CSCE 638 Natural Language Processing Foundation and Techniques

Lecture 10: Pre-Training and Model Distillation

Kuan-Hao Huang

Spring 2025

(Some slides adapted from Vivian Chen)

SIGIR 2025 LiveRAG Challenge

- https://sigir2025.dei.unipd.it/live-rag-challenge.html
- RAG: Retrieval Augmented Generation
- Advance RAG research and compare the performance of their solutions with other teams on a fixed corpus

SIGIR 2025 LiveRAG Challenge

Date (2025)	Details
Feb 24	**Application submission deadline** - SIGIR2025 easychair site (Select: SIGIR2025 LiveRAG Challenge track)
Mar 12	• Application submission notification to selected teams • Opening of easychair site for short paper submission • AWS and Pinecone resources and credits made available to selected teams together with detailed operational instructions
Mar 15	Training and testing tool (DataMorgana) made available to teams
May 8	"Dry" test for participants of live service on a small question set
May 12	**Live Challenge Day** – test questions shared and live service for answers submission opens
May 19	**Short paper submission deadline** - SIGIR2025 easychair site (Select: SIGIR2025 LiveRAG Challenge track)
May 29	Short paper notification and announcement of finalists
July 17	• LiveRAG Workshop at SIGIR'2025 in Padua, Italy • Presentation of research by selected teams • Announcement of winner and runner(s)-up

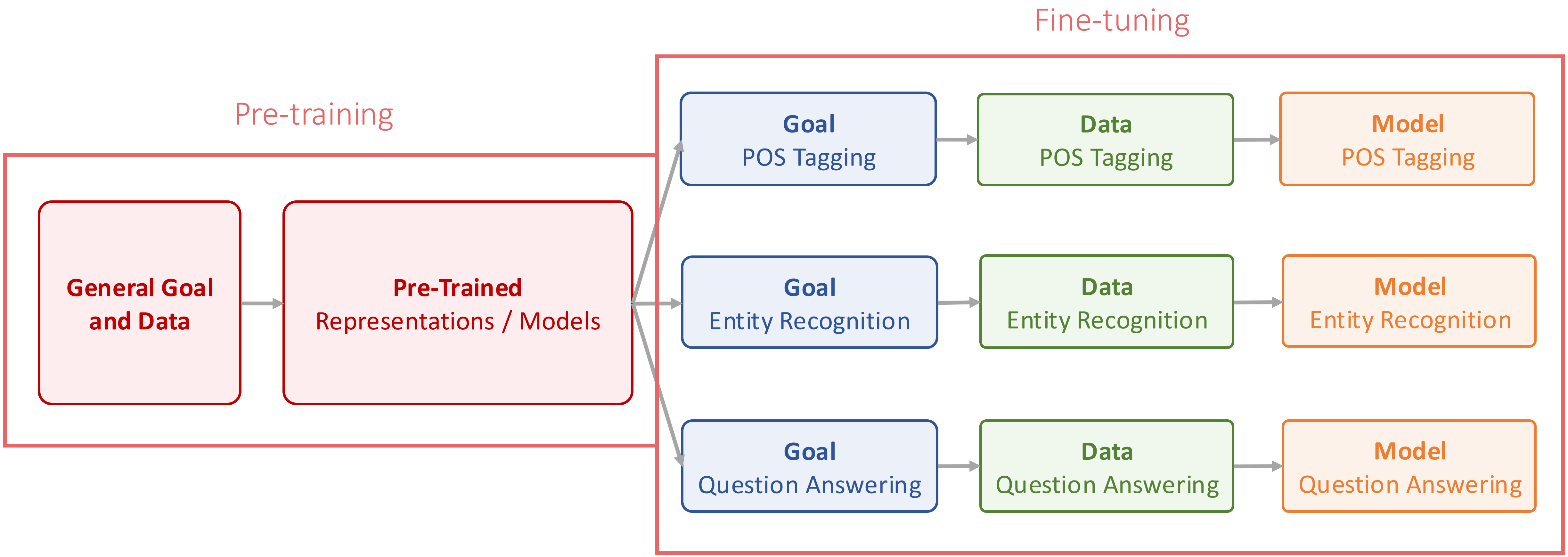
PRIZES

- First Prize: $5000
- Second Prize: $3000
- Third Prize: $2000

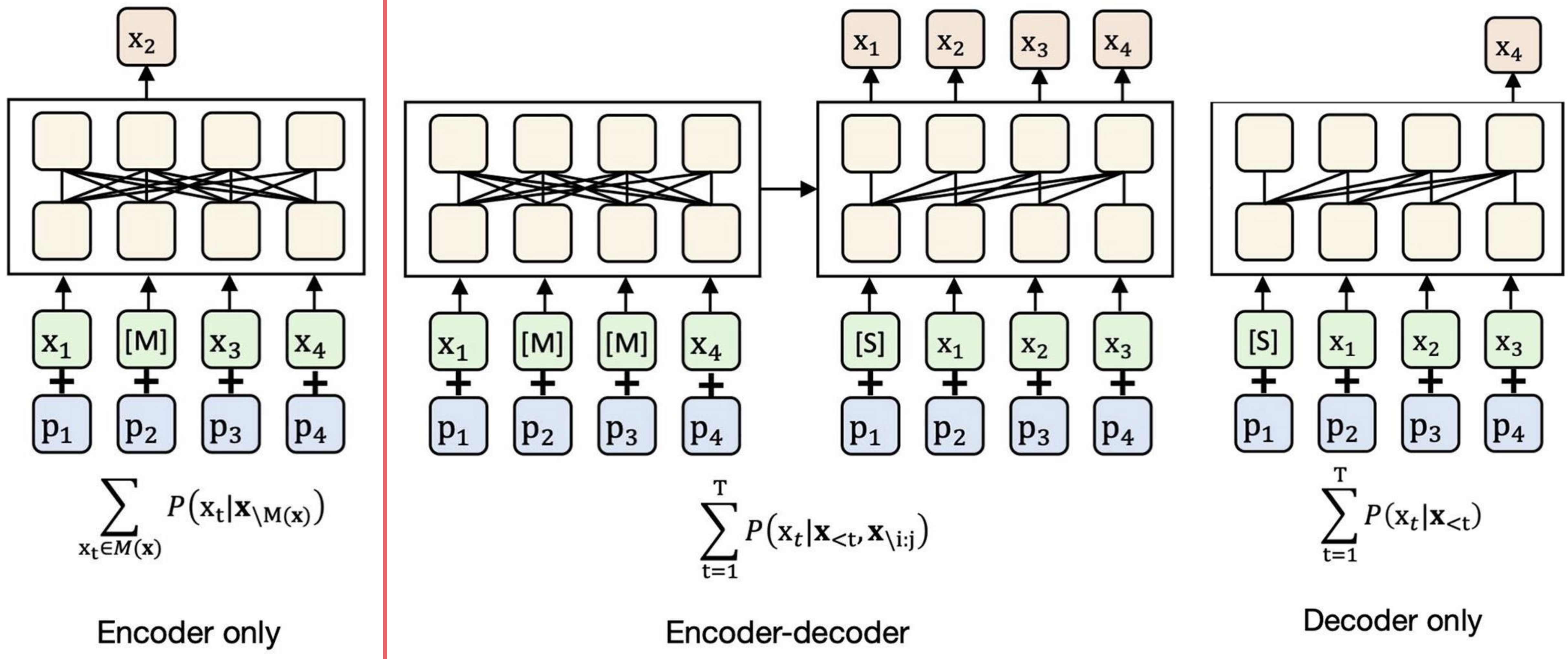
Lecture Plan

- Pre-Training
  - Encoder-Only Pre-Training
  - Encoder-Decoder Pre-Training
  - Decoder-Only Pre-Training
- Model Distillation

