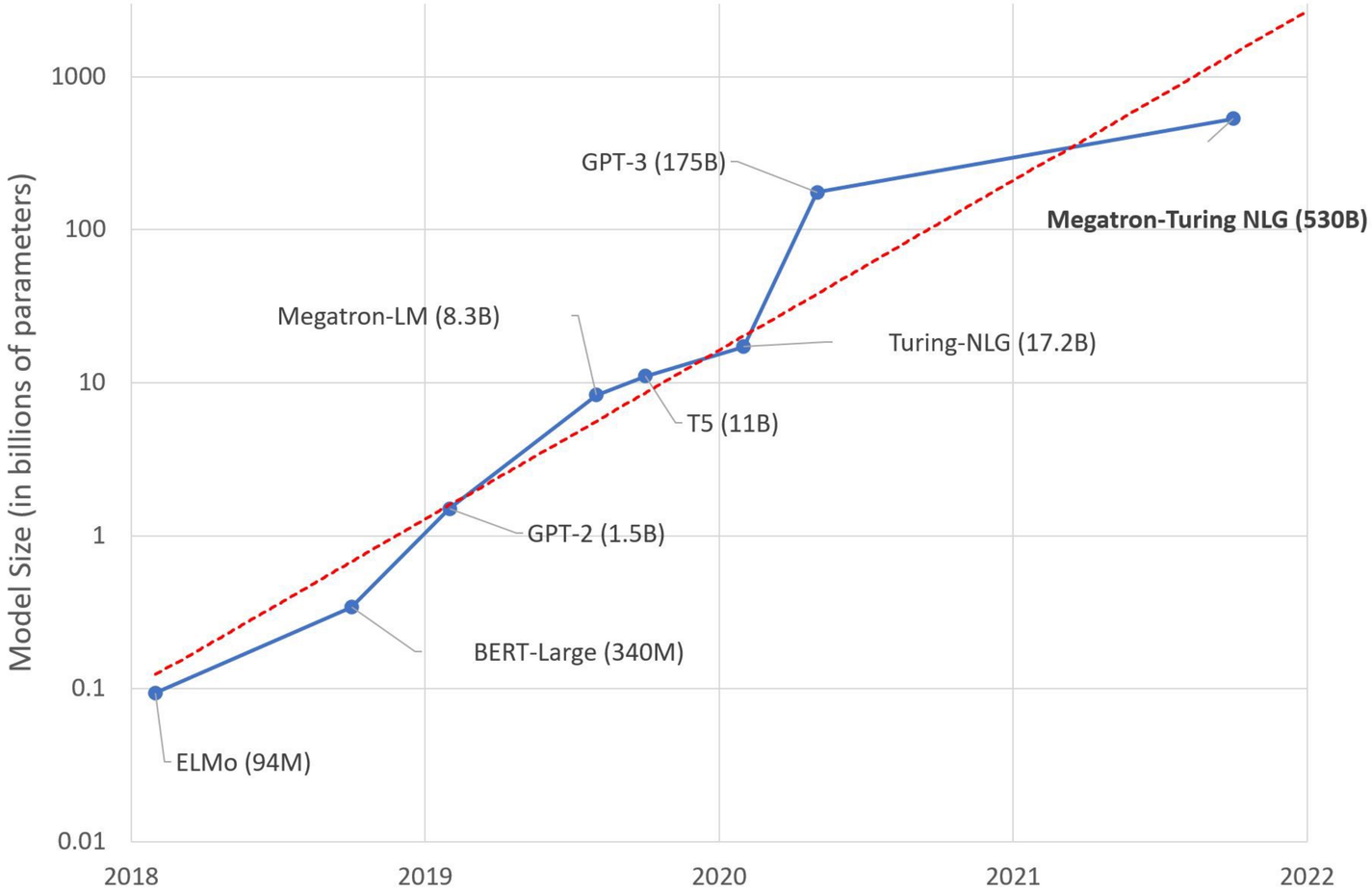
Large Pretrained Models

COMP3361 — Week 9

Lingpeng Kong

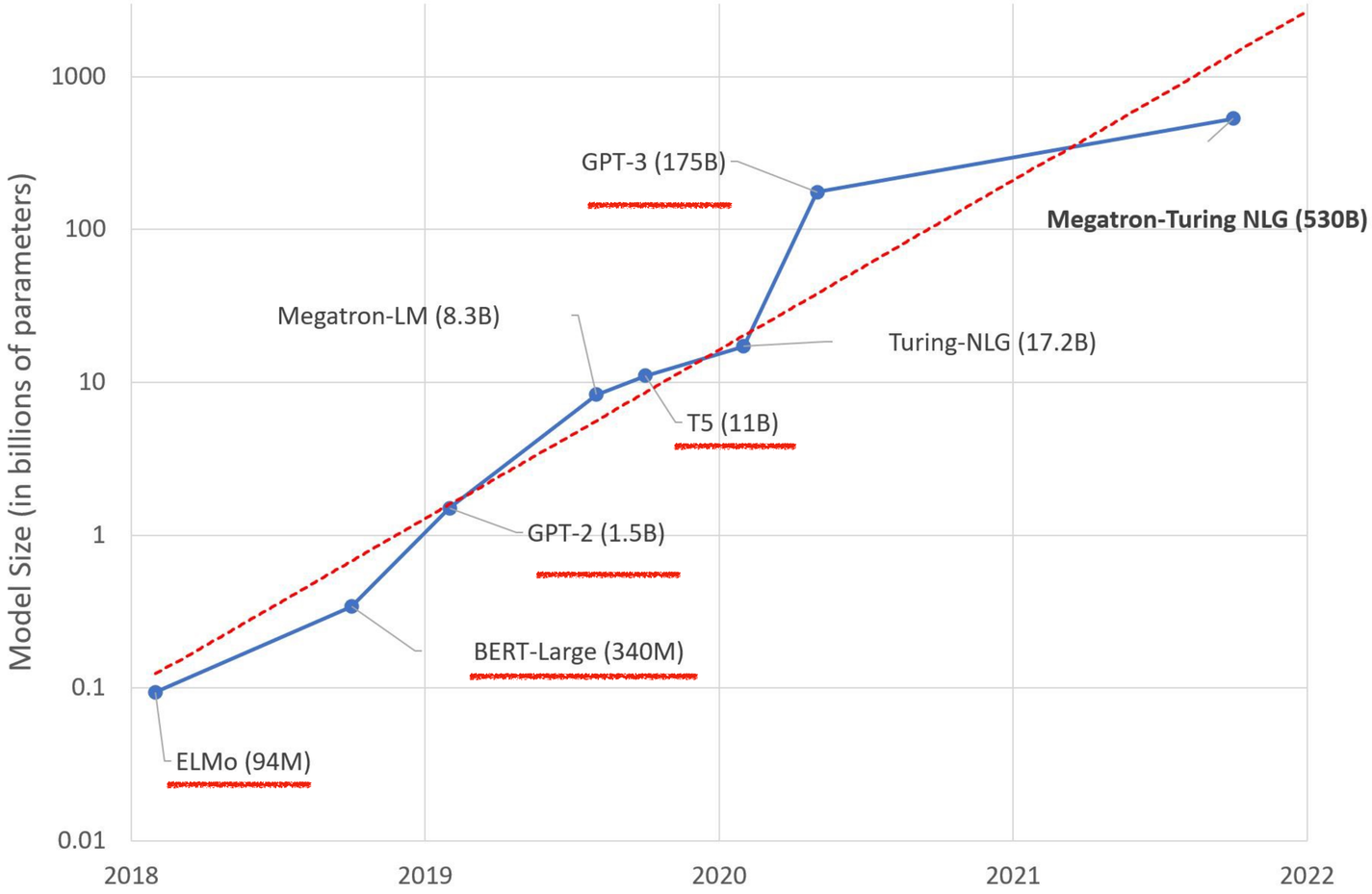
Department of Computer Science, The University of Hong Kong
Many materials from Stanford CS224n with special thanks!

Pretrained Models in the Past Four Years



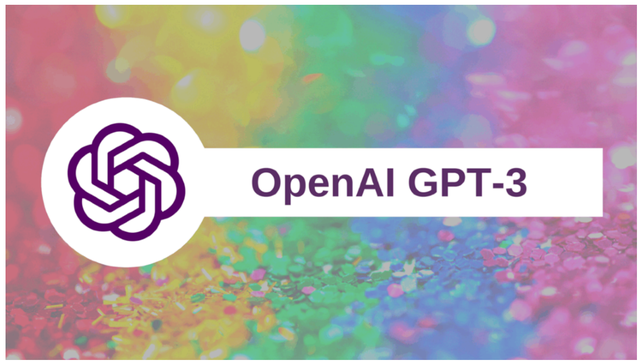
Microsoft Research Blog. Oct 6, 2021.

Pretrained Models in the Past Four Years



Microsoft Research Blog. Oct 6, 2021.

