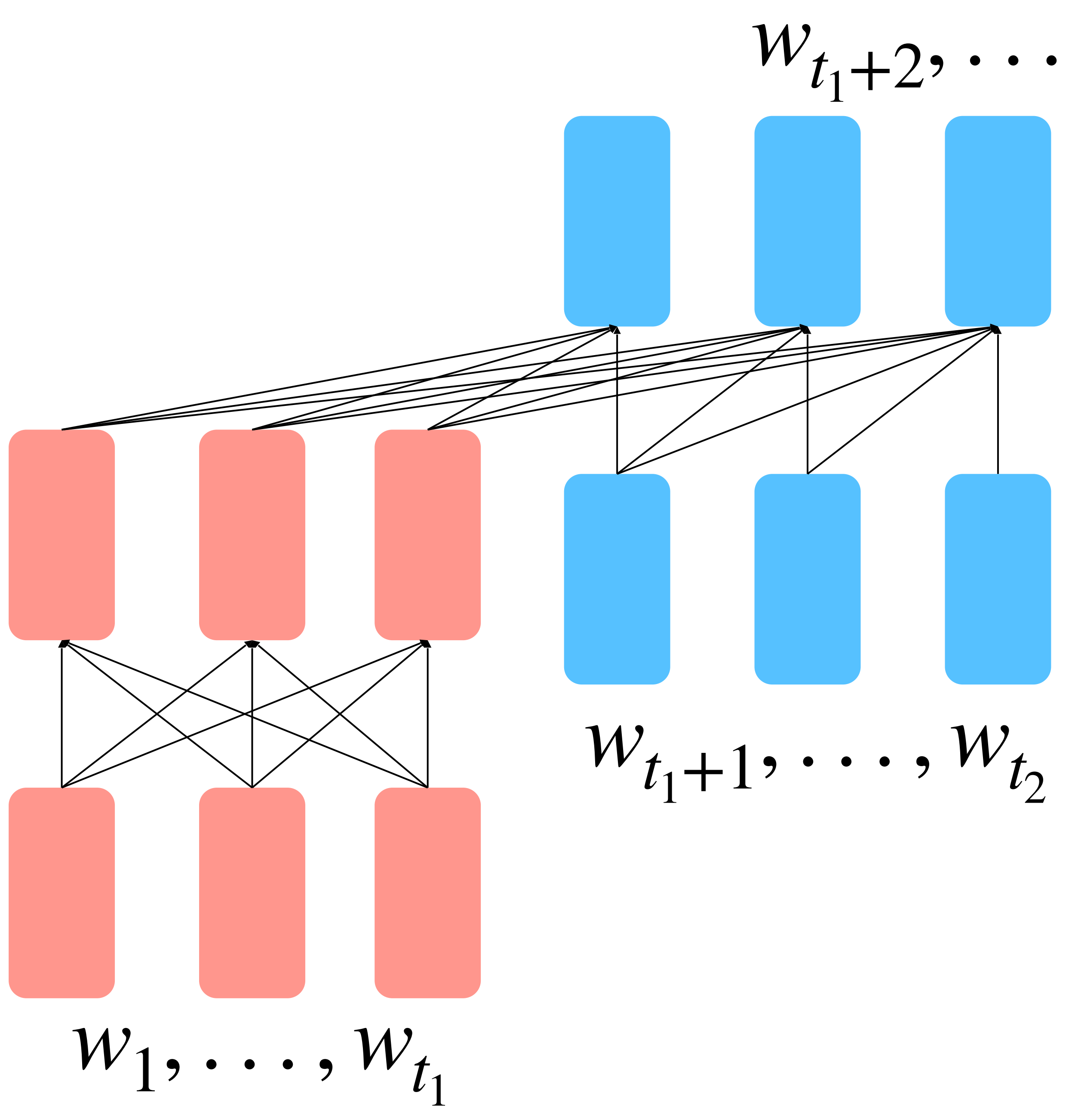
Decoder



- Language modeling; can only condition on the past context

Encoder-Decoder: Architecture

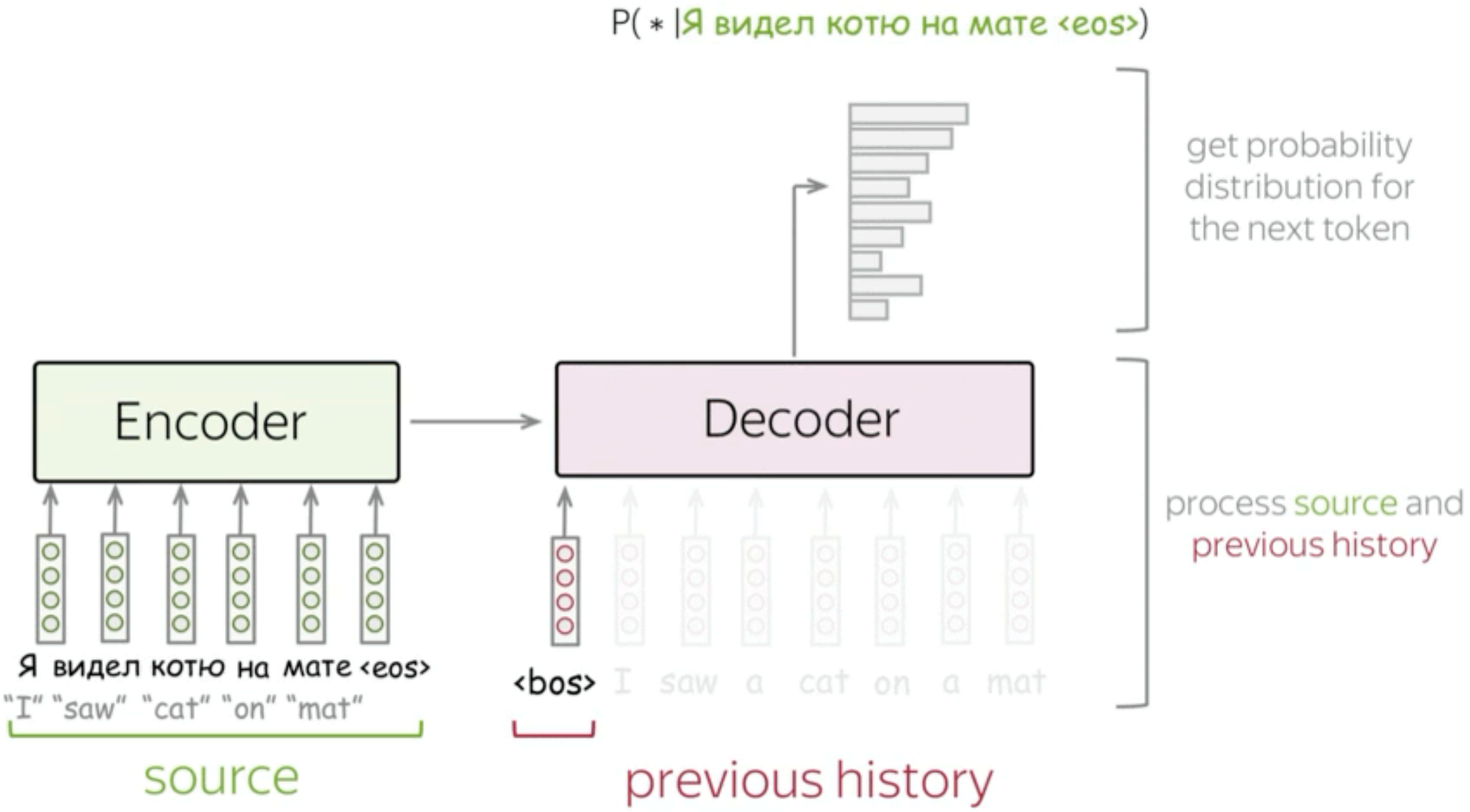
- Moving towards **open-text generation**…
- **Encoder** builds a representation of the source and gives it to the **decoder**
- **Decoder** uses the source representation to generate the target sentence
- The **encoder** portion benefits from **bidirectional** context; the **decoder** portion is used to train the whole model through **language modeling**



$$h_1, \ldots, h_{t_1} = \text{Encoder}(w_1, \ldots, w_{t_1})$$
$$h_{t_1+1}, \ldots, h_{t_2} = \text{Decoder}(w_{t_1+1}, \ldots, w_{t_2}, h_1, \ldots, h_{t_1})$$
$$y_i \sim Ah_i + b, i > t$$