Recap: Fine-Tuning with Pre-Training

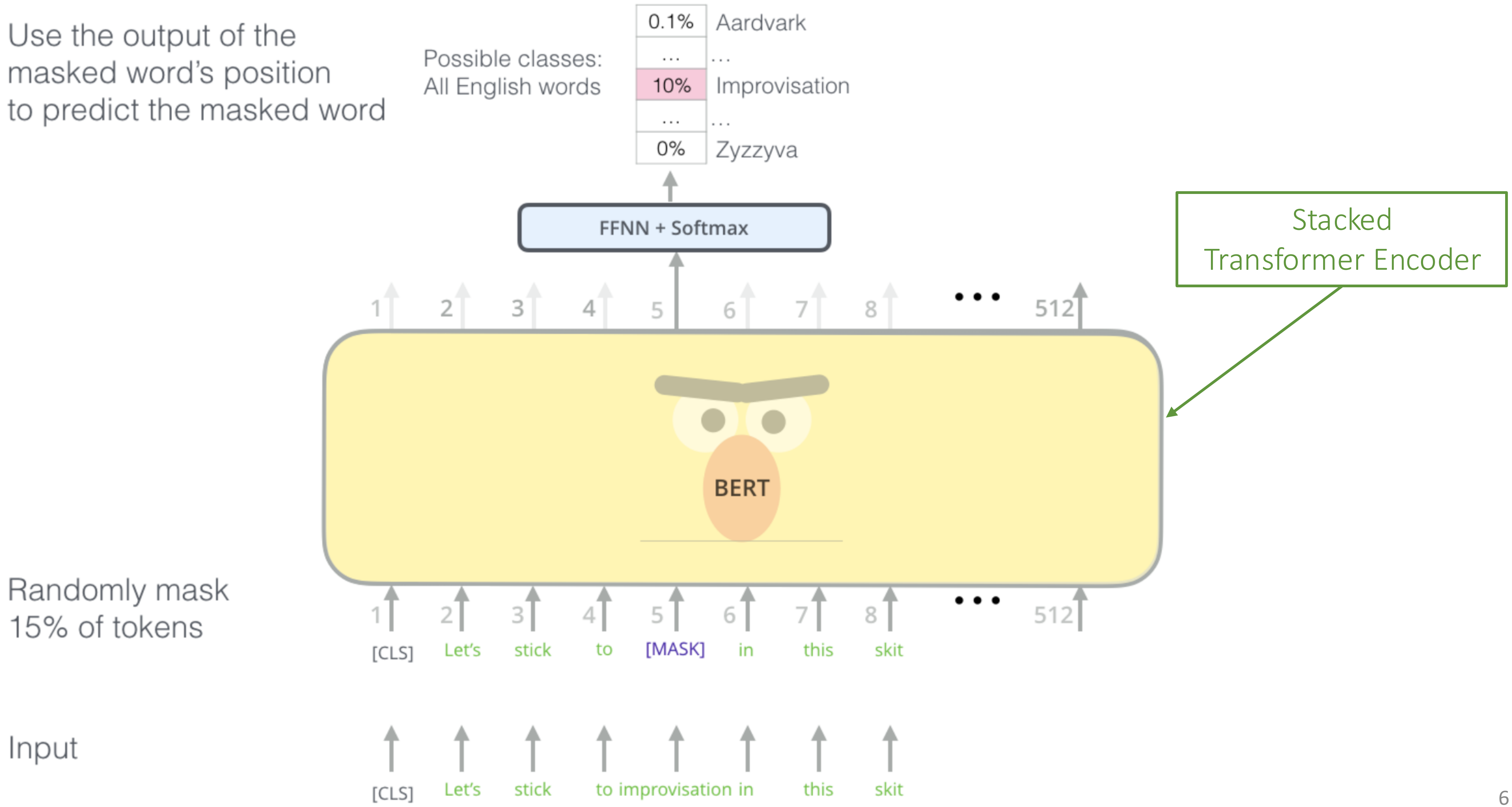


Recap: Types of Pre-Training



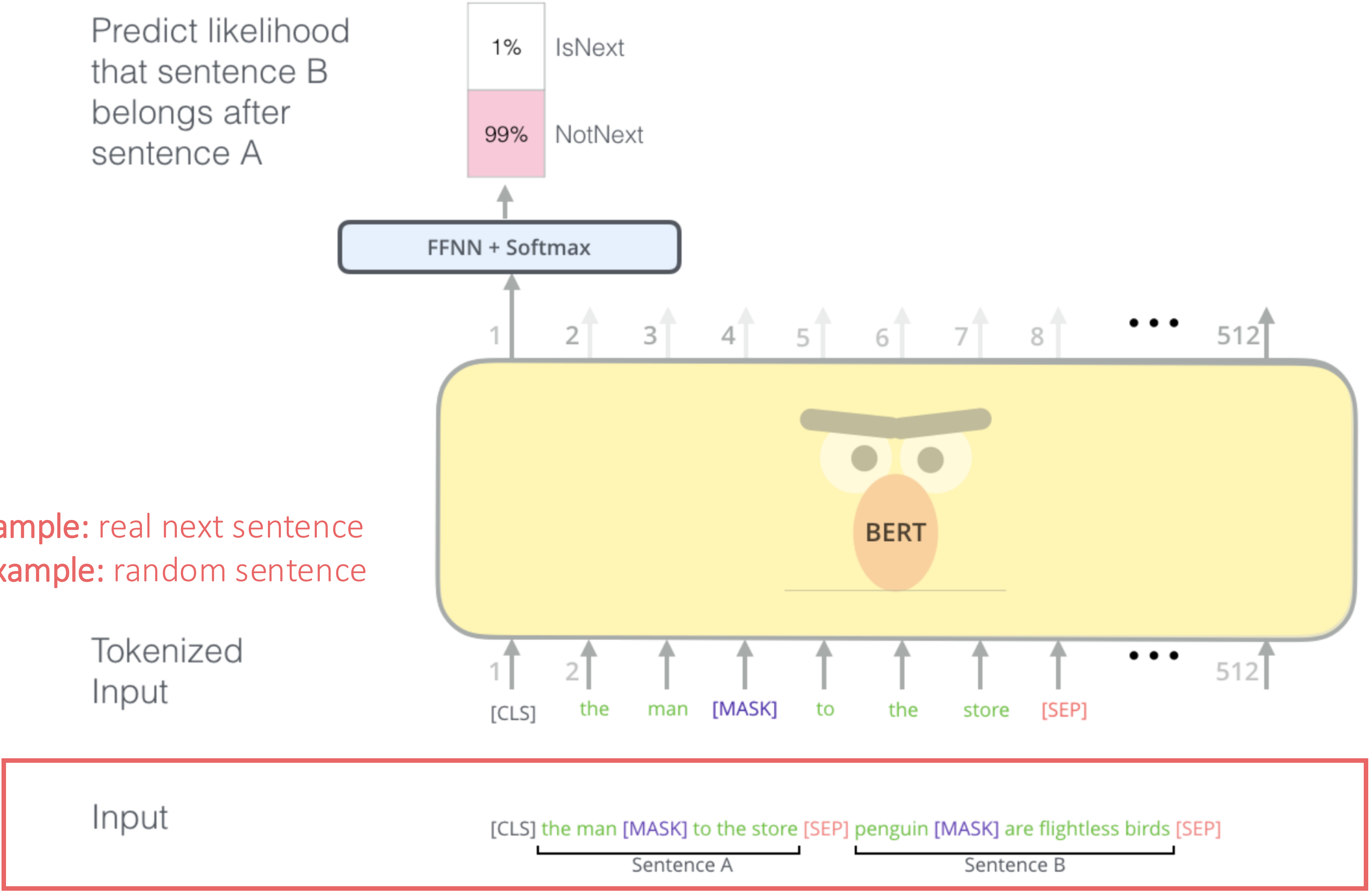
https://www.sciencedirect.com/science/article/pii/S2095809922006324

Recap: BERT – Masked Language Modeling



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

Recap: BERT – Next Sentence Prediction

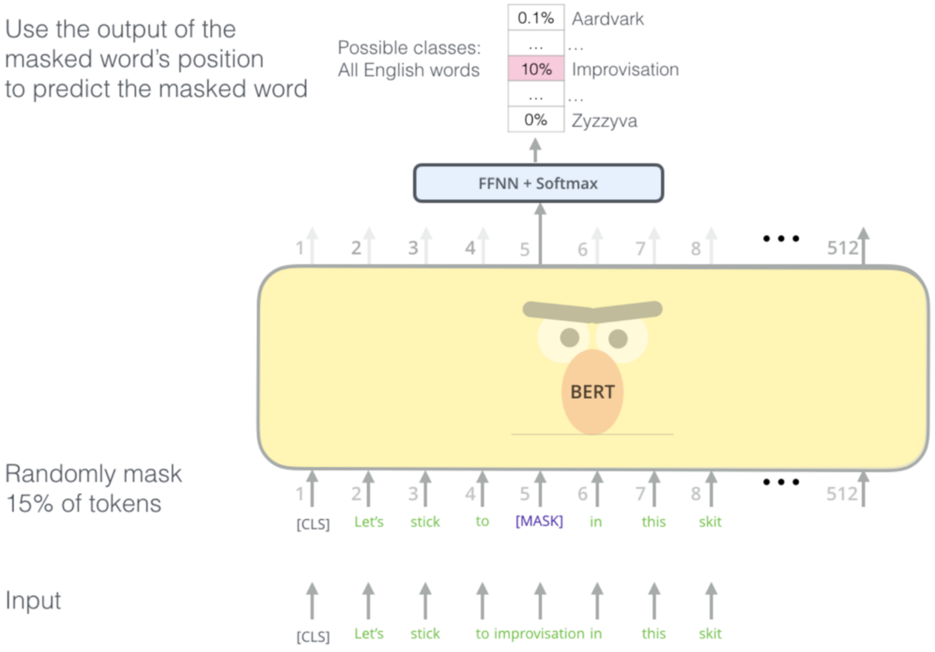


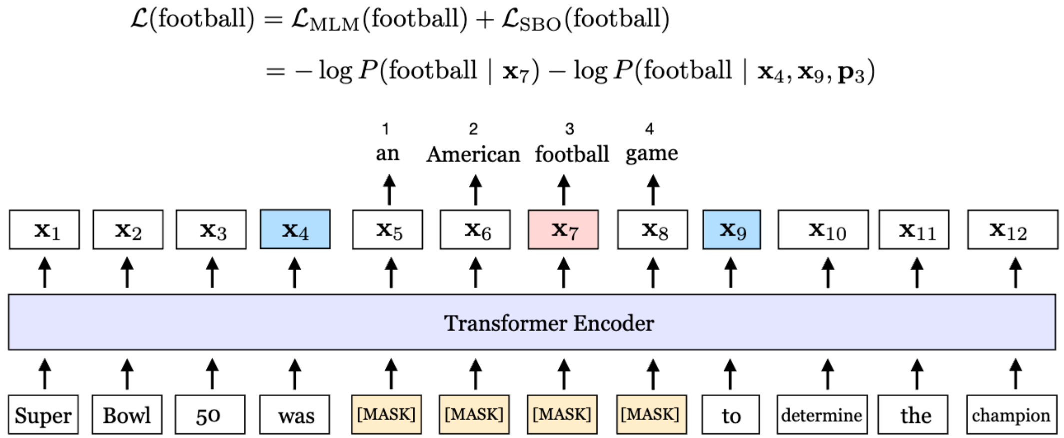
Positive example: real next sentence
Negative example: random sentence

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

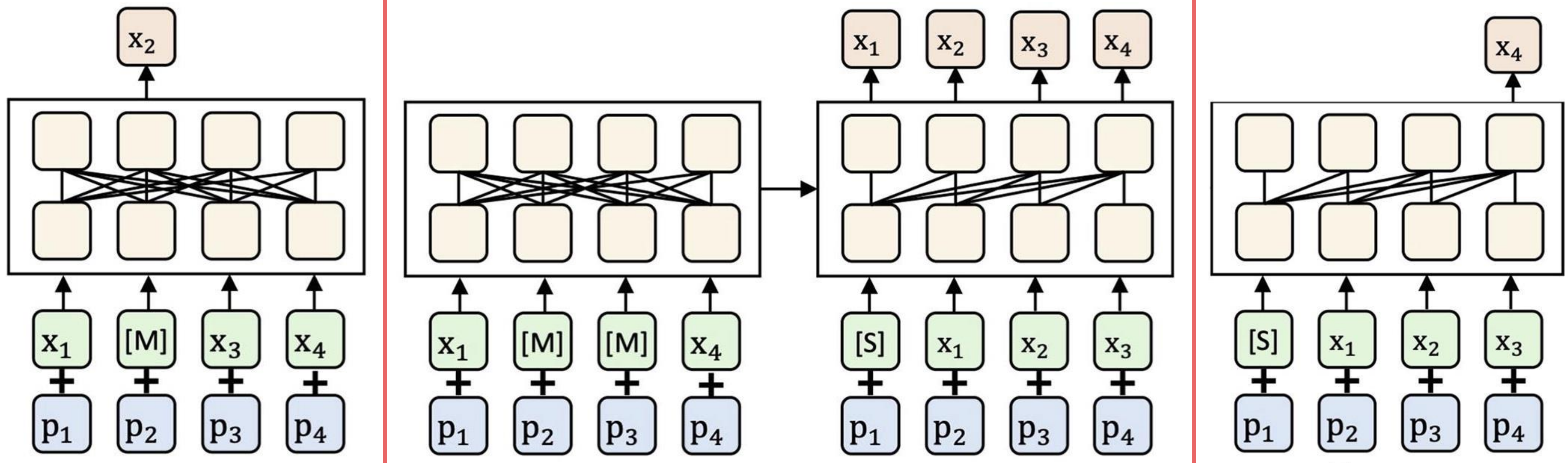
Recap: Other Encoder-Only Pre-Trained Models

- RoBERTa
- SpanBERT





Recap: Types of Pre-Training



$$\sum_{x_t \in M(\mathbf{x})} P\left(x_t | \mathbf{x}_{\setminus M(\mathbf{x})}\right)$$

Encoder only

$$\sum_{t=1}^{T} P\left(x_t | \mathbf{x}_{<t}, \mathbf{x}_{\setminus i:j}\right)$$