Pretrained Models are Expensive



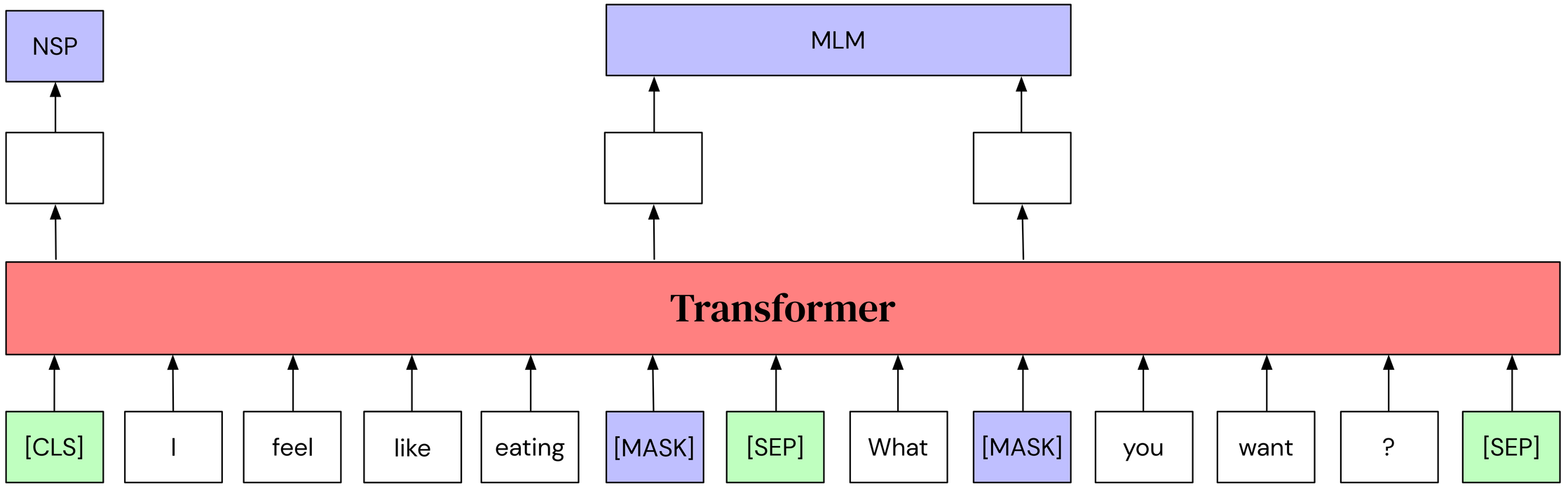
One single training run

$12 million

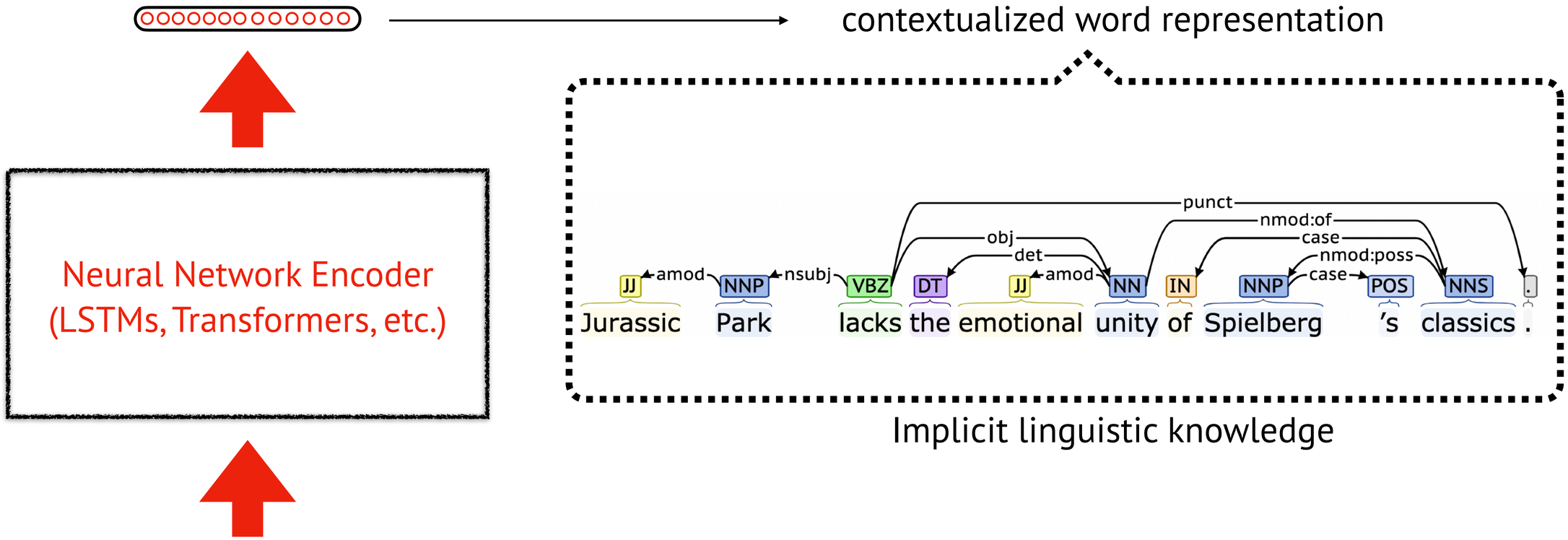
552 metric tons of carbon dioxide
(120 cars per year)

Pretraining and Contextualized Word Representations

$$\mathbb{E}_{p(x_i, \hat{\boldsymbol{x}}_i)}[p(x_i \mid \hat{\boldsymbol{x}}_i)]$$

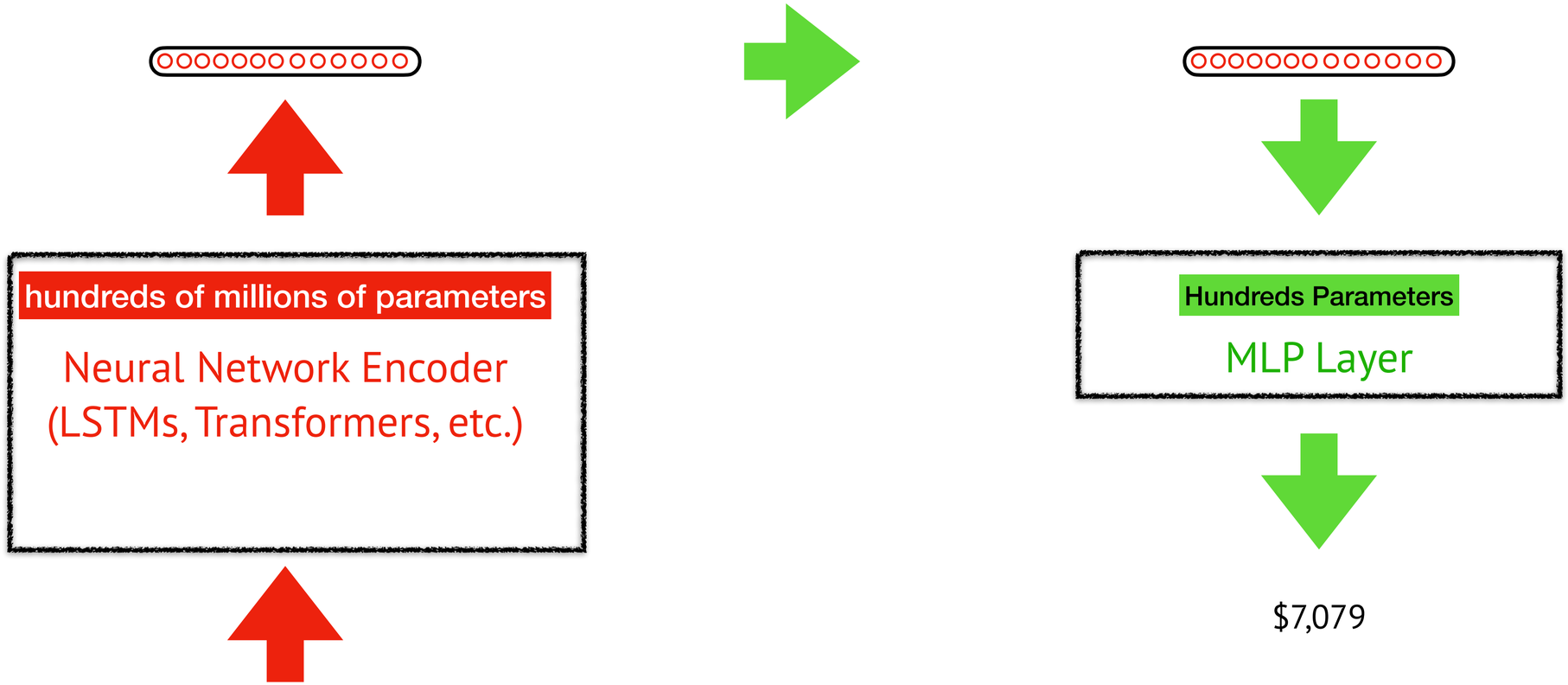


Pretraining and Contextualized Word Representations



Jurassic Park lacks the emotional unity of Spielberg's classics .

Pretraining and Fine-tuning

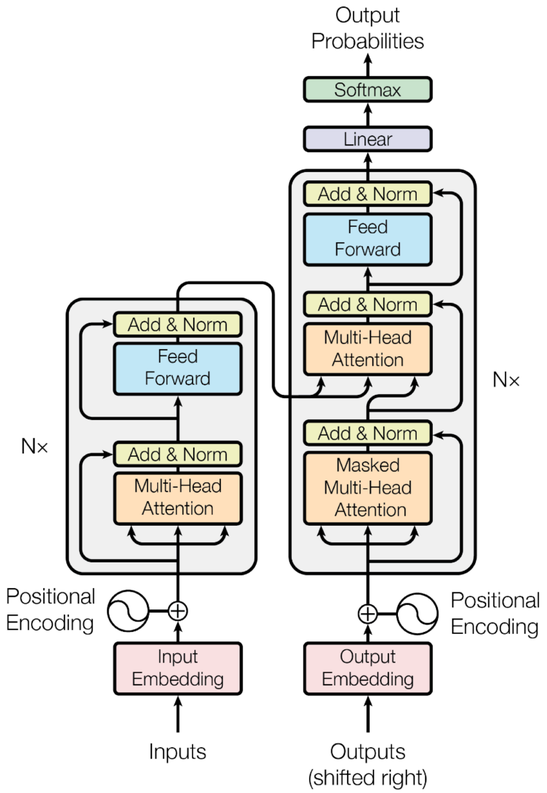


Jurassic Park lacks the emotional unity of Spielberg's classics .

Key Elements in BERT

Transformer	— neural representation learner
Masked Language Modeling (MLM), Next Sentence Prediction (NSP)	— pretraining objective
Bidirectional Encoder	— type of architecture

Neural Representation Learners



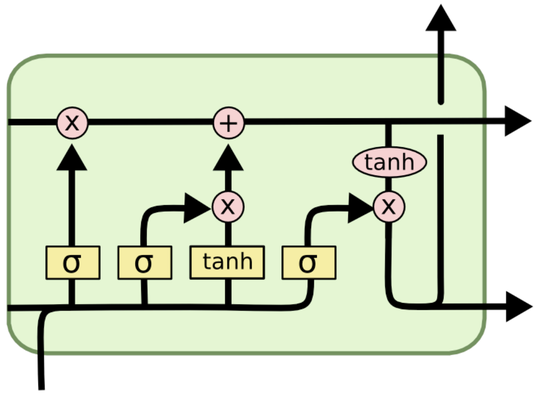
Transformer

BERT GPT-2

GPT-3 BART

T5 XLNet

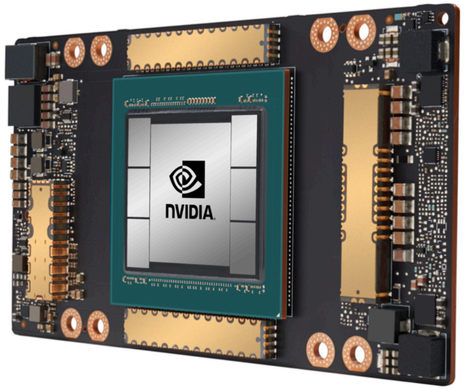
...