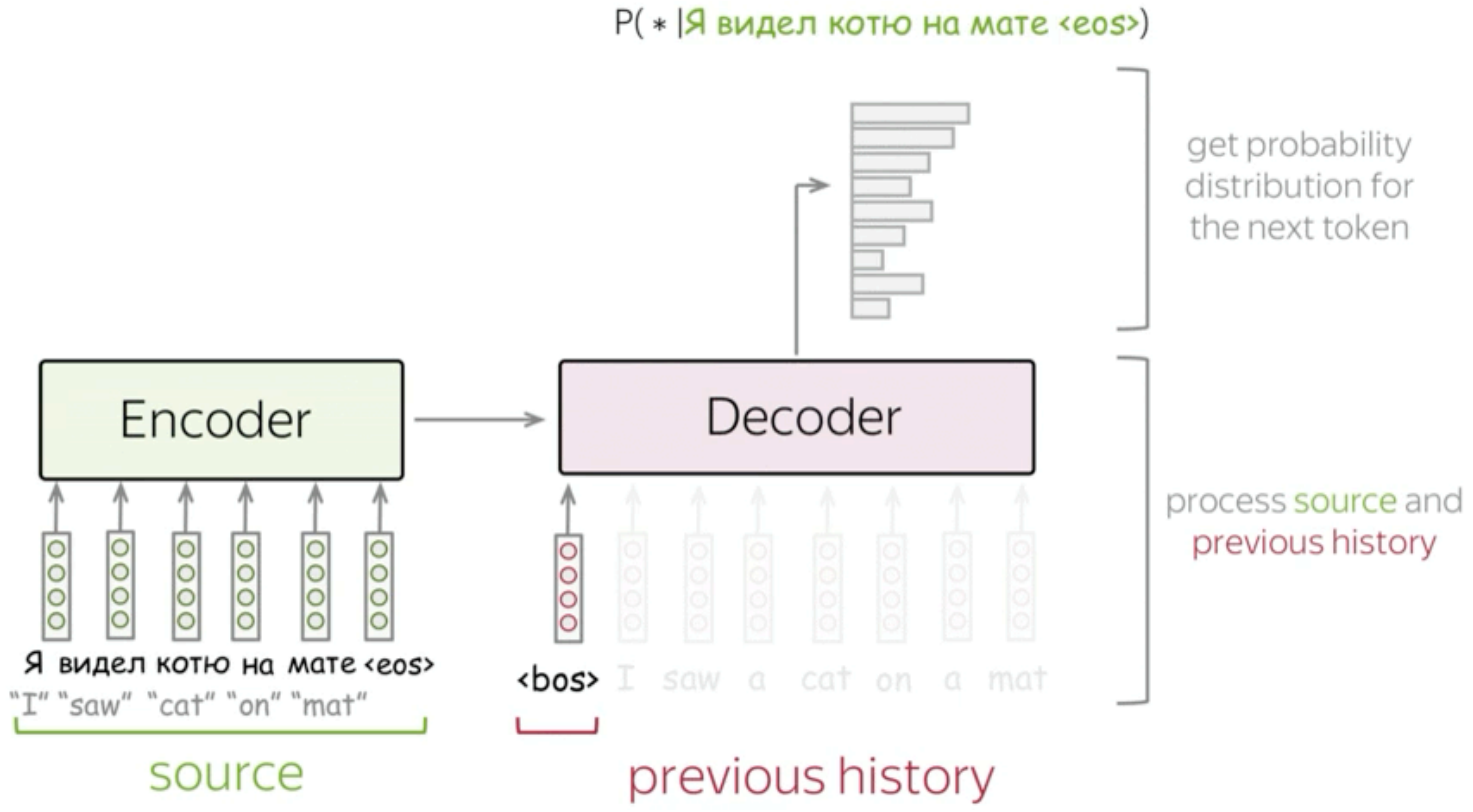
[Raffel et al., 2018]

Encoder-Decoder: An Machine Translation Example



[Lena Viota Blog]

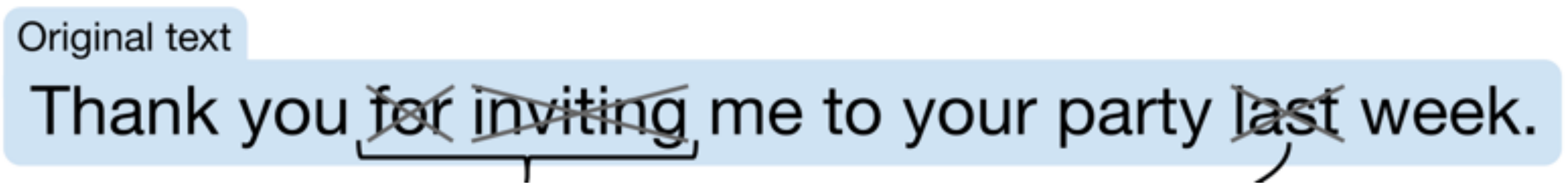
Encoder-Decoder: An Machine Translation Example