Encoder-decoder

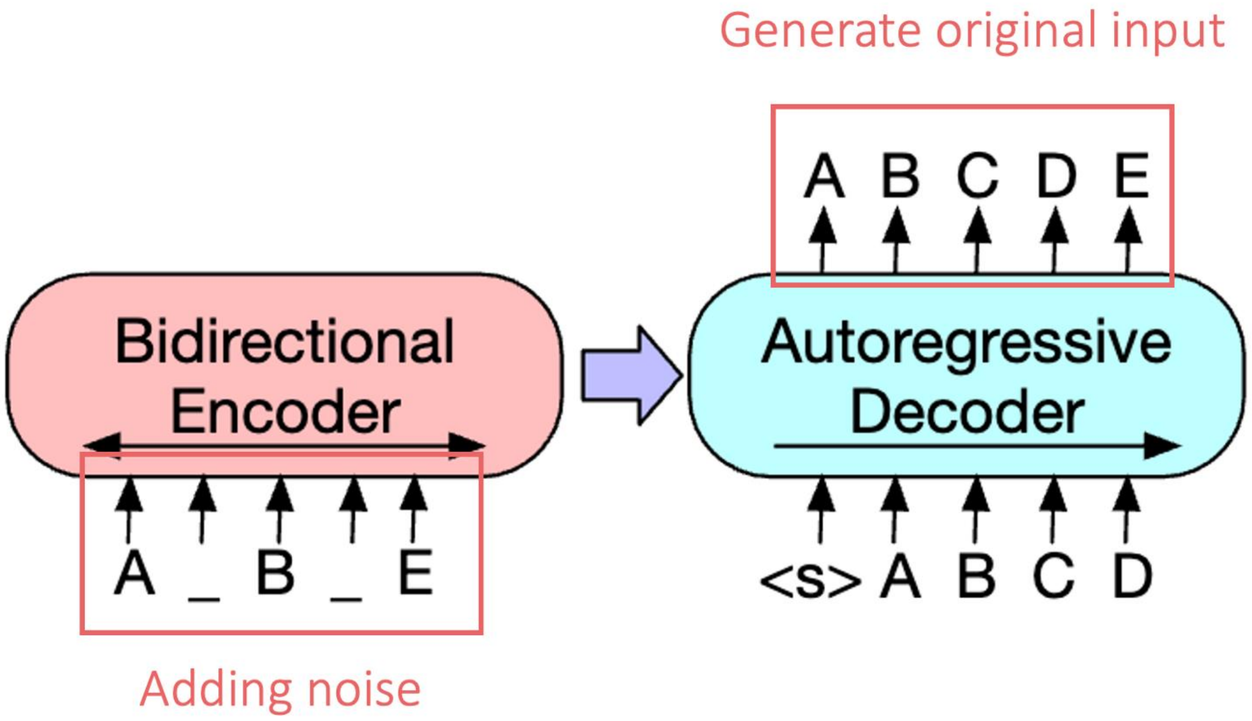
$$\sum_{t=1}^{T} P\left(x_t | \mathbf{x}_{<t}\right)$$

Decoder only

https://www.sciencedirect.com/science/article/pii/S2095809922006324

Recap: BART – Denoising Objective

- Token Masking
  - A<mask>CD<mask>F. → ABCDEF
- Token Deletion
  - ACDF. → ABCDEF.
- Text Infilling
  - A<mask>D<mask>F. → ABCDEF.
- Sentence Permutation
  - FG. ABC. DE. → ABC. DE. FG.
- Document Rotation
  - E. FG. ABC. D → ABC. DE. FG.



Encoder-Decoder: T5

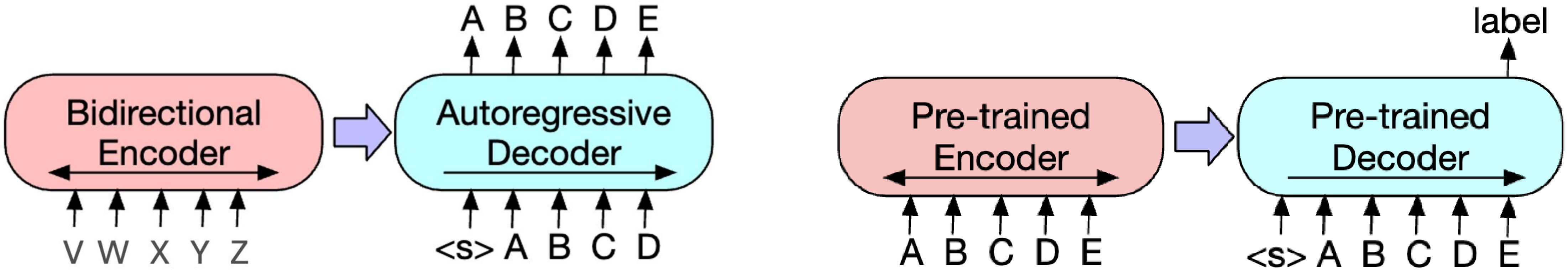
- Text-to-Text Transfer Transformer (T5)

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel*	CRAFFEL@GMAIL.COM
Noam Shazeer*	NOAM@GOOGLE.COM
Adam Roberts*	ADAROB@GOOGLE.COM
Katherine Lee*	KATHERINELEE@GOOGLE.COM
Sharan Narang	SHARANNARANG@GOOGLE.COM
Michael Matena	MMATENA@GOOGLE.COM
Yanqi Zhou	YANQIZ@GOOGLE.COM
Wei Li	MWEILI@GOOGLE.COM
Peter J. Liu	PETERJLIU@GOOGLE.COM

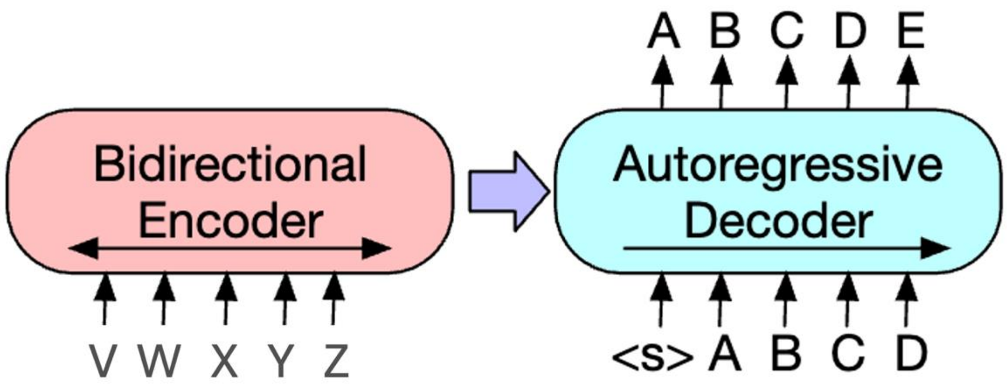
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2020

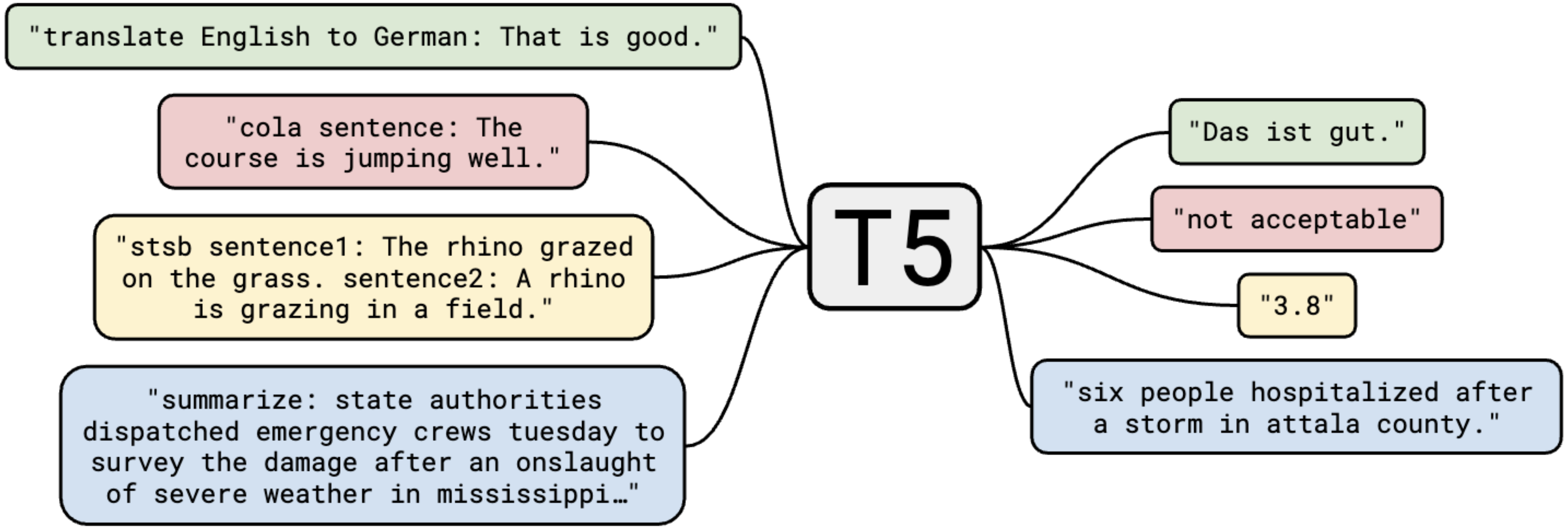
Motivation: BART



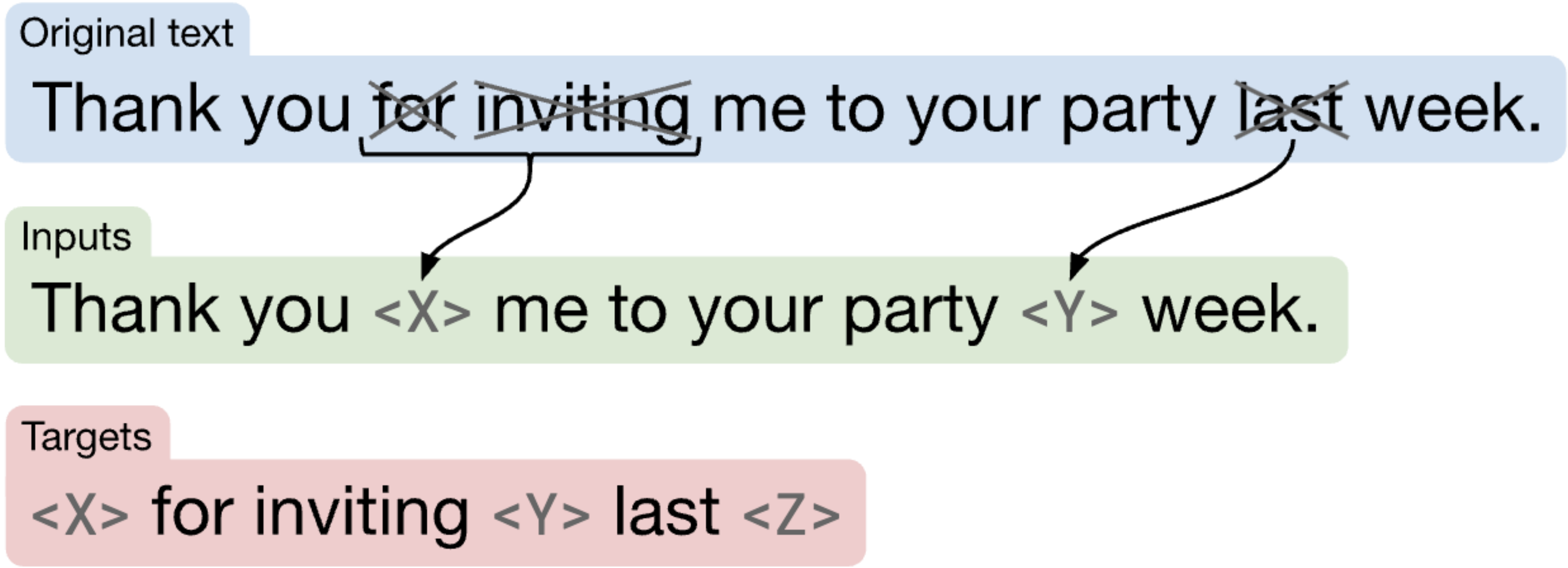
Different ways when considering classification and seq2seq generation