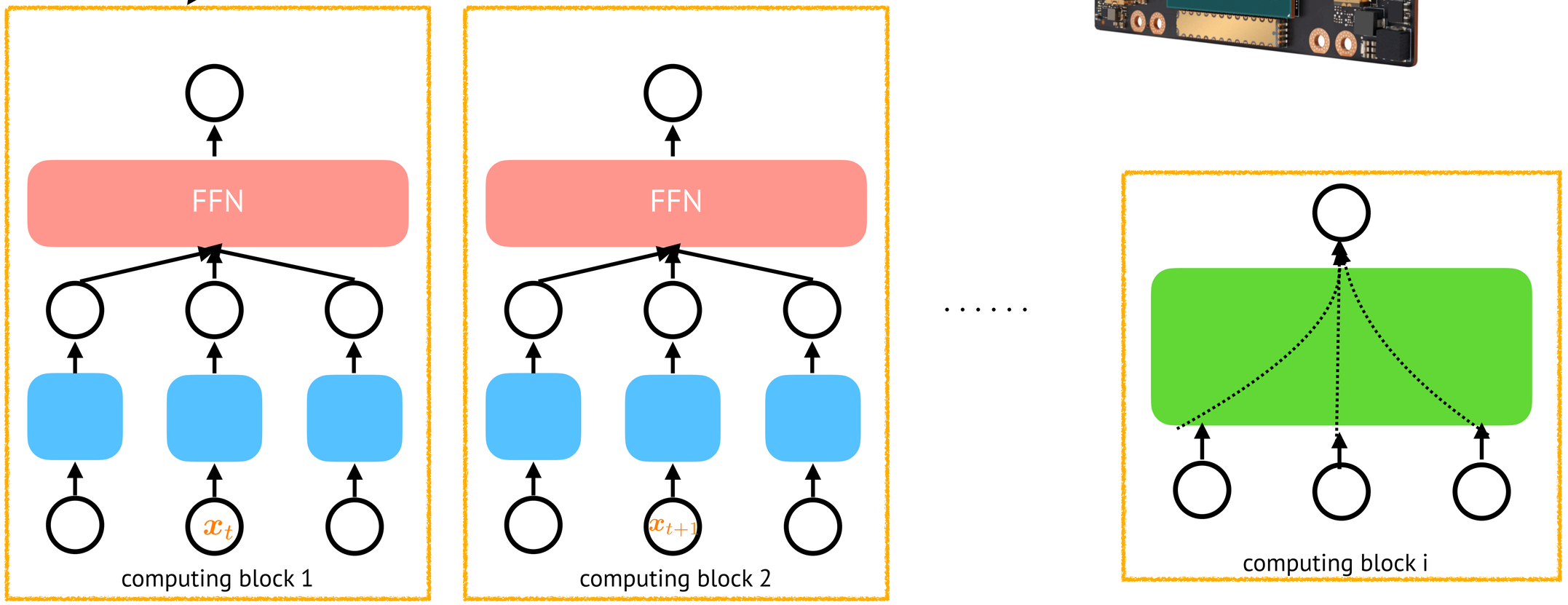


LSTM

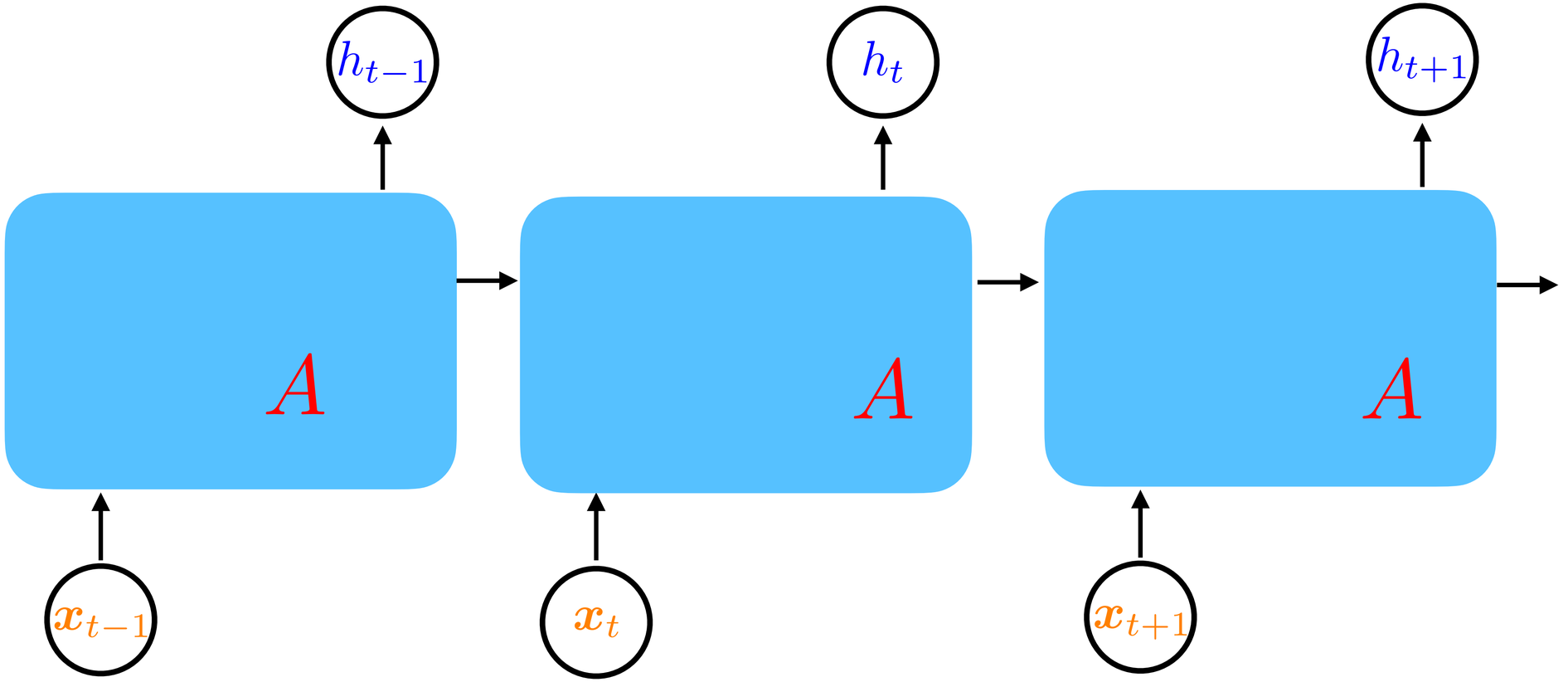
ELMo

Why Transformers?



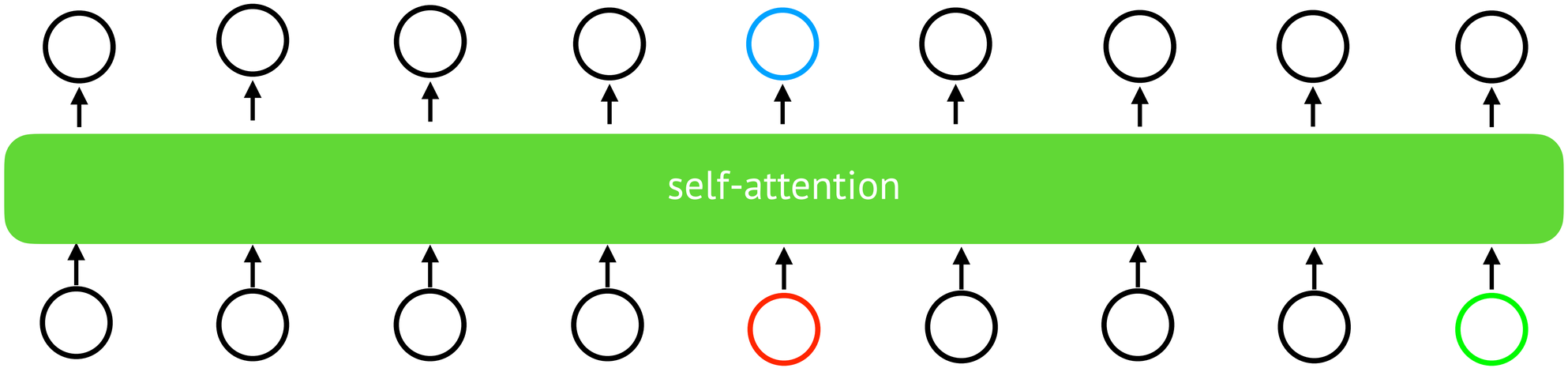


Why Transformers?



Why Transformers?

Direct pair-wise interaction between any tokens in the sequence



Pretraining Objective

training instance (MLM):

x: I feel like eating <MASK> today. What <MASK> you want to eat?

y: noodles, do

training instance (NSP):

x: I feel like eating <MASK> today. ||| What <MASK> you want to eat?

y: True

Pretraining Objective

What makes a good pretraining objective?

1. No human labeling should be involved.

2. Leads to good representations. (How and why?)

Mutual Information

$$I(A, B) = H(A) - H(A \mid B)$$
$$= H(B) - H(B \mid A).$$

$$f_{\boldsymbol{\theta}}(a, b) = g_{\boldsymbol{\psi}}(b)^\top g_{\boldsymbol{\omega}}(a)$$

$$\boldsymbol{\theta} = \{\boldsymbol{\omega}, \boldsymbol{\psi}\}$$

Goal of Training:

$$I(f(A, B)) \geq \mathbb{E}_{p(a,b)} \left[f_{\boldsymbol{\theta}}(a, b) - \mathbb{E}_{q(\tilde{\mathcal{B}})} \left[\log \sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b}) \right] \right] + \log \mid \tilde{\mathcal{B}} \mid,$$

InfoNCE (Logeswaran & Lee, 2018; van den Oord et al., 2019)

$$\mathbb{E}_{p(a,b)} \left[f_{\boldsymbol{\theta}}(a, b) - \log \sum_{\tilde{b} \in \mathcal{B}} \exp f_{\boldsymbol{\theta}}(a, \tilde{b}) \right].$$

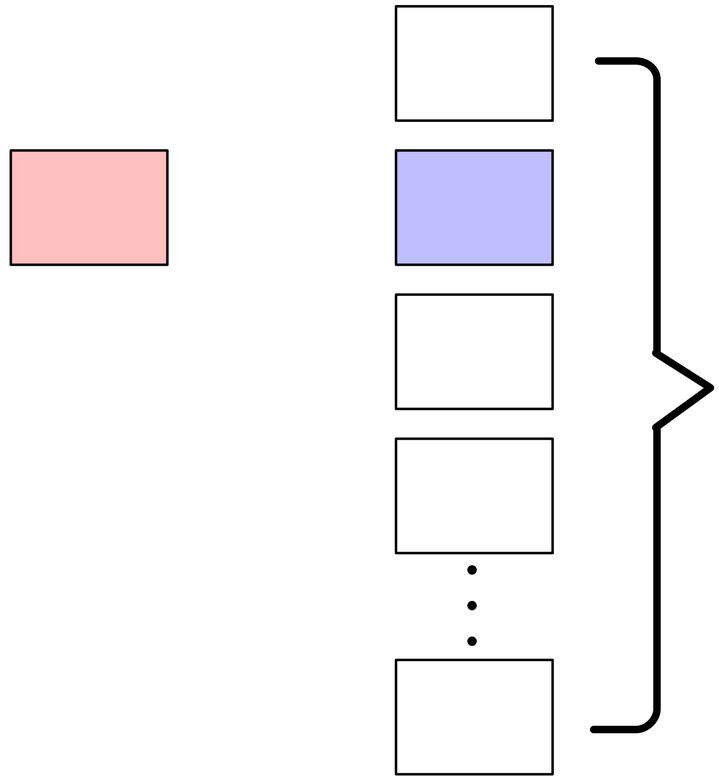
Cross Entropy (Softmax)

Mutual Information

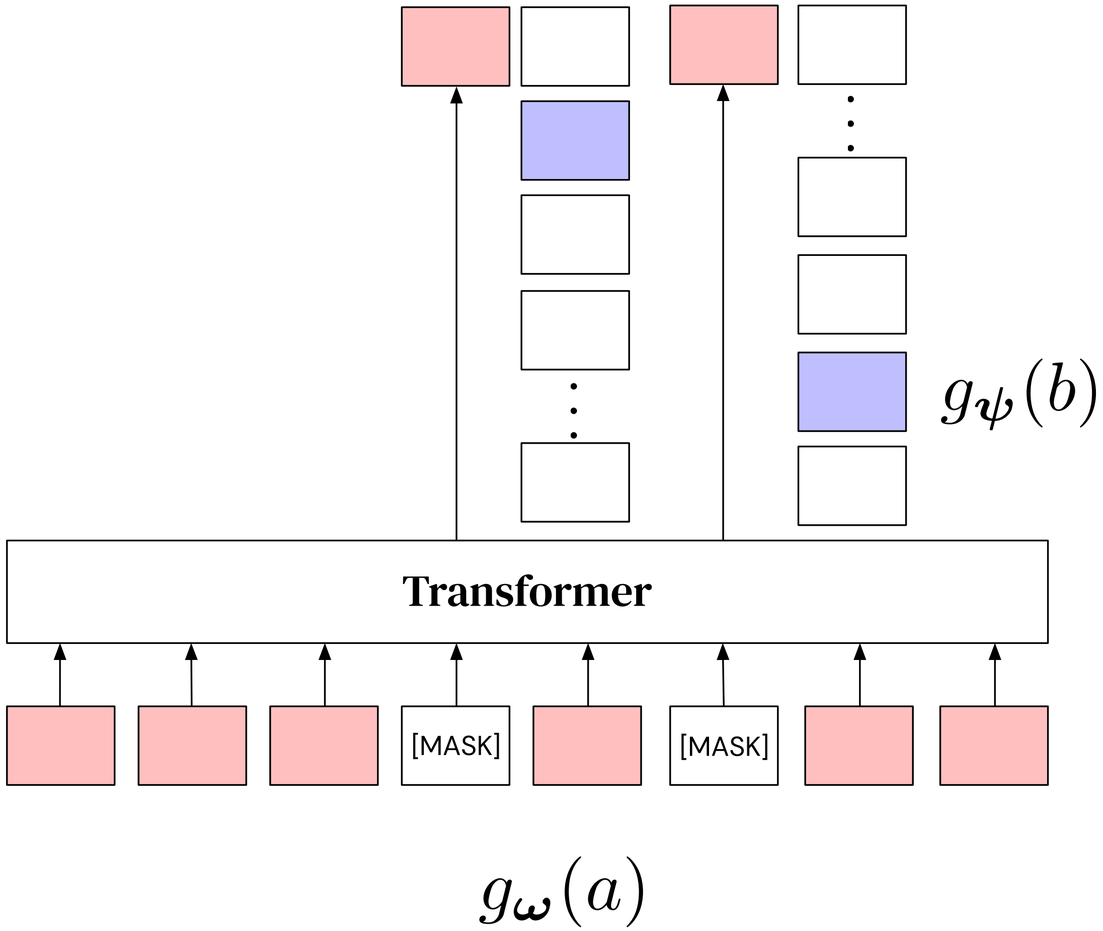
Cross Entropy (Softmax)