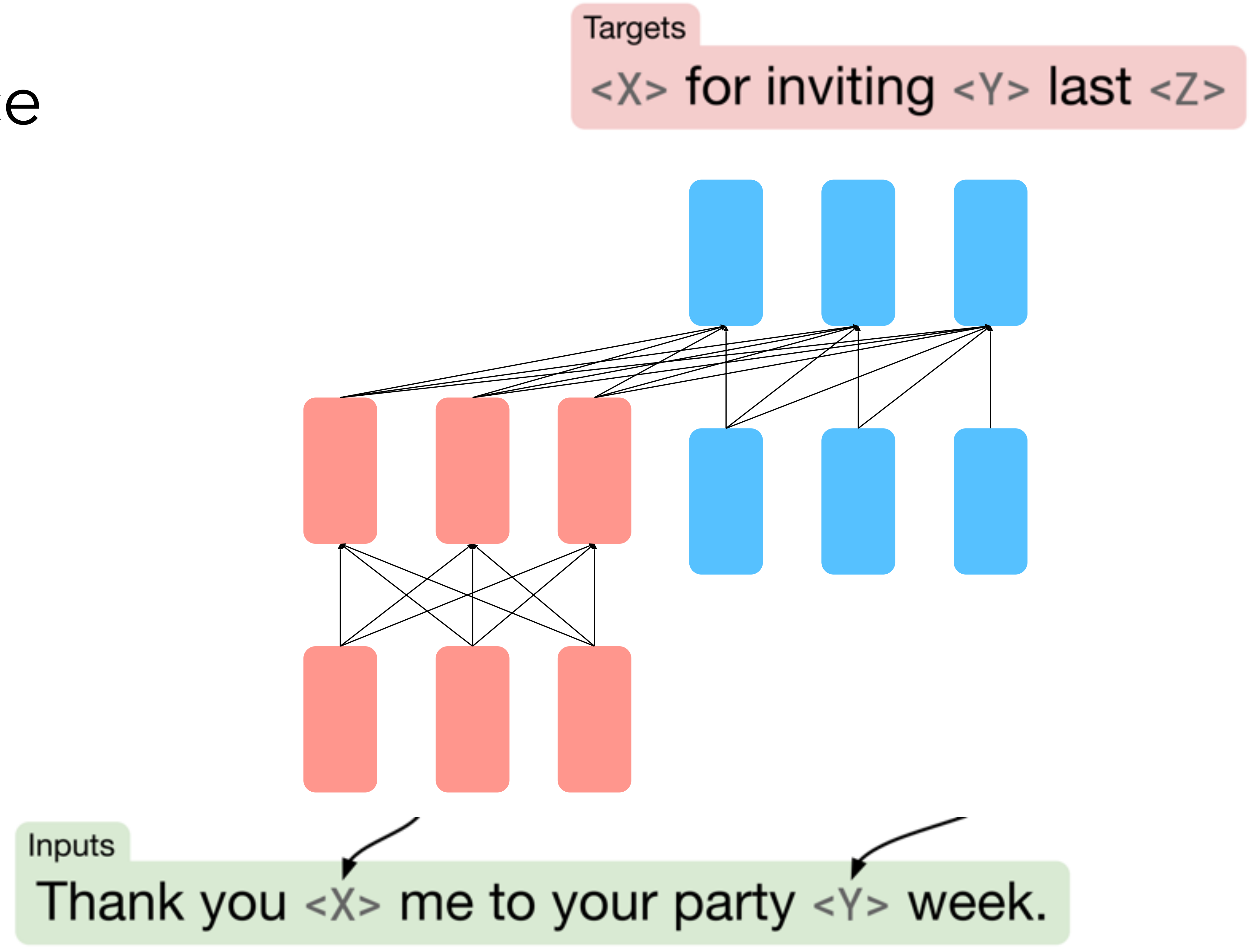


[Lena Viota Blog]

Encoder-Decoder: Training Objective

- **T5 [Raffel et al., 2018]**
- **Text span corruption (denoising):** Replace different-length spans from the input with unique placeholders (e.g., <extra_id_0>); decode out the masked spans.
  - Done during **text preprocessing**: training uses **language modeling** objective at the decoder side





Encoder-Decoder: T5

[Raffel et al., 2018]

- **Encoder-decoders** works better than decoders
- **Span corruption (denoising)** objective works better than language modeling

Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39
Encoder-decoder	LM	$2P$	M	79.56	18.59	76.02	64.29	26.27	39.17	26.86
Enc-dec, shared	LM	P	M	79.60	18.13	76.35	63.50	26.62	39.17	27.05
Enc-dec, 6 layers	LM	P	$M/2$	78.67	18.26	75.32	64.06	26.13	38.42	26.89
Language model	LM	P	M	73.78	17.54	53.81	56.51	25.23	34.31	25.38
Prefix LM	LM	P	M	79.68	17.84	76.87	64.86	26.28	37.51	26.76

Encoder-Decoder: T5

[Raffel et al., 2018]

- **Text-to-Text:** convert NLP tasks into input/output text sequences
- **Dataset:** Colossal Clean Crawled Corpus (C4), 750G text data!
- **Various Sized Models:**
    - Base (222M)
    - Small (60M)
    - Large (770M)
    - 3B
    - 11B
- **Achieved SOTA with scaling & purity of data**

[Google Blog]

Encoder-Decoder: T5

[Raffel et al., 2018]

- **Text-to-Text:** convert NLP tasks into input/output text sequences
- **Dataset:** Colossal Clean Crawled Corpus (C4), 750G text data!
- **Various Sized Models:**
  - Base (222M)
  - Small (60M)
  - Large (770M)
  - 3B
  - 11B
- **Achieved SOTA with scaling & purity of data**

[Google Blog]

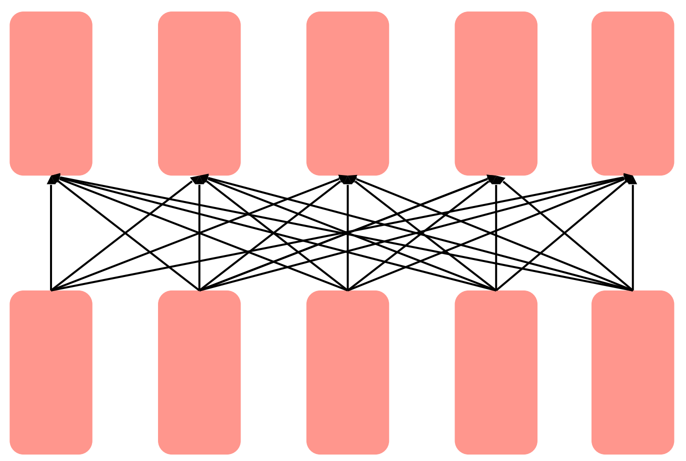
Encoder-Decoder: Pros & Cons

- A nice middle ground between leveraging **bidirectional** contexts and **open-text** generation
- Good for **multi-task** fine-tuning