Convert Everything to Text-to-Text Tasks





Masked Span Reconstruction (Seq2Seq Version)



Multi-Task Learning

- Convert everything to text-to-text tasks
- Jointly fine-tune them together

Multi-Task Learning

D.7 SST2

Original input:

Sentence: `it confirms fincher 's status as a film maker who artfully bends technical know-how to the service of psychological insight .`

Processed input: `sst2 sentence: it confirms fincher 's status as a film maker who artfully bends technical know-how to the service of psychological insight .`

Original target: `1`

Processed target: `positive`

Multi-Task Learning

D.4 MRPC

Original input:

Sentence 1: `We acted because we saw the existing evidence in a new light , through the prism of our experience on 11 September , " Rumsfeld said .`

Sentence 2: `Rather , the US acted because the administration saw " existing evidence in a new light , through the prism of our experience on September 11 " .`

Processed input: `mrpc sentence1: We acted because we saw the existing evidence in a new light , through the prism of our experience on 11 September , " Rumsfeld said . sentence2: Rather , the US acted because the administration saw " existing evidence in a new light , through the prism of our experience on September 11 " .`

Original target: `1`

Processed target: `equivalent`

Multi-Task Learning

D.16 WMT English to German

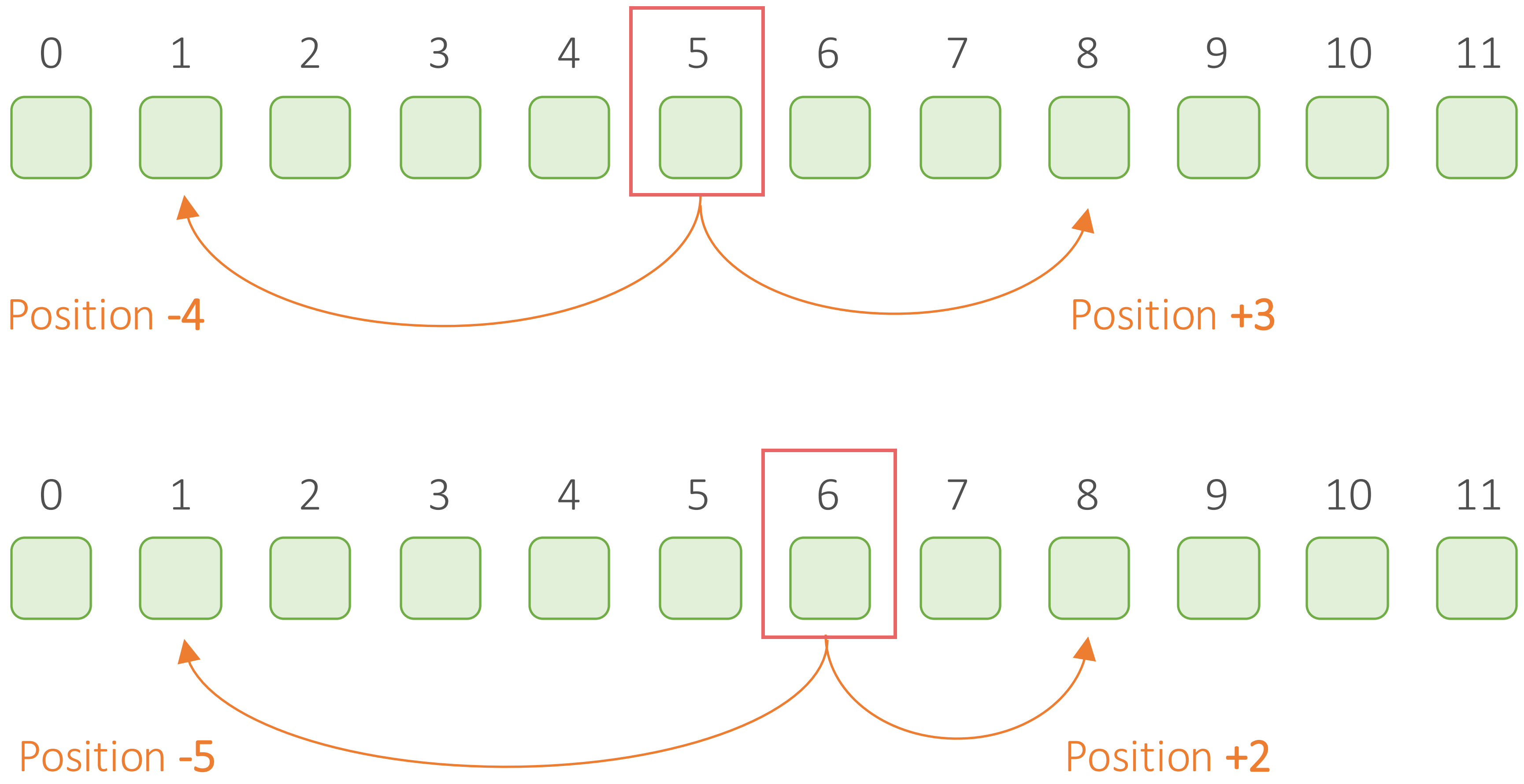
Original input: "Luigi often said to me that he never wanted the brothers to end up in court," she wrote.

Processed input: translate English to German: "Luigi often said to me that he never wanted the brothers to end up in court," she wrote.

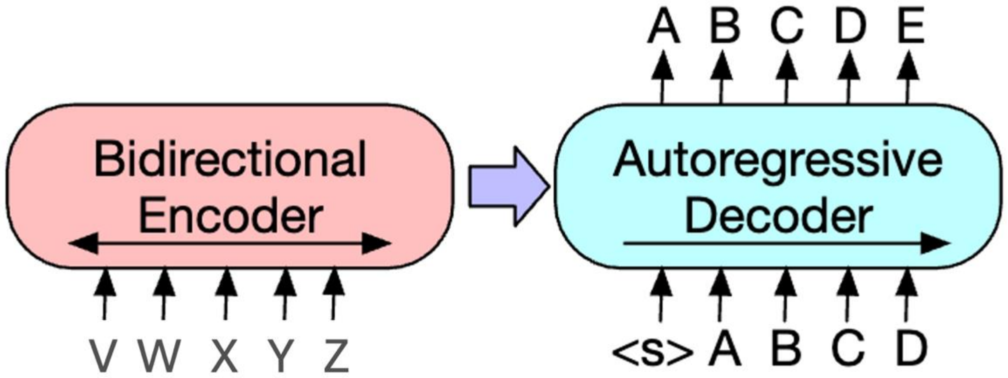
Original target: "Luigi sagte oft zu mir, dass er nie wollte, dass die Brüder vor Gericht landen", schrieb sie.

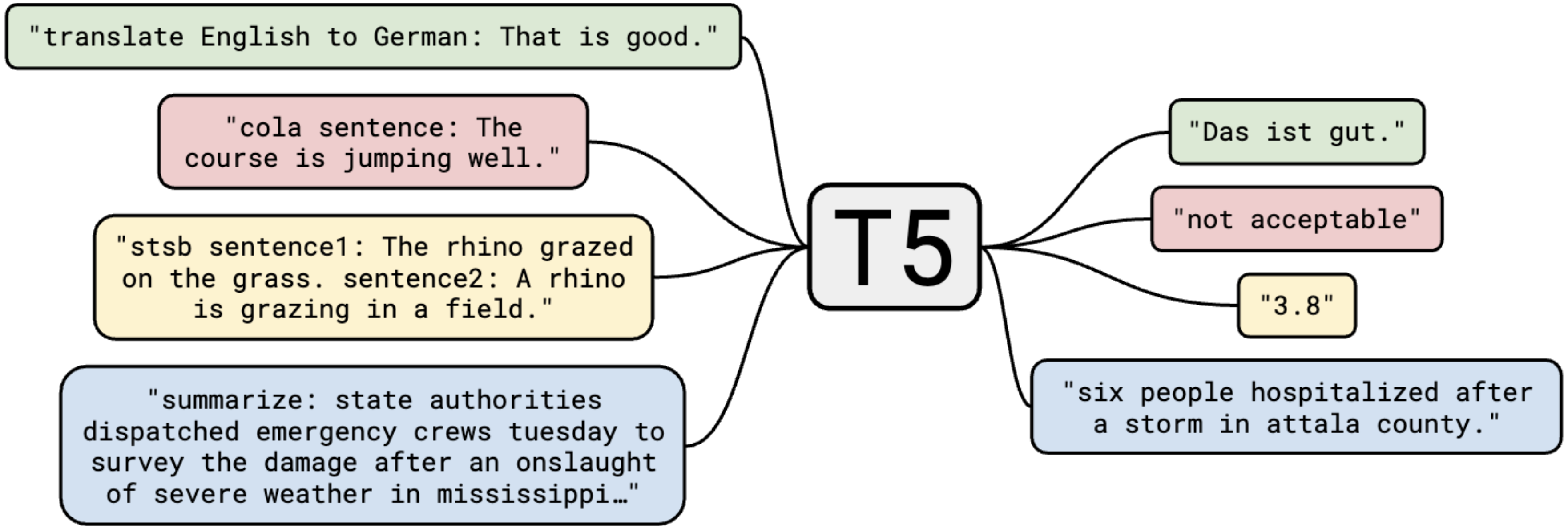
Processed target: "Luigi sagte oft zu mir, dass er nie wollte, dass die Brüder vor Gericht landen", schrieb sie.

Relative Position



Fine-Tuning: Text-to-Text For Everything





Promising Results

Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	74.8[c]	**90.7**[b]	91.3[a]	91.0[a]	**99.2**[a]	89.2[a]	91.8[a]
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.6	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	91.2	96.3	91.1	89.7
T5-11B	**75.1**	90.6	**92.2**	**91.9**	96.9	**92.8**	**94.5**

Model	SQuAD EM	SQuAD F1	SuperGLUE Average	BoolQ Accuracy	CB F1	CB Accuracy	COPA Accuracy
Previous best	90.1[a]	95.5[a]	84.6[d]	87.1[d]	90.5[d]	95.2[d]	90.6[d]
T5-Small	79.10	87.24	63.3	76.4	56.9	81.6	46.0
T5-Base	85.44	92.08	76.2	81.4	86.2	94.0	71.2
T5-Large	86.66	93.79	82.3	85.4	91.6	94.8	83.4
T5-3B	88.53	94.95	86.4	89.9	90.3	94.4	92.0
T5-11B	**91.26**	**96.22**	**88.9**	**91.2**	**93.9**	**96.8**	**94.8**