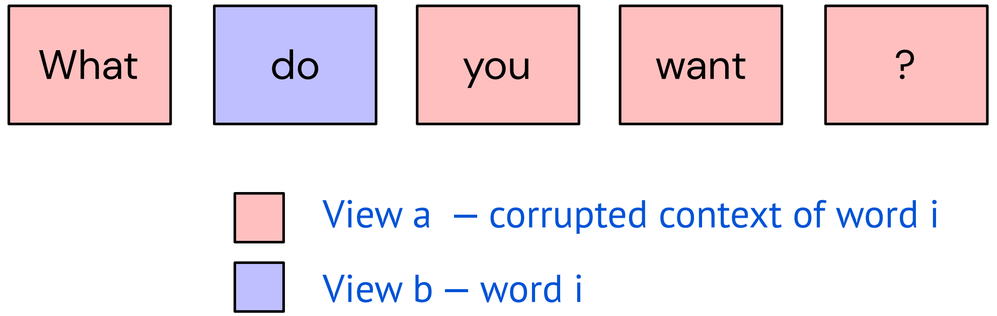
$$\mathbb{E}_{p(a,b)}\left[f_{\boldsymbol{\theta}}(a,b) - \log \sum_{\tilde{b}\in\mathcal{B}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})\right].$$

“Hope” “Fear”

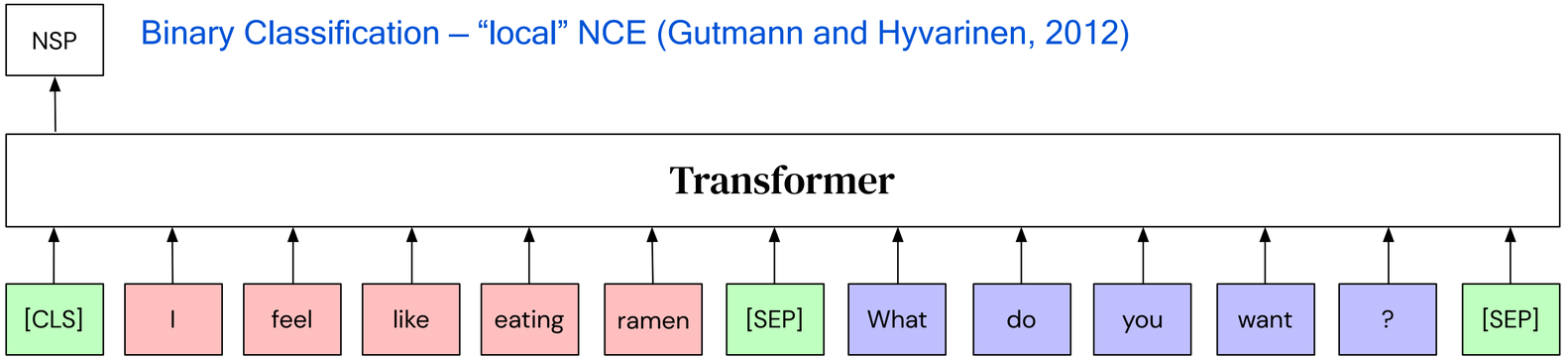


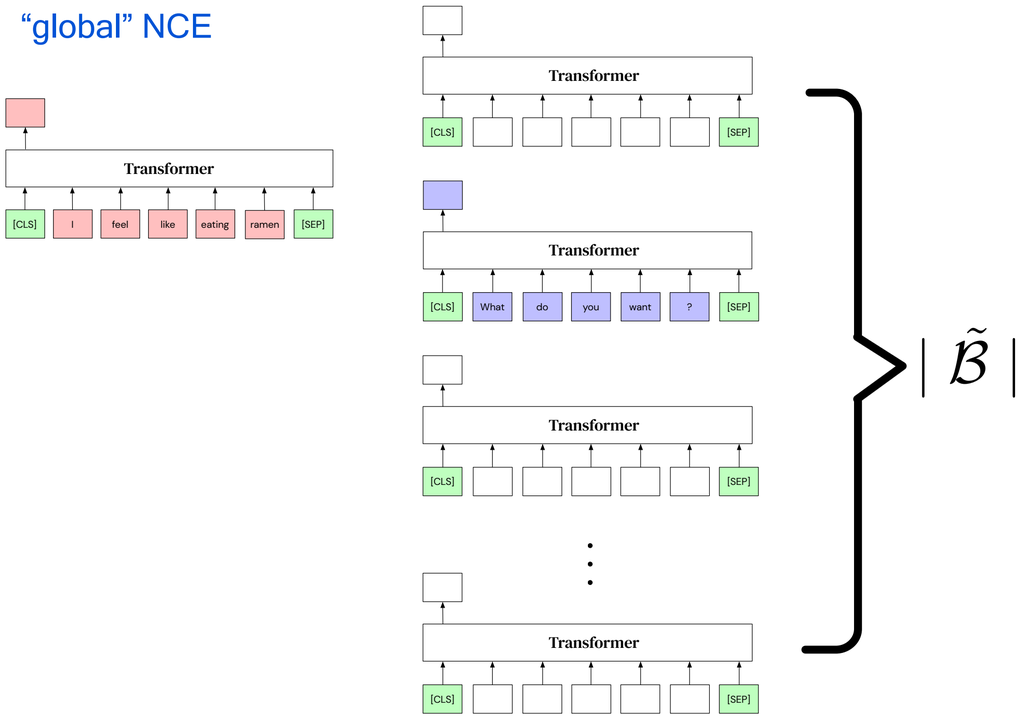
Masked Language Modeling



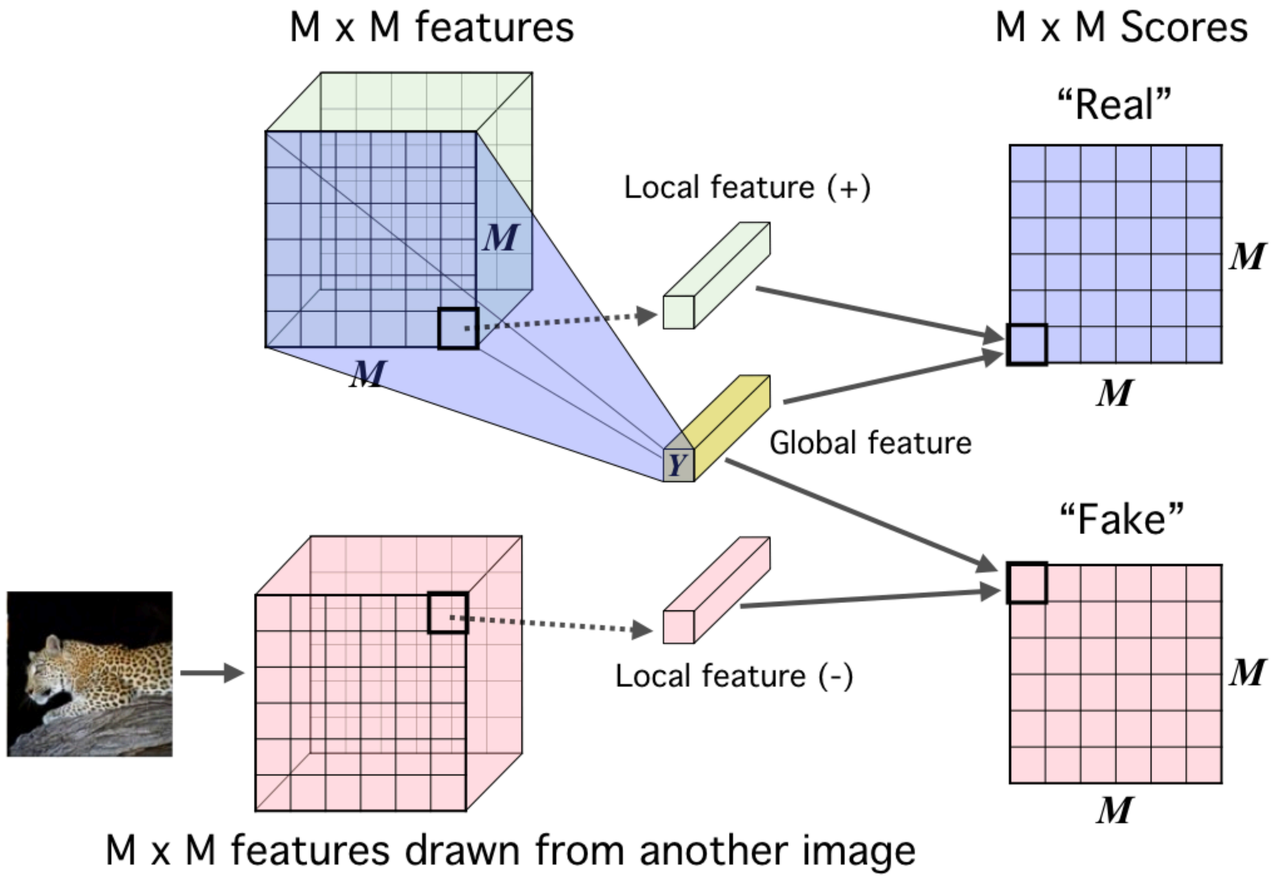


Next Sentence Prediction



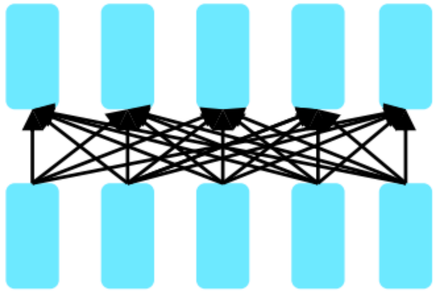


Connections with Computer Vision



Deep InfoMax (DIM; Hjelm et al., 2019)

Type of Architecture



Encoders

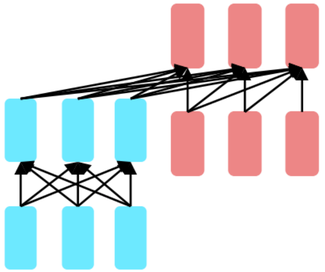
Parameters are what we get from the pretraining process.

Pros for the “encoders” architecture:

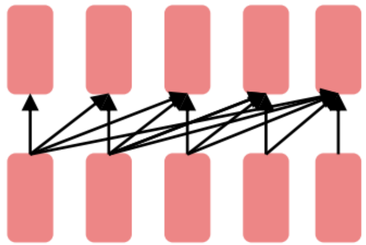
Gets bidirectional context.

Easy to use in language understanding tasks!