- Require more **text wrangling**
- **Harder to train**
- **Less flexible** for natural language generation

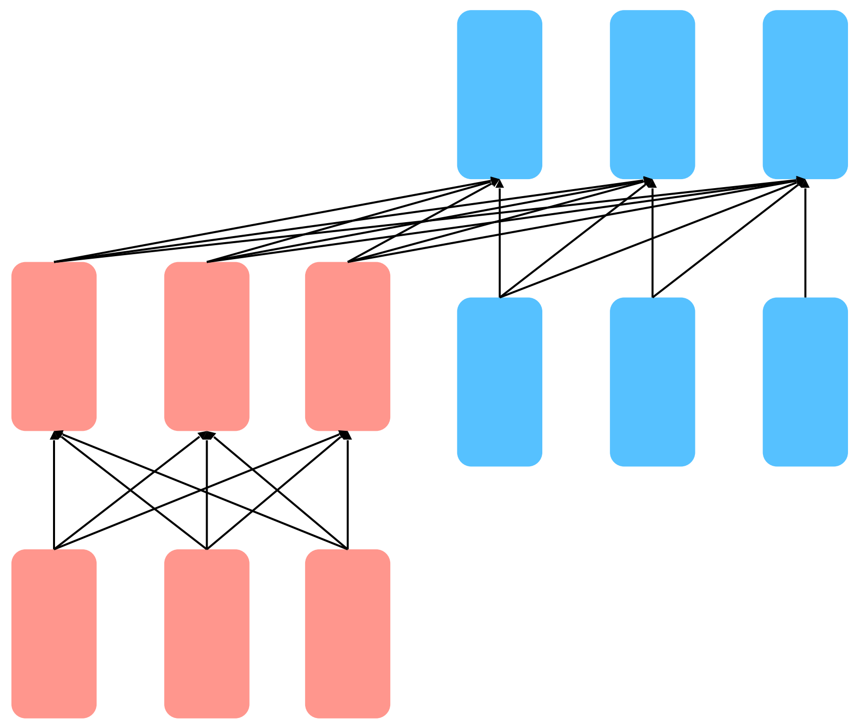
3 Pre-training Paradigms/Architectures

Encoder



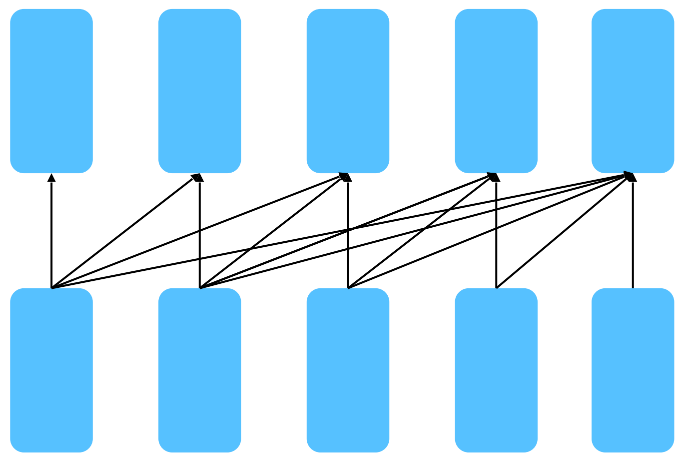
- Bidirectional; can condition on the future context