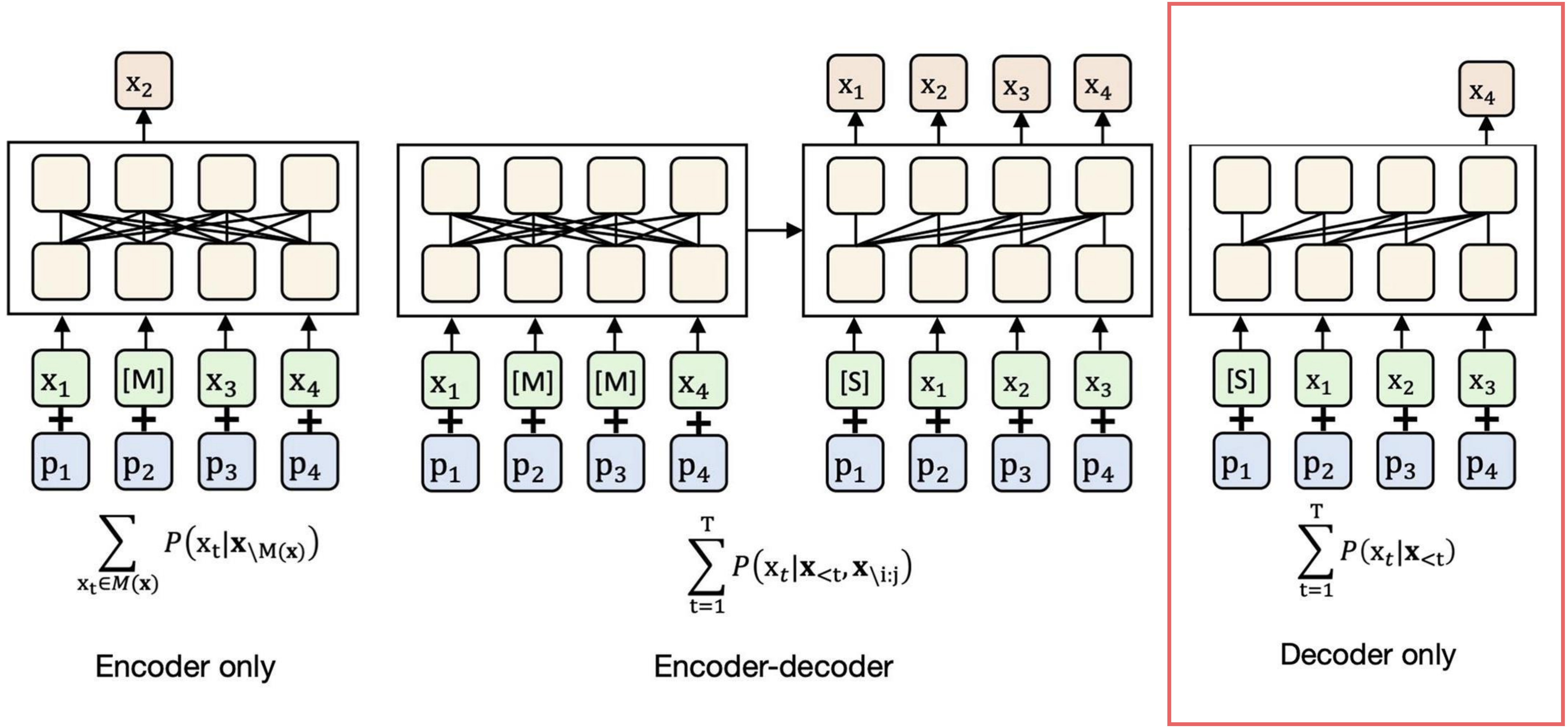
Model	MultiRC F1a	MultiRC EM	ReCoRD F1	ReCoRD Accuracy	RTE Accuracy	WiC Accuracy	WSC Accuracy
Previous best	84.4[d]	52.5[d]	90.6[d]	90.0[d]	88.2[d]	69.9[d]	89.0[d]
T5-Small	69.3	26.3	56.3	55.4	73.3	66.9	70.5
T5-Base	79.7	43.1	75.0	74.2	81.5	68.3	80.8
T5-Large	83.3	50.7	86.8	85.9	87.8	69.3	86.3
T5-3B	86.8	58.3	91.2	90.4	90.7	72.1	90.4
T5-11B	**88.1**	**63.3**	**94.1**	**93.4**	**92.5**	**76.9**	**93.8**

Use T5

Hugging Face

- T5-small:
  - # parameters ≈ 60M
- T5-base:
  - # parameters ≈ 220M
- T5-large:
  - # parameters ≈ 770M
- T5-3B: #
  - parameters ≈ 3B
- T5-11B:
  - # parameters ≈ 11B

Types of Pre-Training



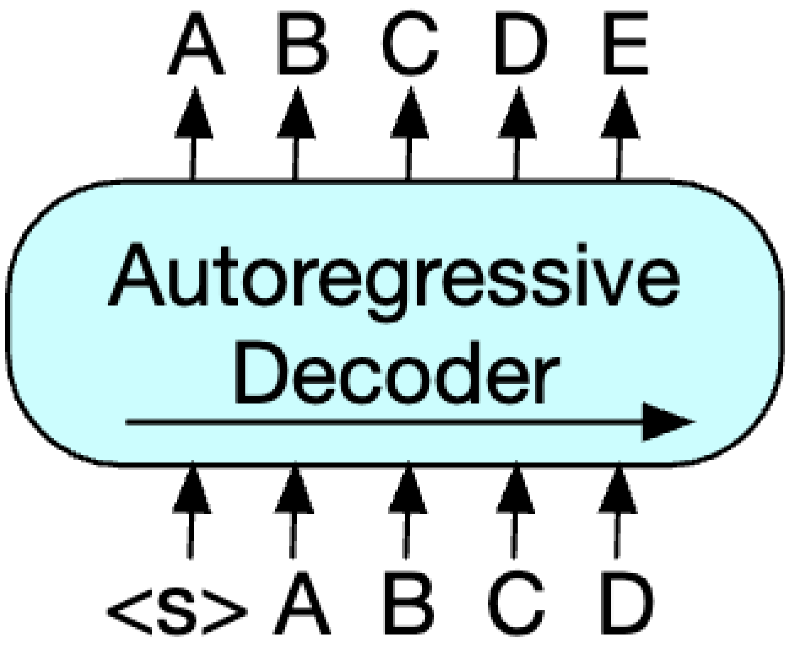
https://www.sciencedirect.com/science/article/pii/S2095809922006324

Decoder-Only: GPT

- Improving Language Understanding by Generative Pre-Training, OpenAI 2018
  - **G**enerative **P**re-trained **T**ransformer (GPT)
- Language Models are Unsupervised Multitask Learners, OpenAI 2019
  - GPT-2
- Language Models are Few-Shot Learners, OpenAI 2020
  - GPT-3

Language Modeling

- Next word prediction
- Trained with large corpus

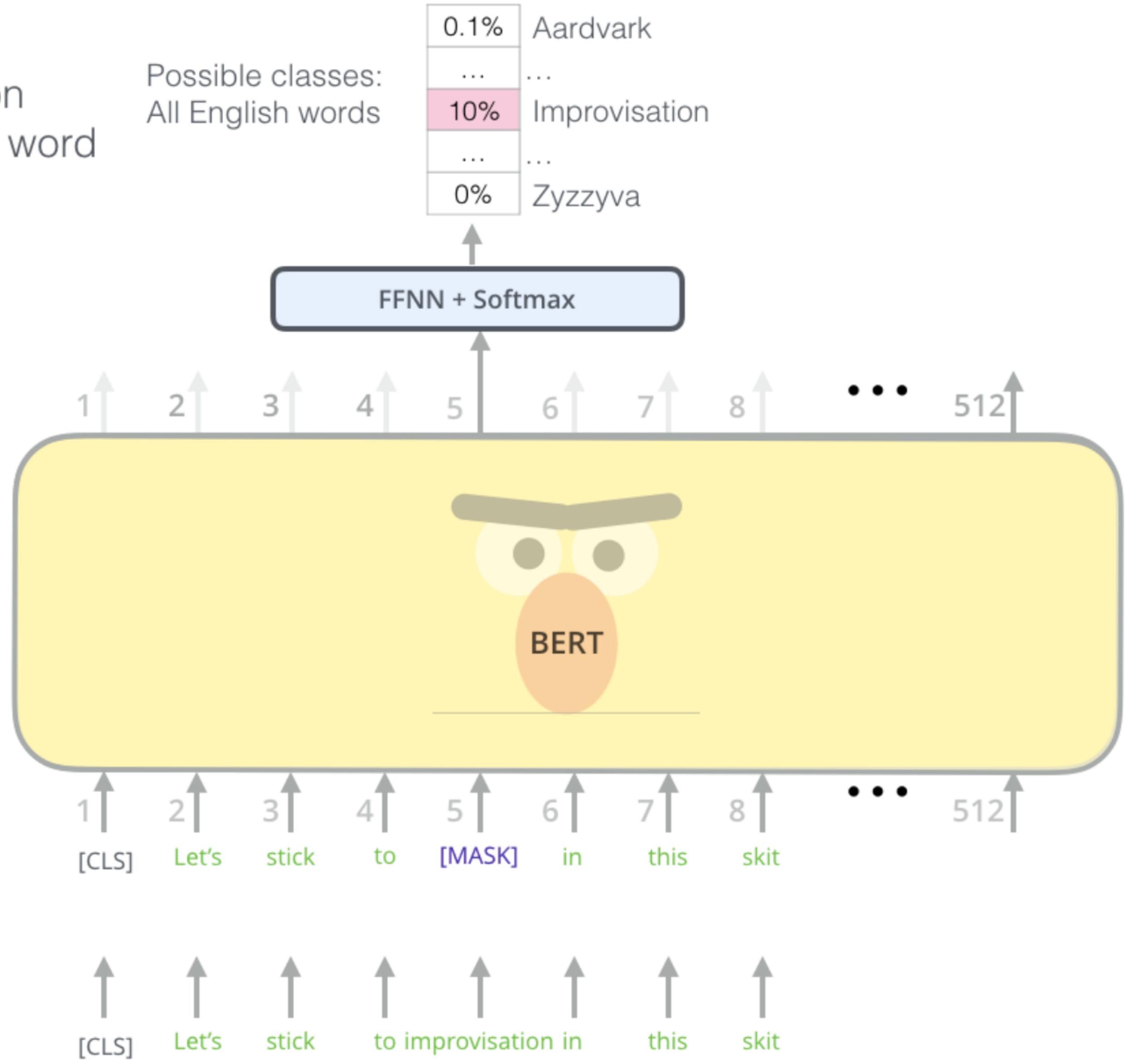


Comparison: Masked Language Models

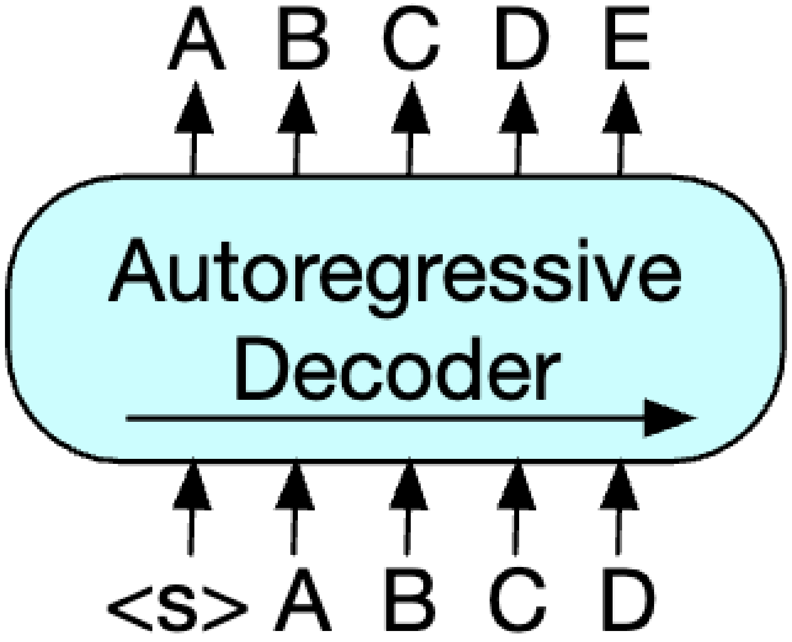
Use the output of the masked word's position to predict the masked word