Other members in the family:

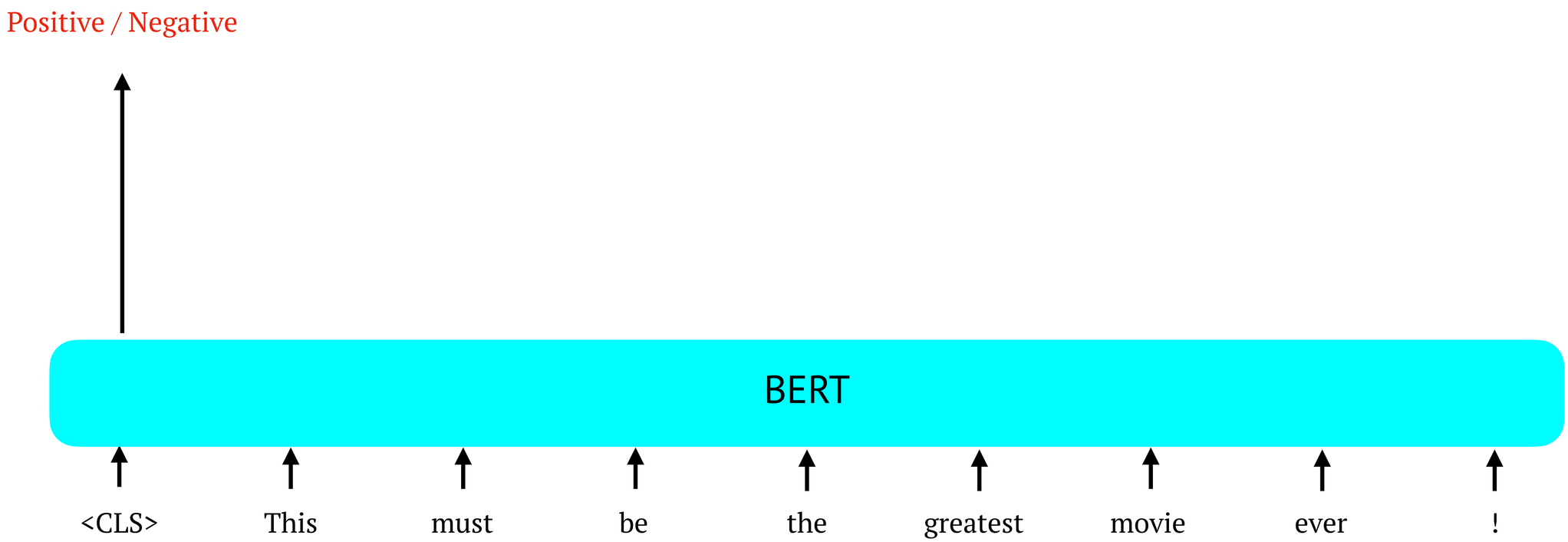


Encoder-Decoders

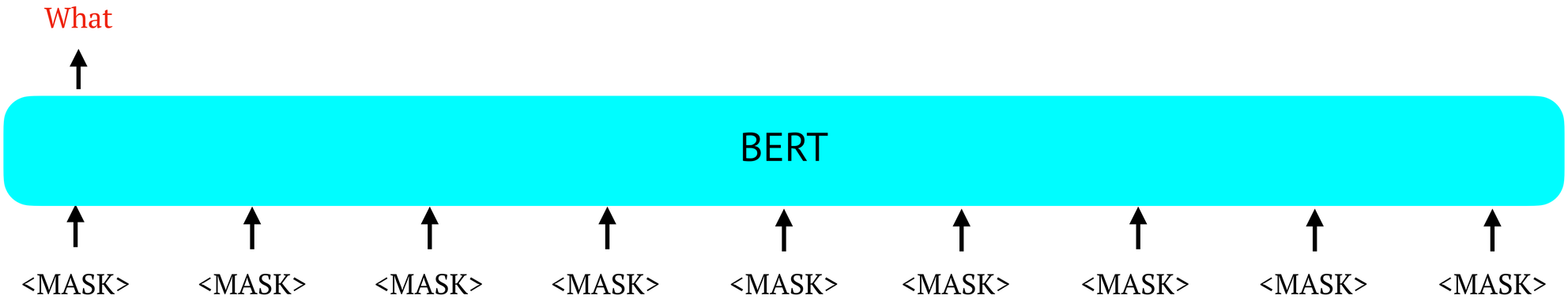


Decoders

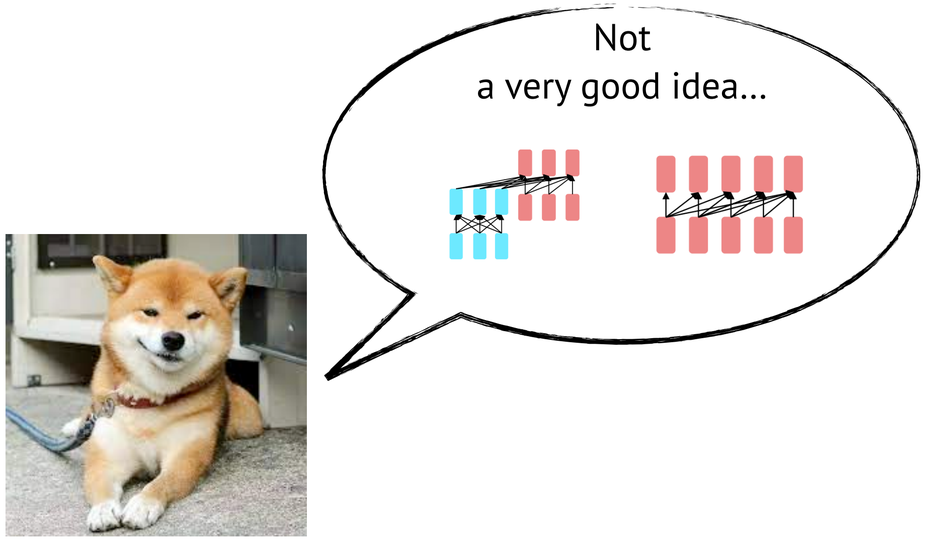
BERT for Understanding



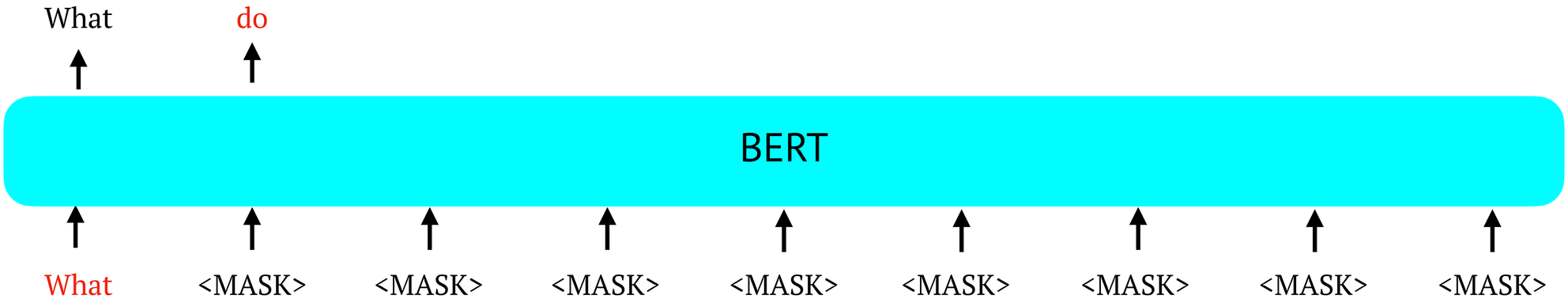
BERT for Generation



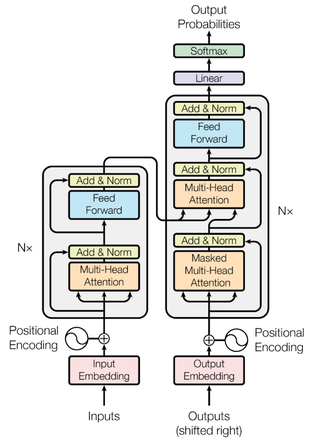
BERT for Generation



Input has been changed. The representations will need to be recomputed!



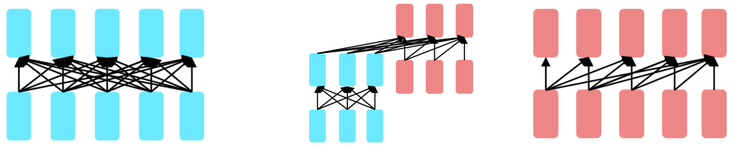
Pretrained Models



— neural representation learner

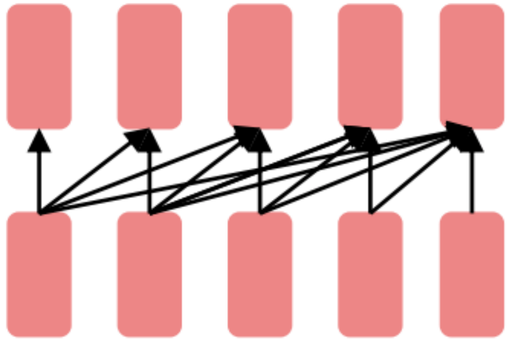
$$\mathbb{E}_{p(x_i, \hat{\boldsymbol{x}}_i)}[p(x_i \mid \hat{\boldsymbol{x}}_i)]$$

— pretraining objective

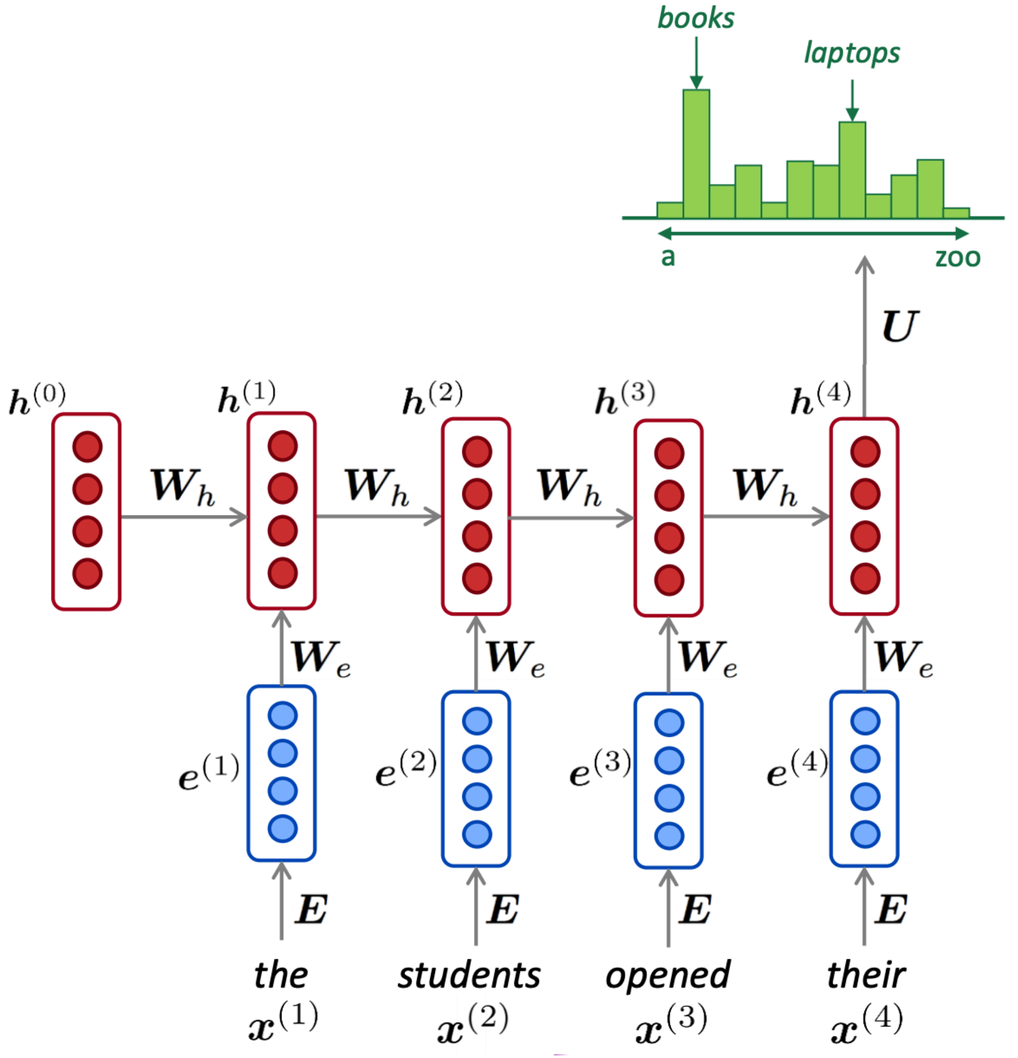


— type of architecture

GPT (Generative Pretrained Transformer)