Encoder-Decoder



- Map two sequences of different length together

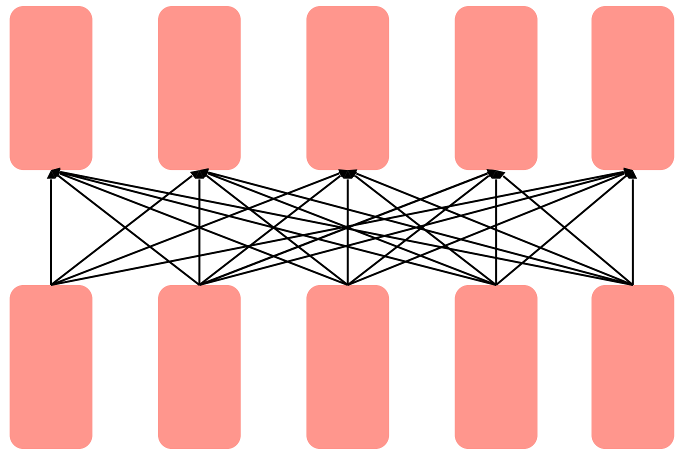
Decoder



- Language modeling; can only condition on the past context

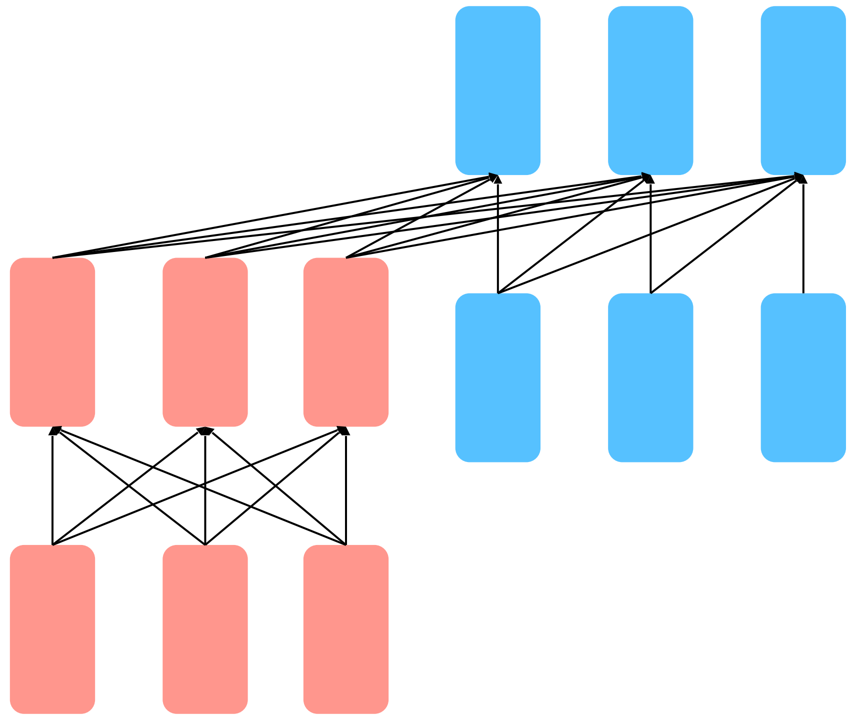
3 Pre-training Paradigms/Architectures

Encoder



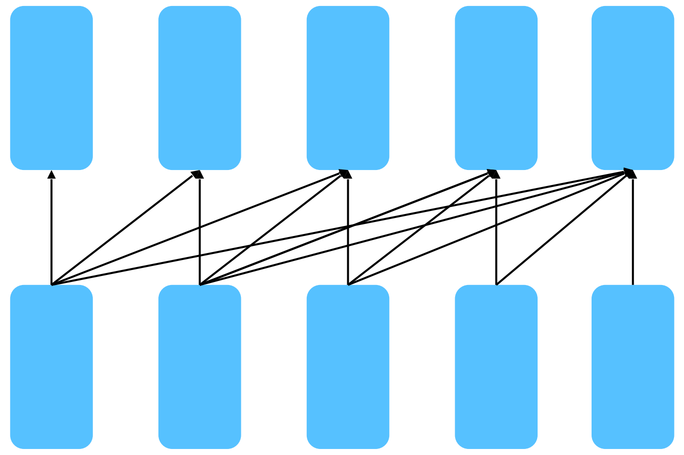
- Bidirectional; can condition on the future context