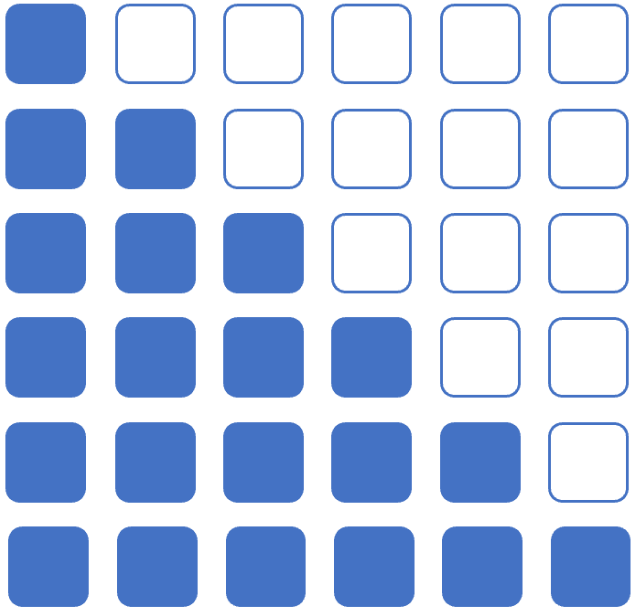
Randomly mask 15% of tokens

Input



Comparison: Causal Language Models

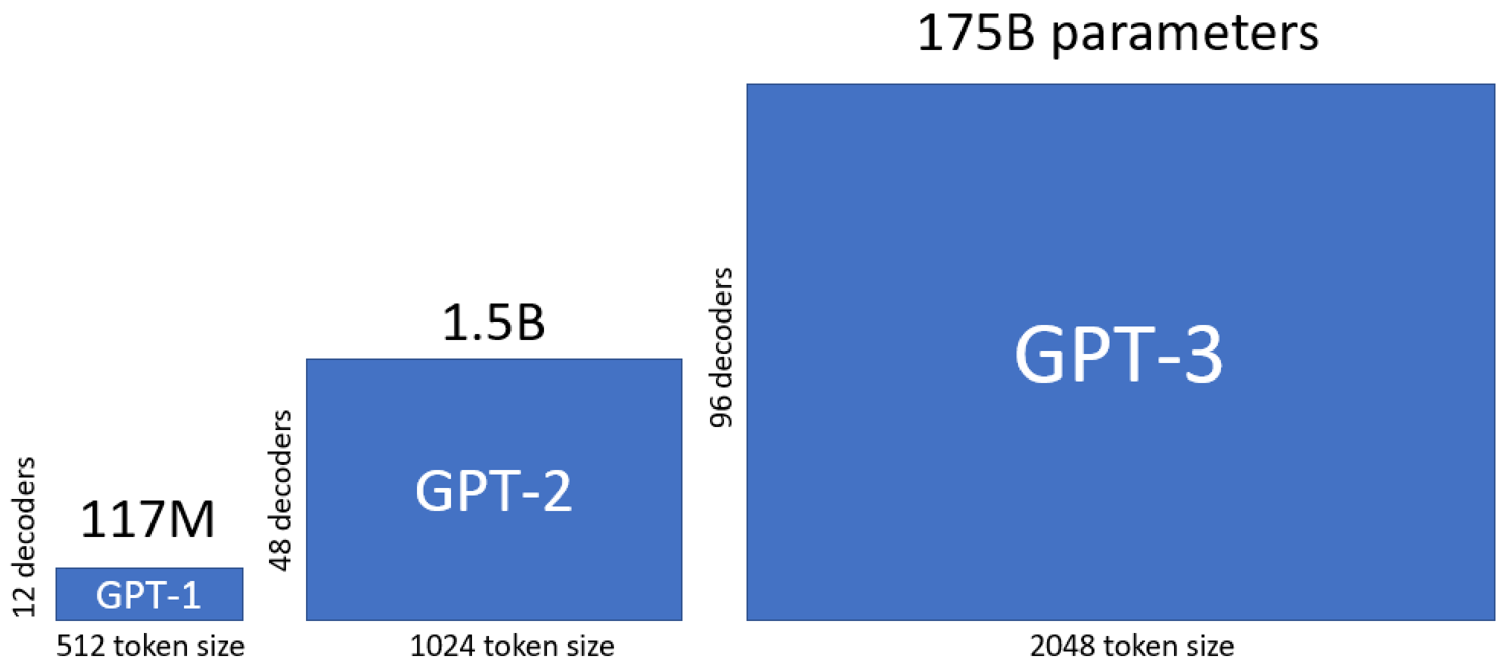




Causal Masking

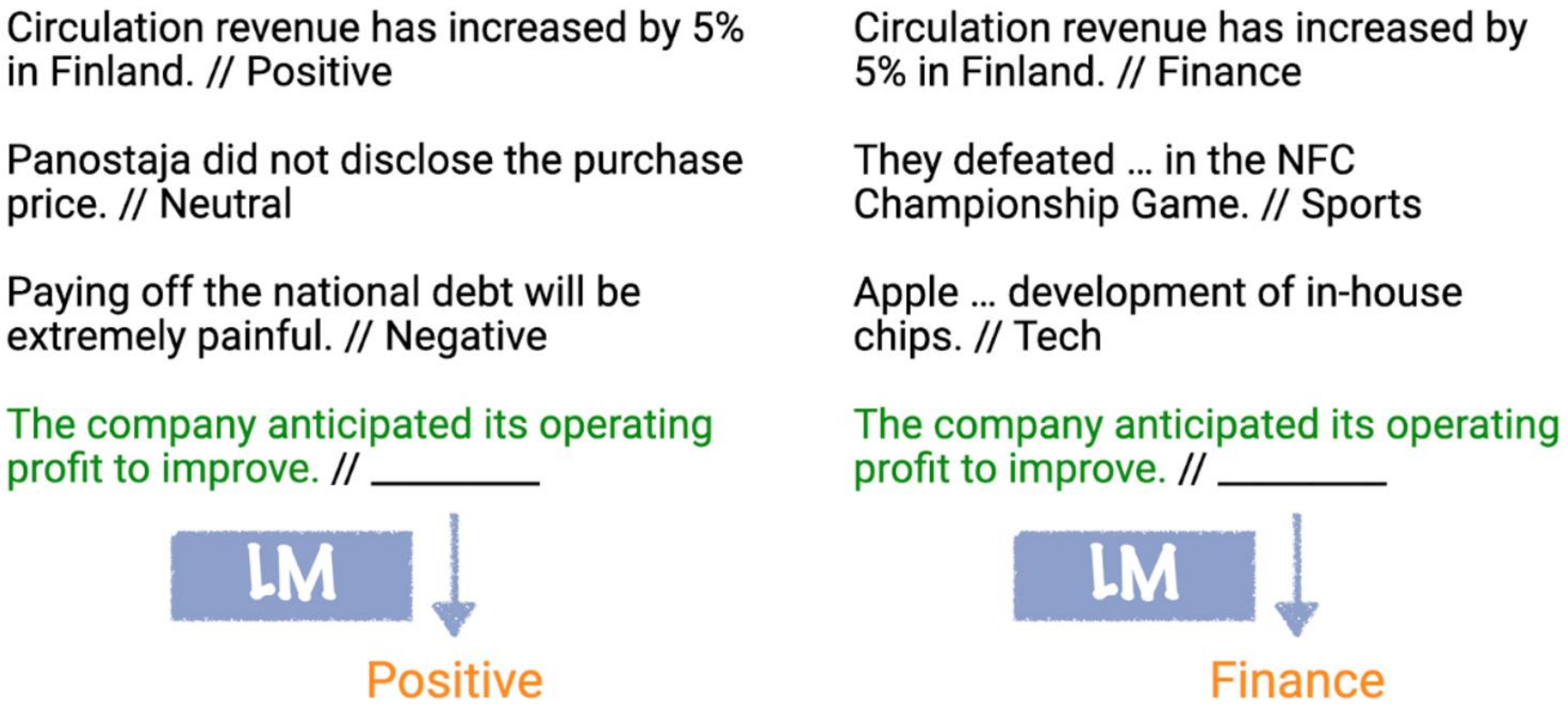
GPT-3: From Fine-Tuning to Few-Shot Learning