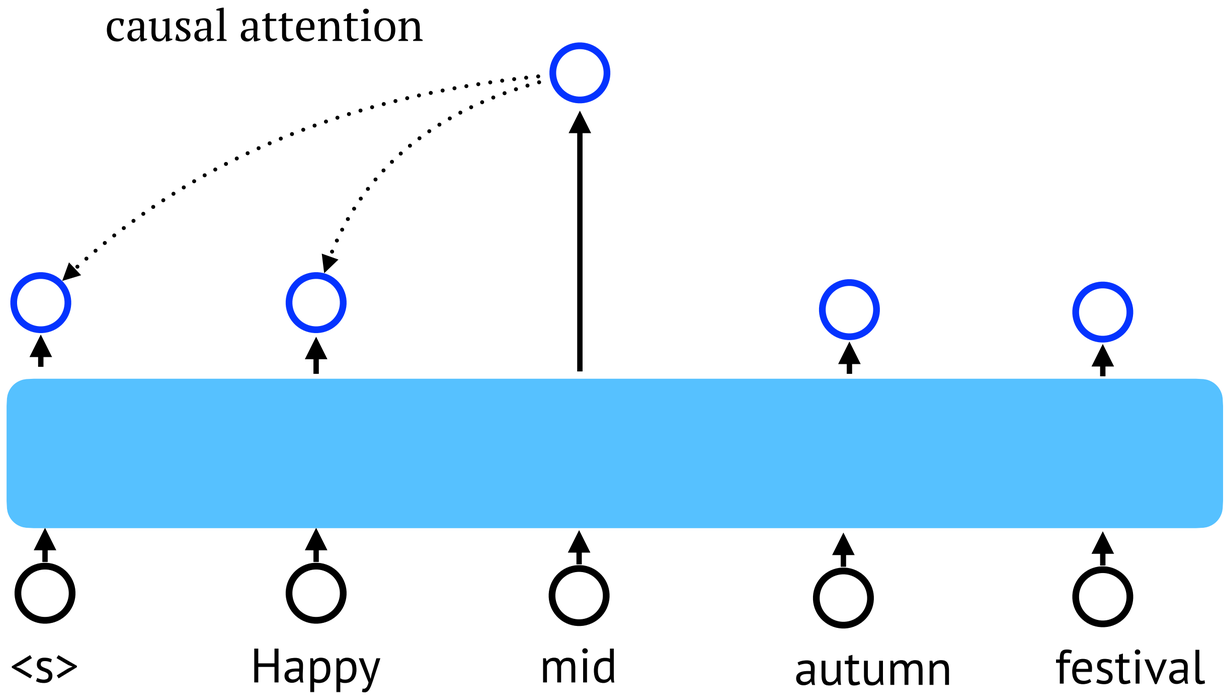


Decoders



Radford et al., 2018

Transformer as Decoder



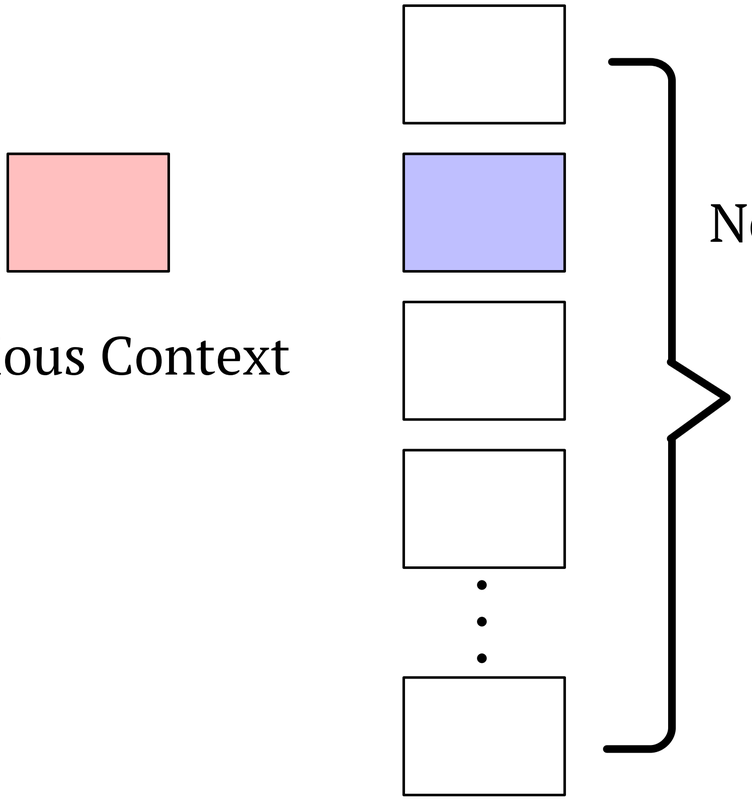
Need to prevent the attention the future words.

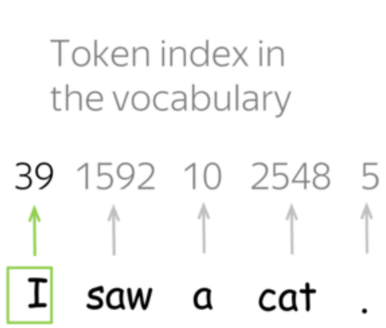
	Happy	mid	autumn	festival
Happy	$-\infty$	$-\infty$	$-\infty$	$-\infty$
mid		$-\infty$	$-\infty$	$-\infty$
autumn			$-\infty$	$-\infty$
festival				$-\infty$

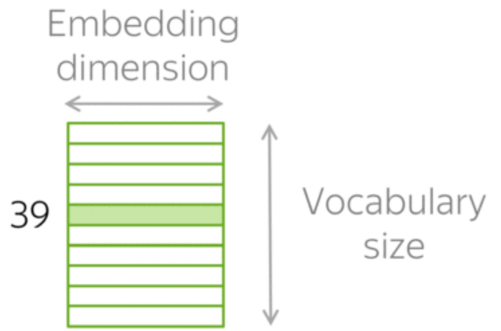
$$e_{ij} = \begin{cases} q_i^\top k_j, j < i \\ -\infty, j \geq i \end{cases}$$

GPT (Generative Pretrained Transformer)