Encoder-Decoder



- Map two sequences of different length together

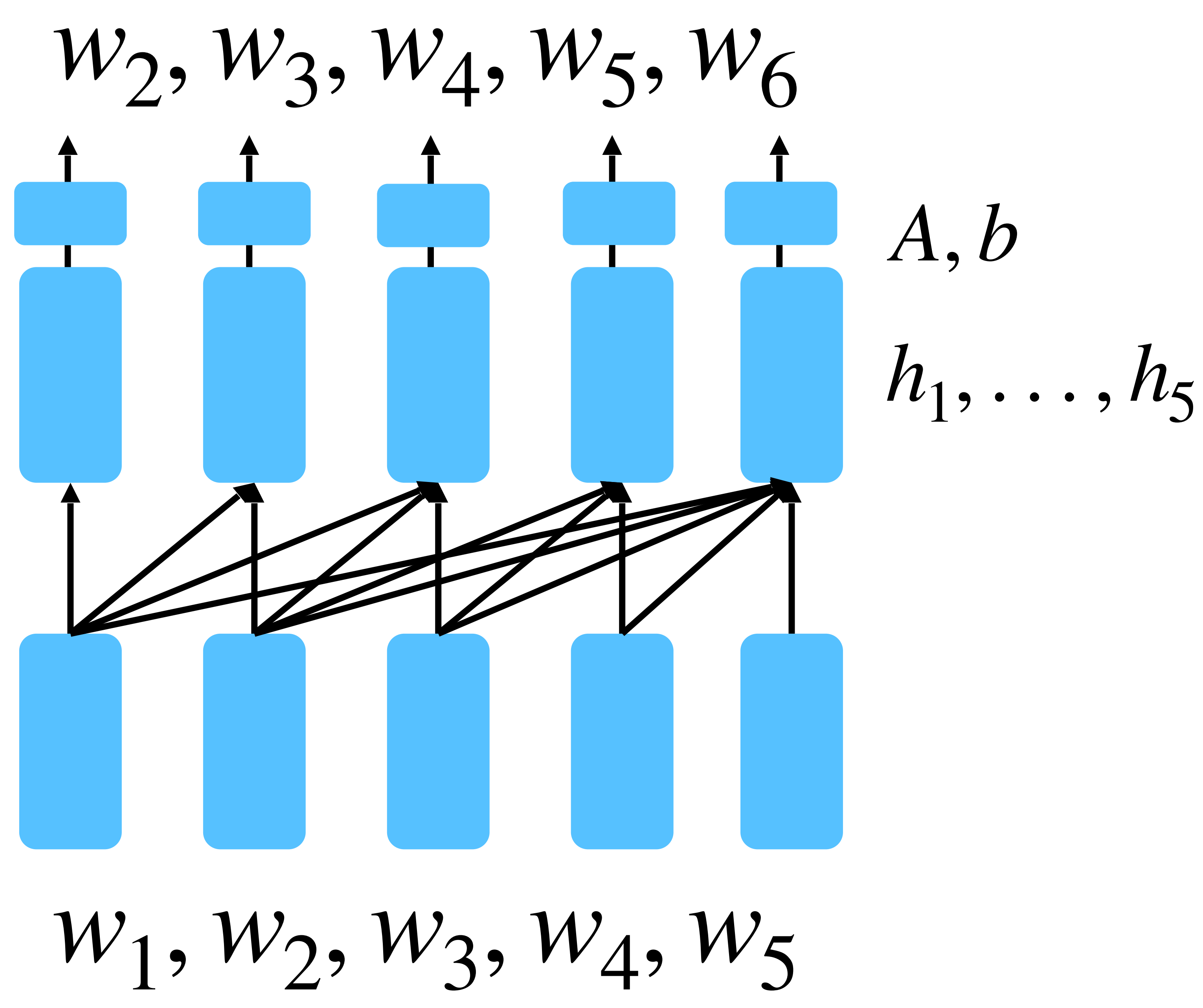
Decoder



- Language modeling; can only condition on the past context

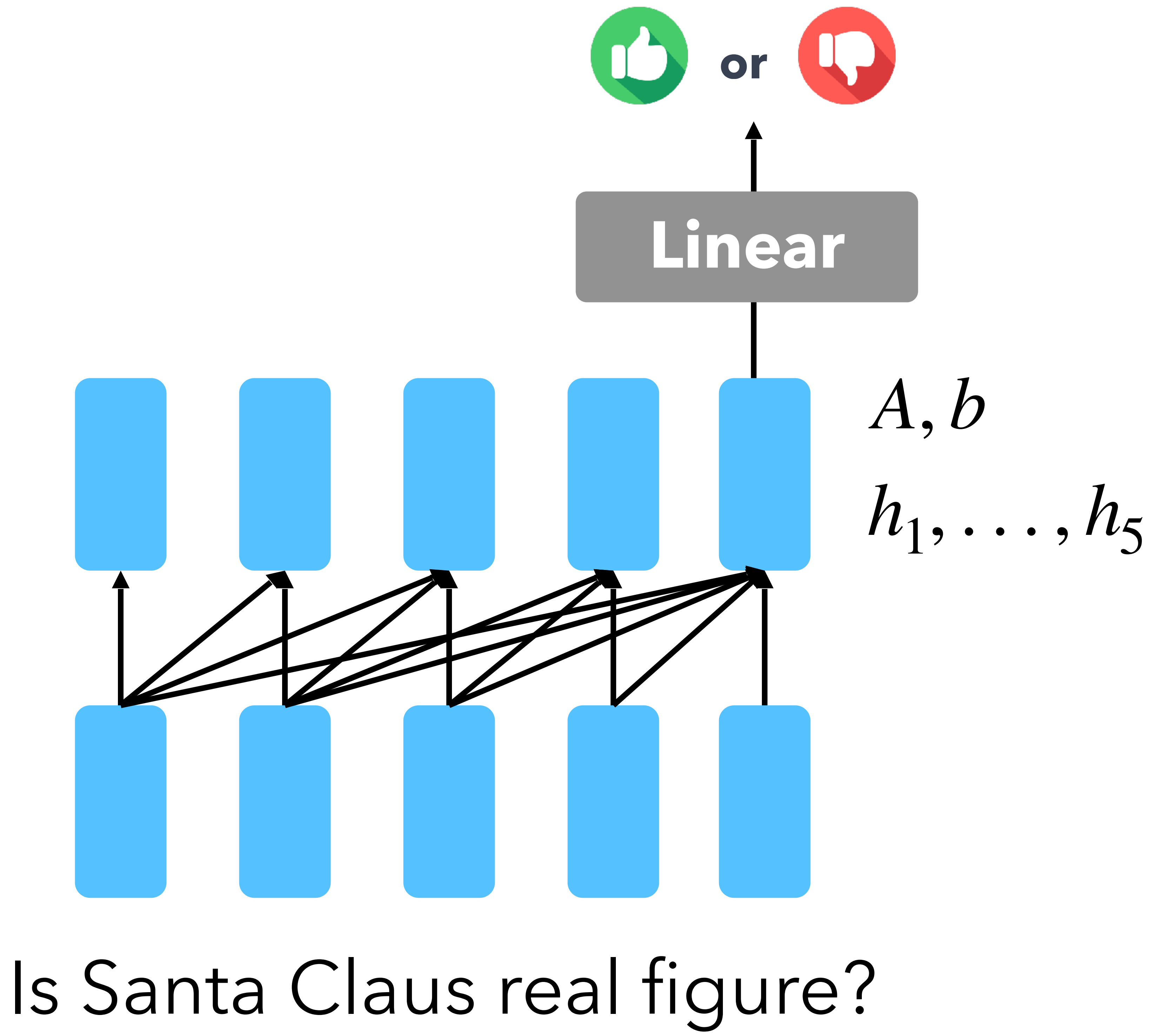
Decoder: Training Objective

- Many most famous generative LLMs are **decoder-only**
  - e.g., GPT1/2/3/4, Llama1/2
- **Language modeling!** Natural to be used for **open-text generation**
- **Conditional LM:** $p(w_t \mid w_1, \ldots, w_{t-1}, x)$
  - Conditioned on a source context x to generate from left-to-right
- Can be fine-tuned for **natural language generation (NLG)** tasks, e.g., dialogue, summarization.



Decoder: Training Objective