- Even larger training data, even larger model size

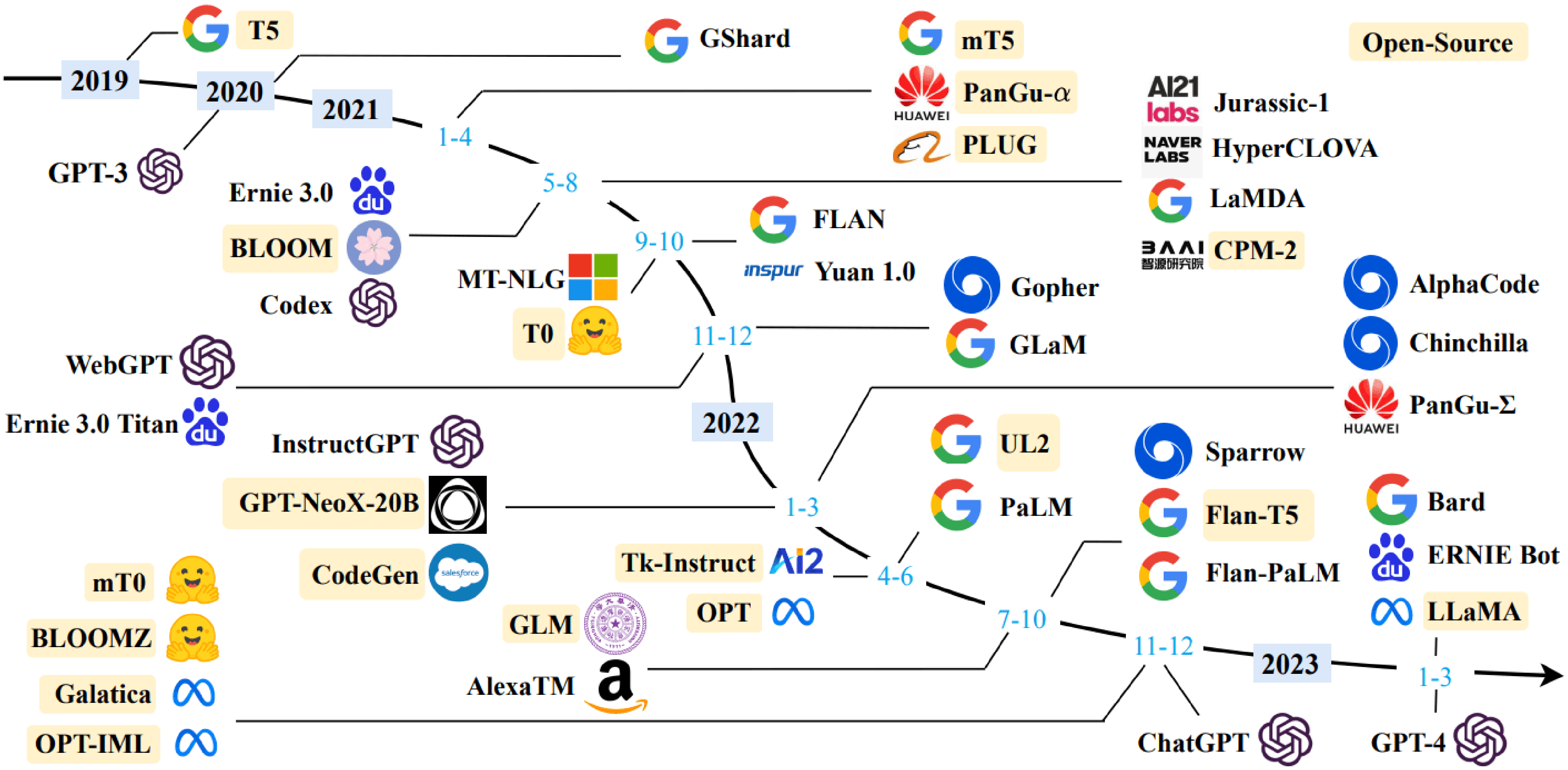


GPT-3: From Fine-Tuning to Few-Shot Learning

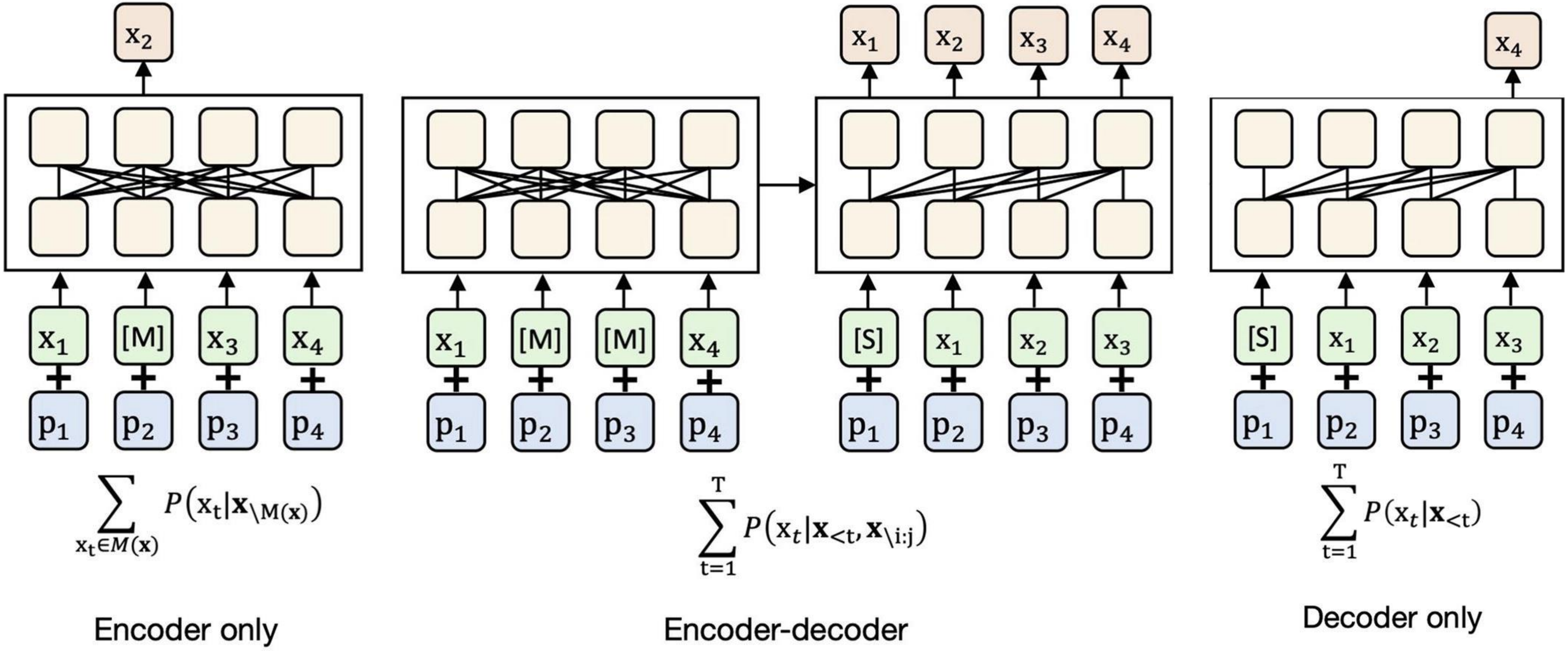
- Solve entirely new tasks by few-shot learning (in-context learning)



Large Language Models



Types of Pre-Training



https://www.sciencedirect.com/science/article/pii/S2095809922006324

Use GPT

Hugging Face

- GPT-2-small
  - # parameters ≈ 117M
- GPT-2-medium
  - # parameters ≈ 345M
- GPT-2-large
  - # parameters ≈ 762M
- GPT-2-xl
  - # parameters ≈ 1.5B

Lecture Plan

- Pre-Training
  - Encoder-Only Pre-Training
  - Encoder-Decoder Pre-Training
  - Decoder-Only Pre-Training
- Model Distillation

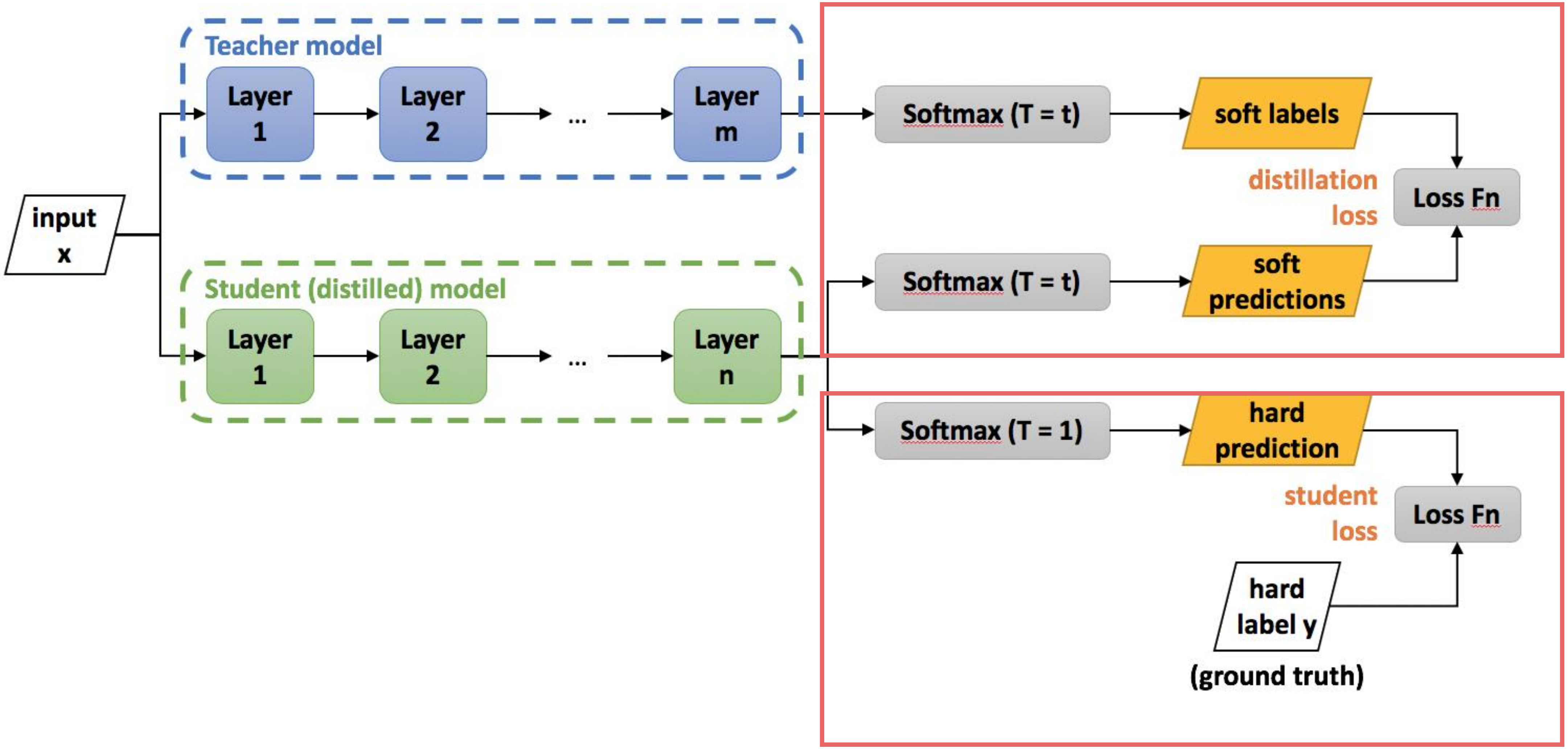
Model Distillation

- Distill knowledge from a large model to a small model while maintaining similar capability
  - **Large model:** teacher model
  - **Small model:** student model
- Train a student model to mimic the behavior of the teacher model
- Reduce the number of parameters

Why don’t we train a student model directly from data?

Model Distillation

Mimic teacher's behavior

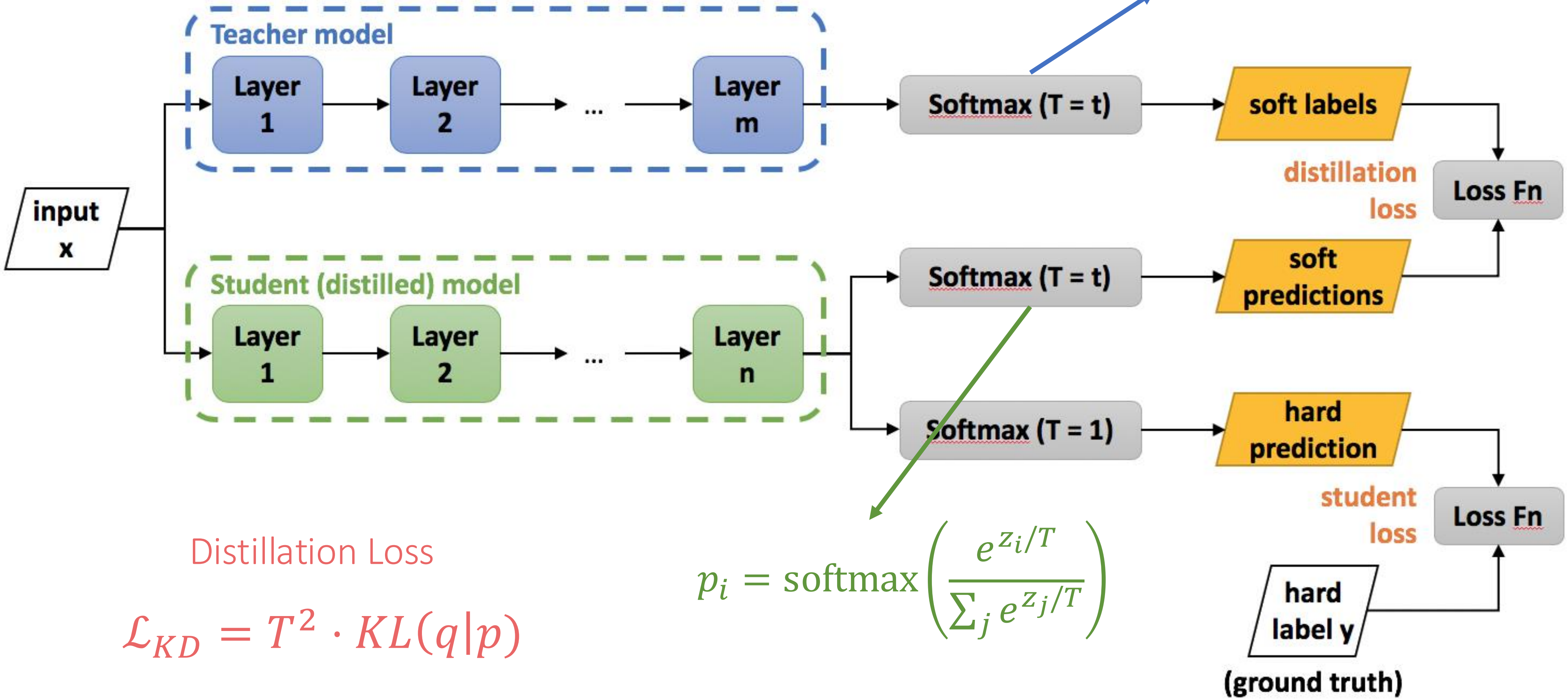


Learn from data

https://onedreame.github.io/2020/12/10/%E6%A8%A1%E5%9E%8B%E8%92%B8%E9%A6%8F%E6%8E%A2%E7%B4%A2(Bert)/