Transformer





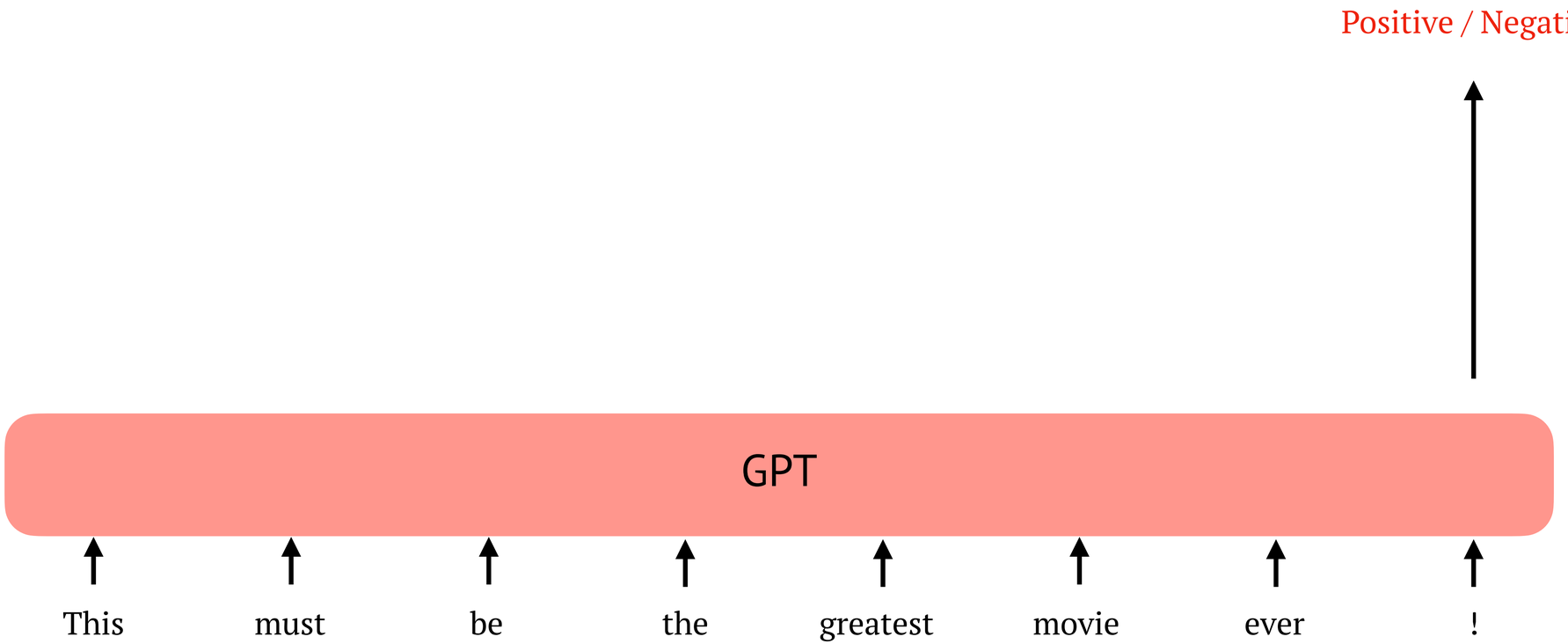


GIF credit: Lena Voita

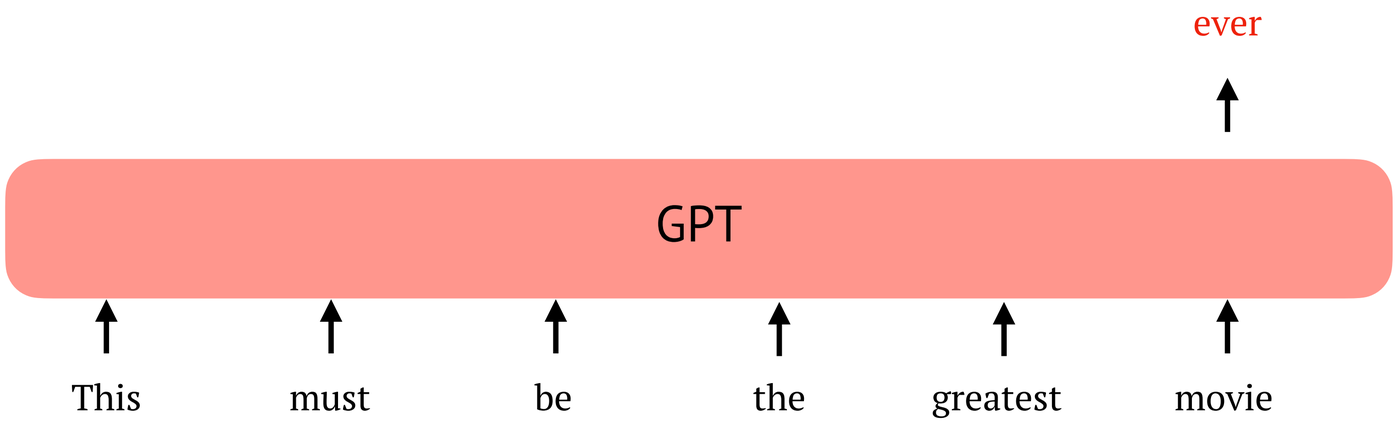
lookup table

Radford et al., 2018

GPT for Understanding

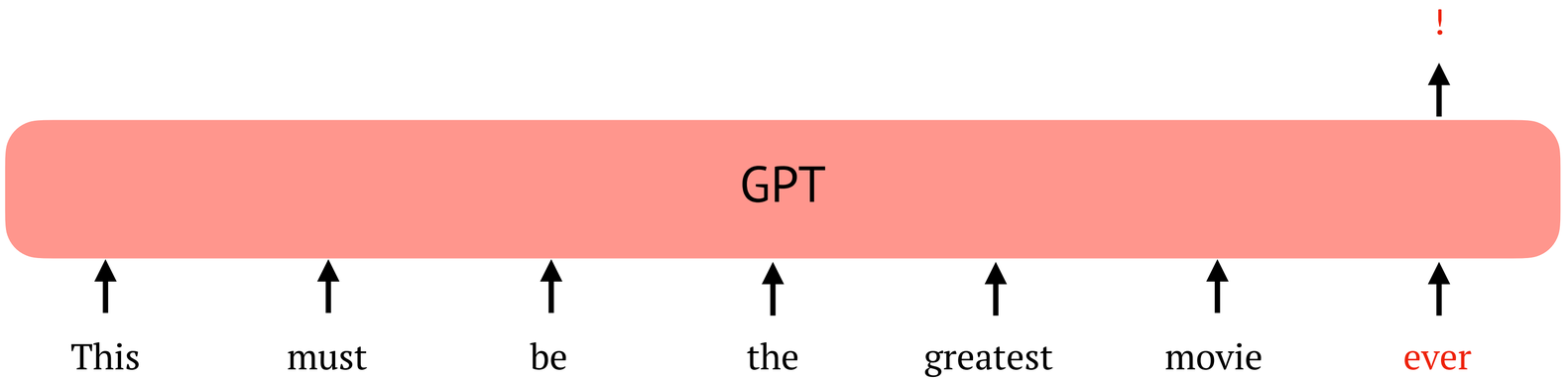


GPT for Generation

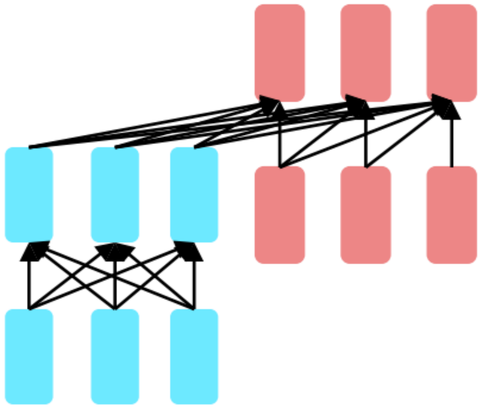


GPT for Generation

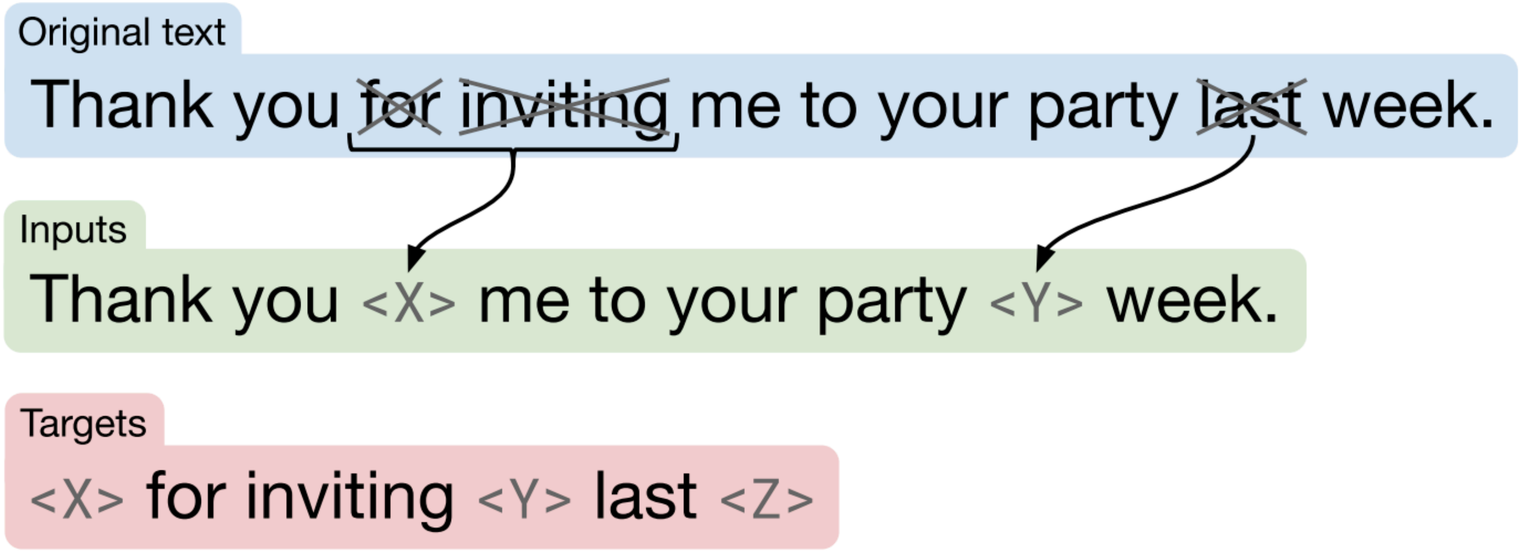
Just “grow” the transformer!



T5 (Text-to-Text Transfer Transformer)



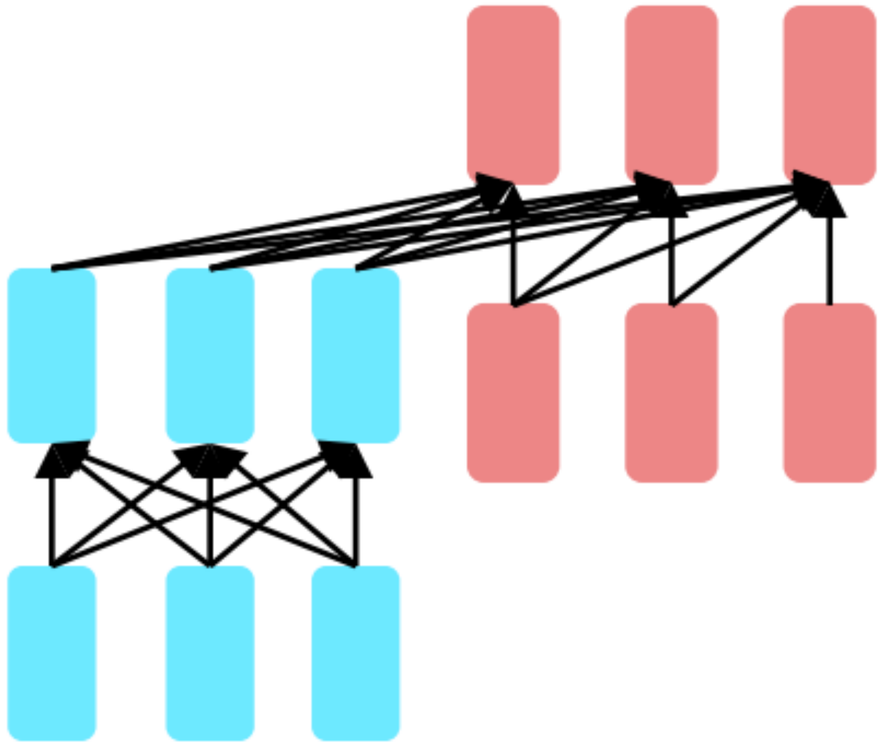
Encoder-Decoders



Raffel et al., 2020

T5 (Text-to-Text Transfer Transformer)

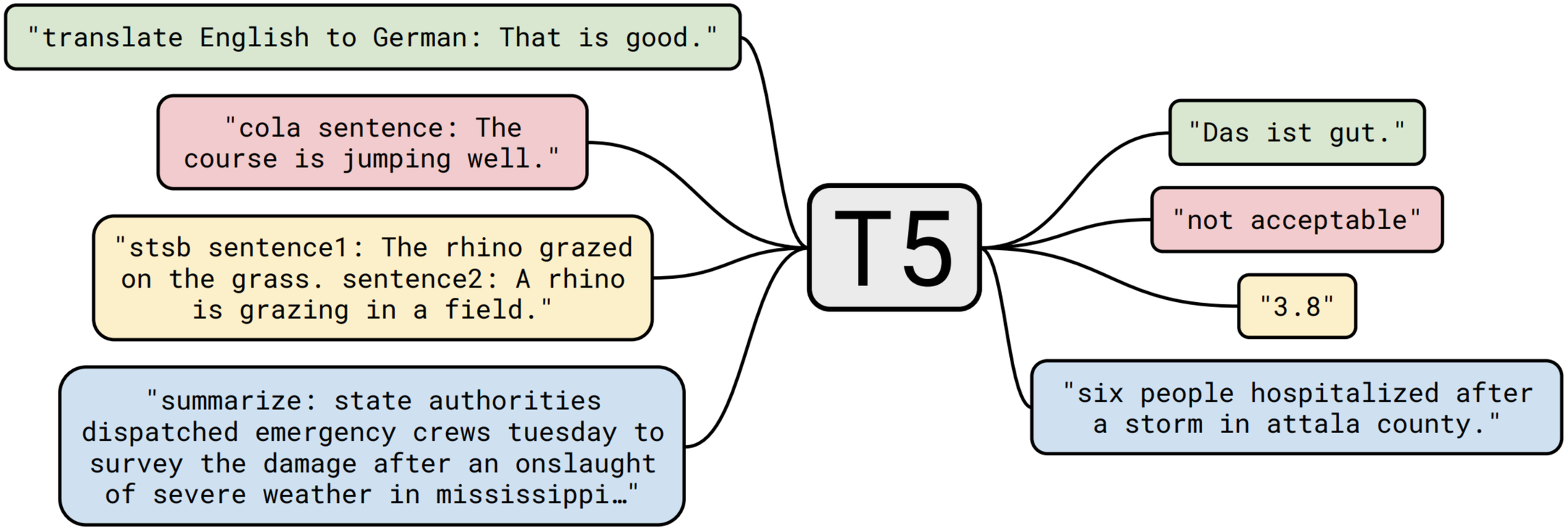
<X> for inviting <Y> last <Z>



Thank you <X> me to your party <Y> week.