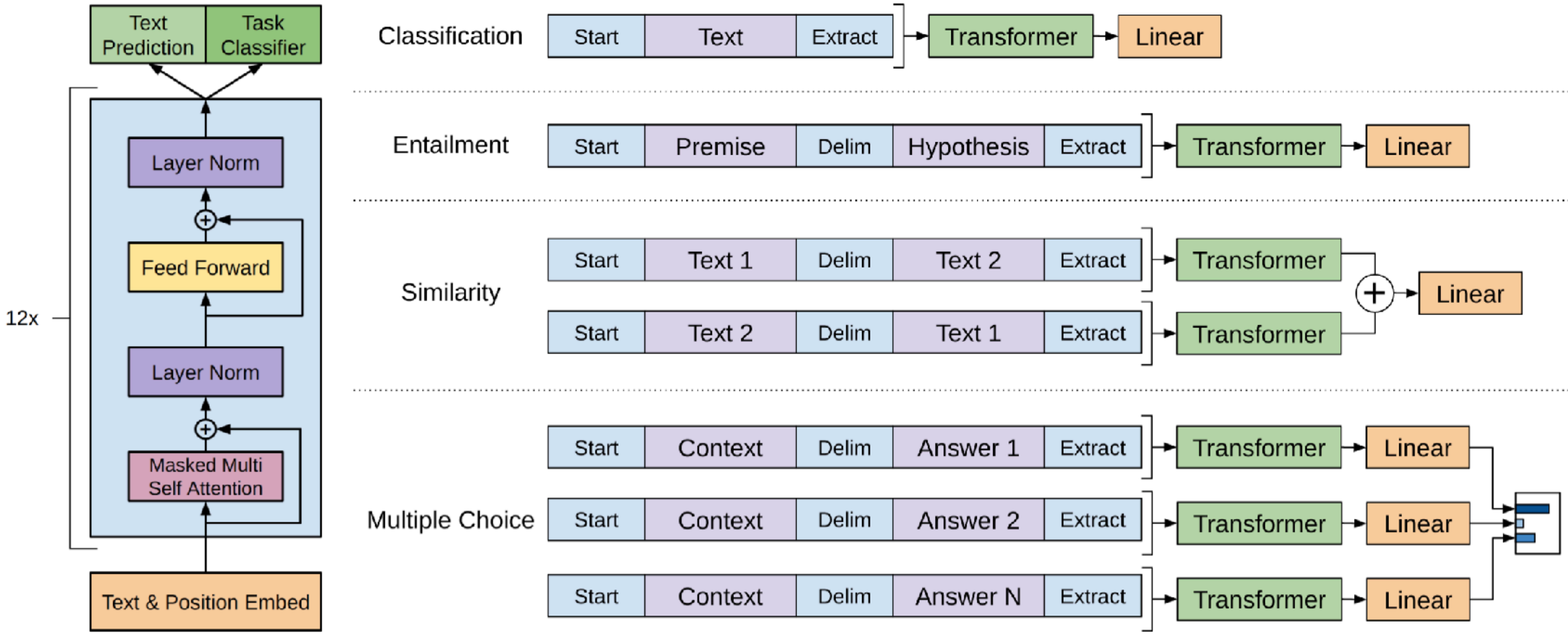
- Customizing the pre-trained model for downstream tasks:
  - Add a **linear layer** on top of the last hidden layer to make it a classifier!
  - During fine-tuning, trained the randomly **initialized linear layer**, along with **all parameters** in the neural net.



Decoder: GPT

Generative **P**re-trained **T**ransformer

[Radford et al., 2018]



How to pick a proper architecture for a given task?

- Right now **decoder-only** models seem to dominant the field at the moment
    - e.g., GPT1–5, Mistral, Llama1–3, Claude, etc.
- T5 (seq2seq) works well with multi-tasking
- **Picking the best model architecture remains an open research question!**