Raffel et al., 2020

T5 (Text-to-Text Transfer Transformer)



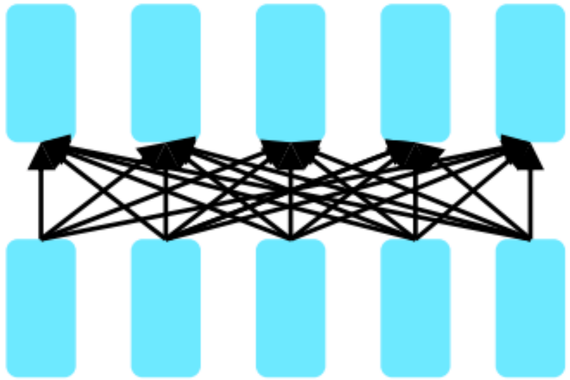
Raffel et al., 2020

T5 (Text-to-Text Transfer Transformer)

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
MASS-style Song et al. (2019)	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>

Raffel et al., 2020

ELMo (Embeddings from Language Models)



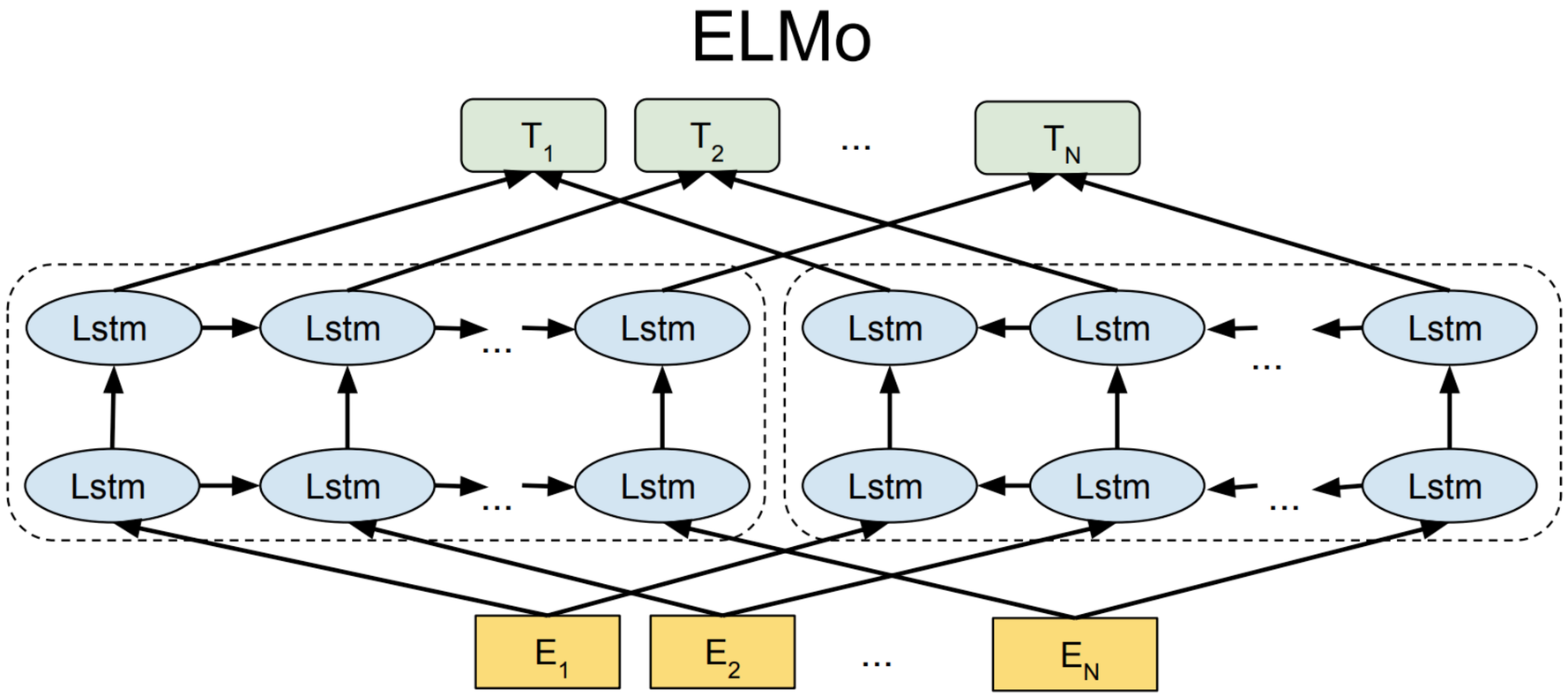
Encoders

$$\sum_{k=1}^{N} \Big(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big)$$

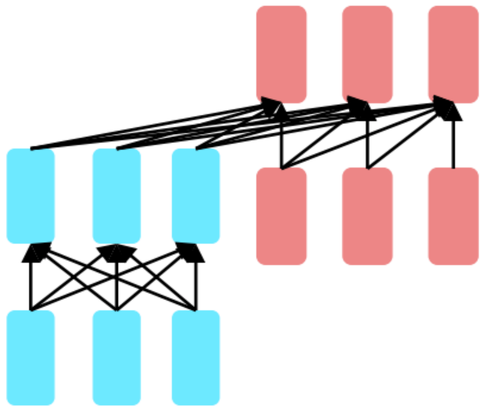
Bidirectional Language Model

Peters et al., 2018

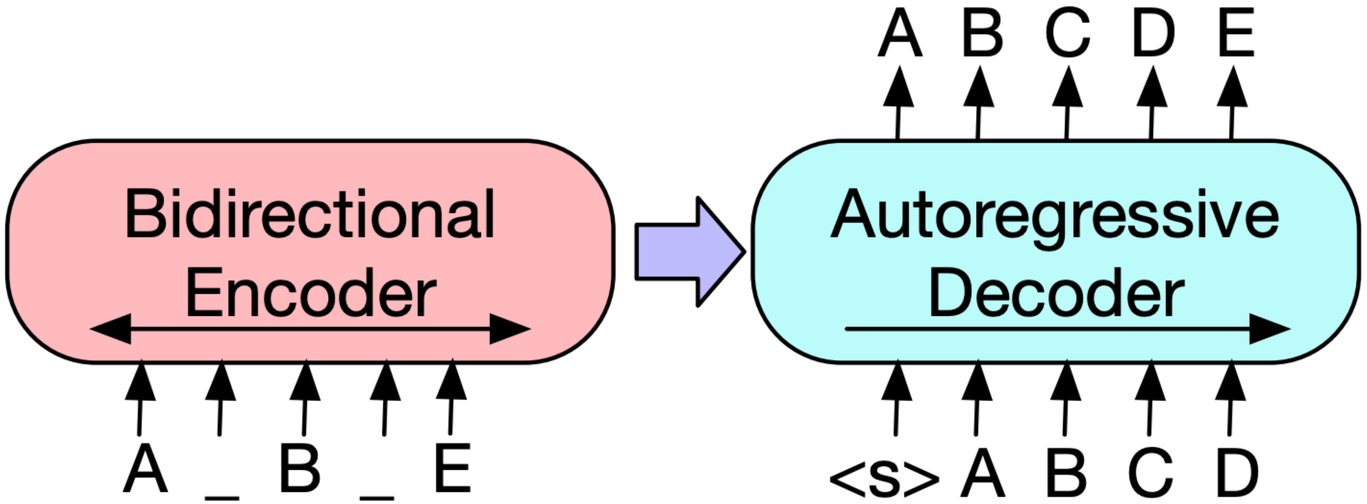
ELMo (Embeddings from Language Models)



BART (Denoising Sequence-to-Sequence Pre-training)

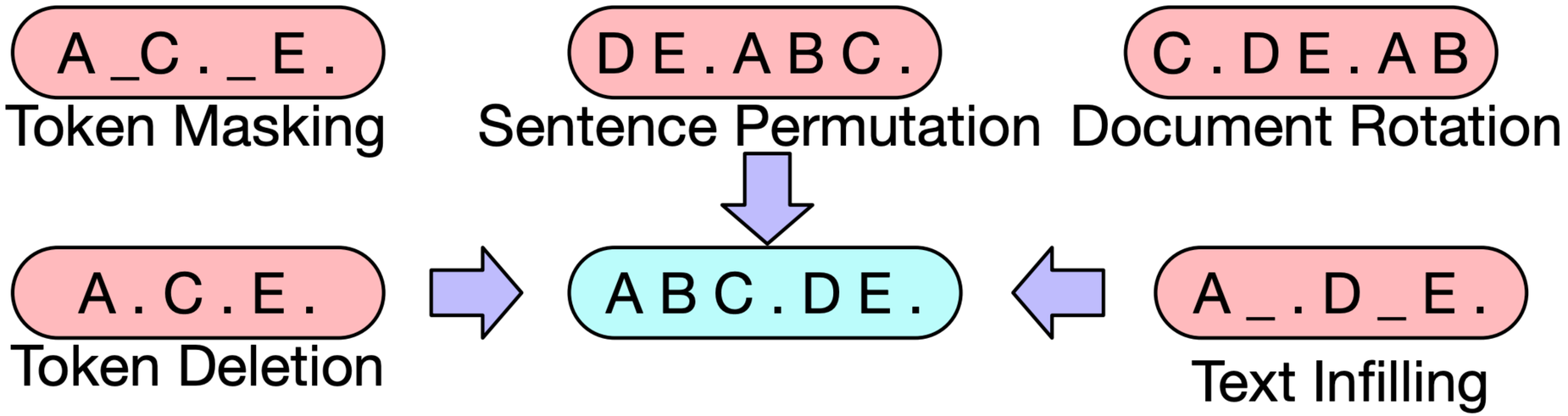


Encoder-Decoders



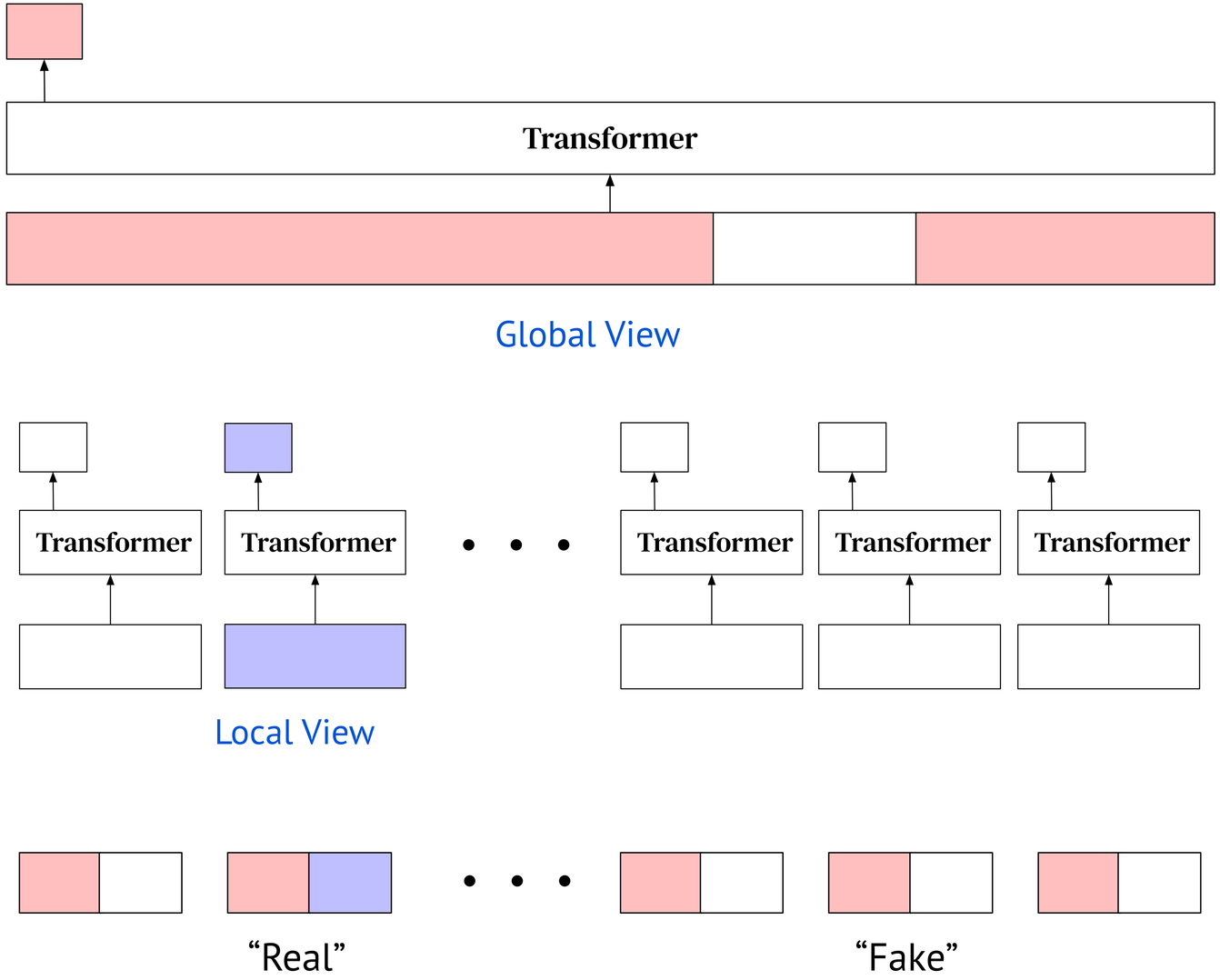
Lewis et al., 2018

BART (Denoising Sequence-to-Sequence Pre-training)



Lewis et al., 2018

InfoWord



Kong et al., 2019