Model Distillation

$$q_i = \text{softmax}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}}\right)$$



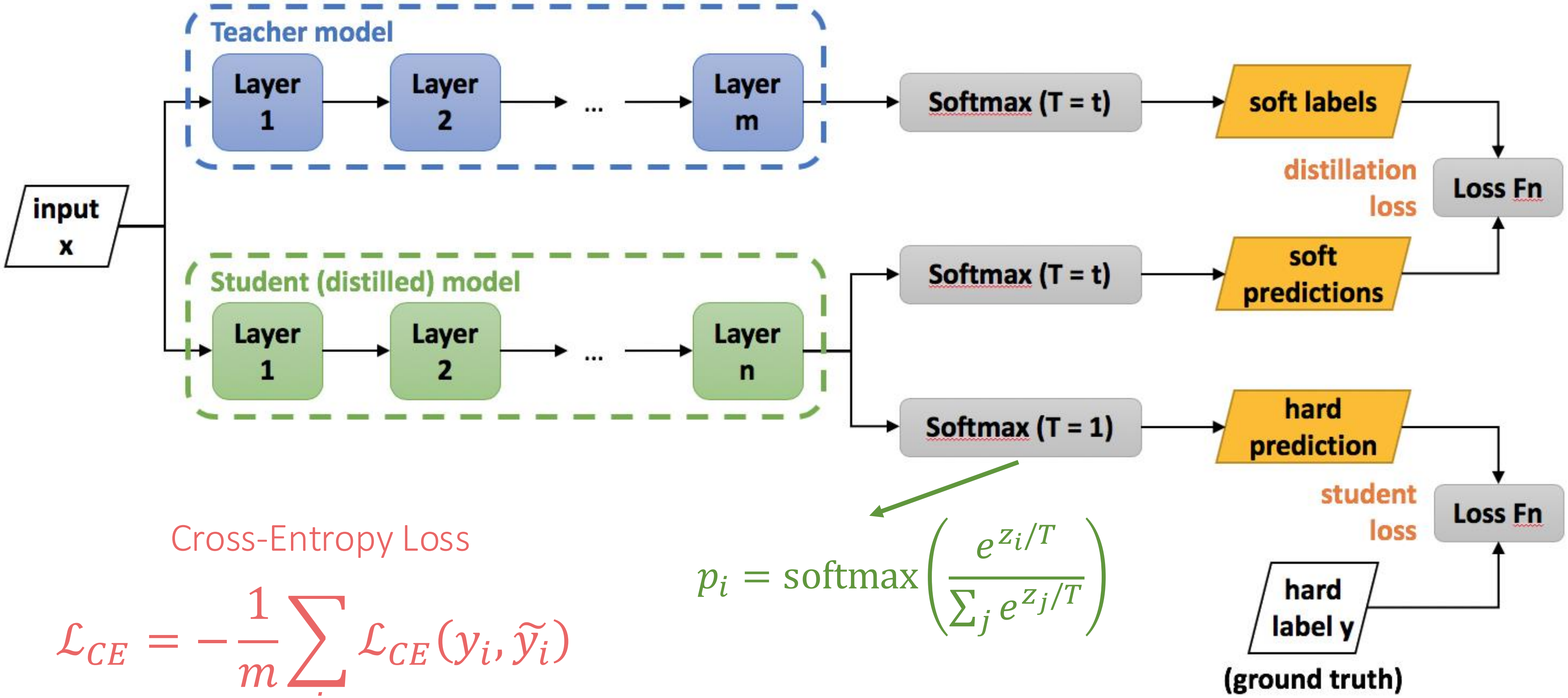
Distillation Loss

$$\mathcal{L}_{KD} = T^2 \cdot KL(q|p)$$

$$p_i = \text{softmax}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}}\right)$$

https://onedreame.github.io/2020/12/10/%E6%A8%A1%E5%9E%8B%E8%92%B8%E9%A6%8F%E6%8E%A2%E7%B4%A2(Bert)/

Model Distillation



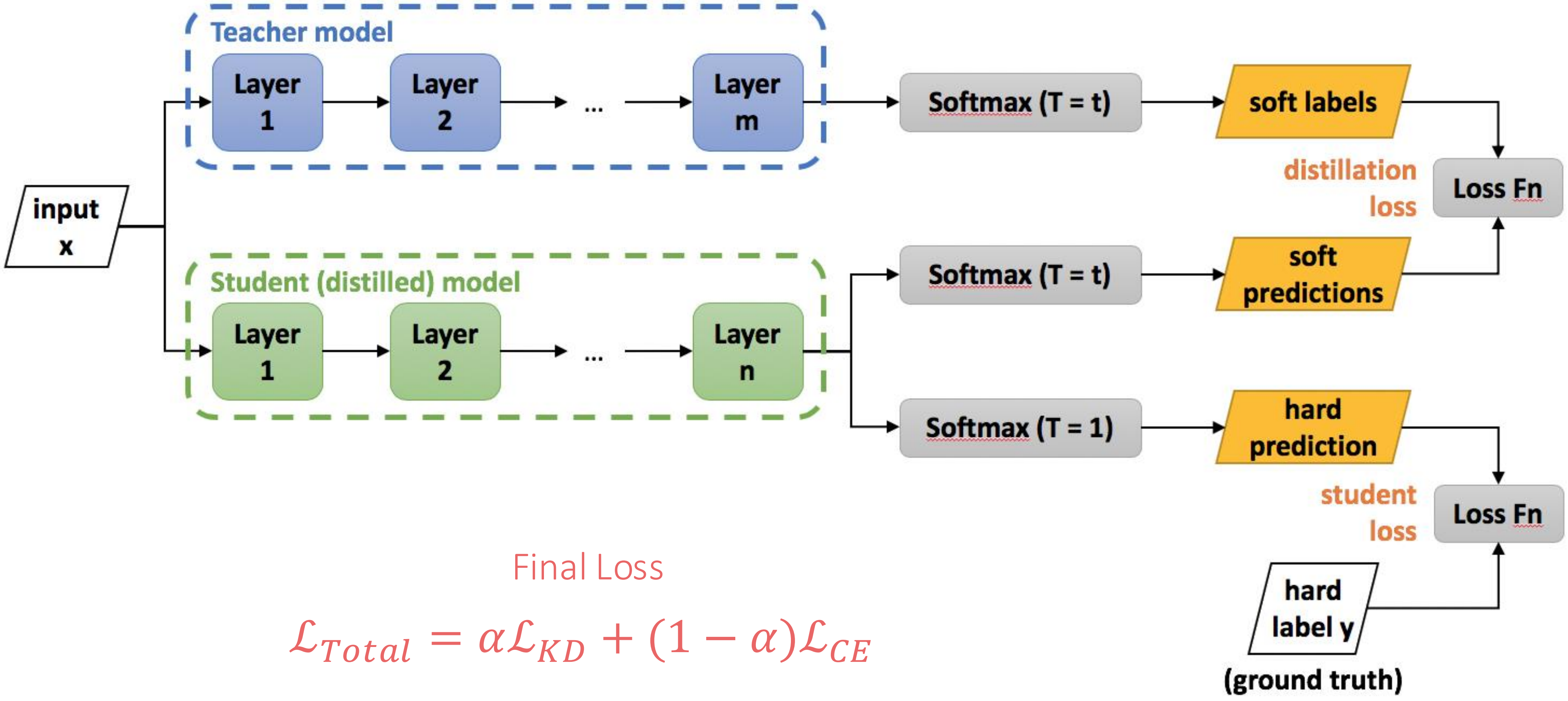
$$p_i = \mathrm{softmax}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}}\right)$$

Cross-Entropy Loss

$$\mathcal{L}_{CE} = -\frac{1}{m}\sum_i \mathcal{L}_{CE}(y_i, \tilde{y}_i)$$

https://onedreame.github.io/2020/12/10/%E6%A8%A1%E5%9E%8B%E8%92%B8%E9%A6%8F%E6%8E%A2%E7%B4%A2(Bert)/

Model Distillation



Final Loss

$$\mathcal{L}_{Total} = \alpha\mathcal{L}_{KD} + (1-\alpha)\mathcal{L}_{CE}$$

https://onedreame.github.io/2020/12/10/%E6%A8%A1%E5%9E%8B%E8%92%B8%E9%A6%8F%E6%8E%A2%E7%B4%A2(Bert)/

DistilBERT

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF
Hugging Face
{victor,lysandre,julien,thomas}@huggingface.co

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020

DistilBERT

Smaller Size

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410

- BERT-base
  - 12 layers, hidden size = 768, 12 attention heads
- DistilBERT
  - 6 layers, hidden size = 768, 12 attention heads

Almost similar performance

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

MobileBERT

MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices

Zhiqing Sun[1]*, Hongkun Yu[2], Xiaodan Song[2], Renjie Liu[2], Yiming Yang[1], Denny Zhou[2]

[1]Carnegie Mellon University {zhiqings, yiming}@cs.cmu.edu
[2]Google Brain {hongkuny, xiaodansong, renjieliu, dennyzhou}@google.com

MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices, 2020

MobileBERT

- Instead of less layers, reduce the hidden size

			$\text{BERT}_{\text{LARGE}}$	$\text{BERT}_{\text{BASE}}$	$\text{IB-BERT}_{\text{LARGE}}$	MobileBERT	$\text{MobileBERT}_{\text{TINY}}$
embedding		$h_{\text{embedding}}$	1024	768	128		
			no-op	no-op	3-convolution		
		h_{inter}	1024	768	512		
body	Linear	h_{input}			512	512	512
		h_{output}			1024	128	128
	MHA	h_{input}	1024	768	512	512	128
		#Head	16	12	4	4	4
		h_{output}	1024	768	1024	128	128
	FFN	h_{input}	1024	768	1024	128 ($\times 4$)	128 ($\times 2$)
		h_{FFN}	4096	3072	4096	512	512
		h_{output}	1024	768	1024	128	128
	Linear	h_{input}			1024	128	128
		h_{output}			512	512	512
			$\times 24$	$\times 12$	$\times 24$	$\times 24$	$\times 24$
#Params			334M	109M	293M	25.3M	15.1M

MobileBERT

	#Params	#FLOPS	Latency	CoLA 8.5k	SST-2 67k	MRPC 3.7k	STS-B 5.7k	QQP 364k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	GLUE
ELMo-BiLSTM-Attn	-	-	-	33.6	90.4	84.4	72.3	63.1	74.1/74.5	79.8	58.9	70.0
OpenAI GPT	109M	-	-	47.2	93.1	87.7	84.8	70.1	80.7/80.6	87.2	69.1	76.9
$\text{BERT}_{\text{BASE}}$	109M	22.5B	342 ms	**52.1**	**93.5**	**88.9**	**85.8**	71.2	**84.6/83.4**	90.5	66.4	78.3
$\text{BERT}_{\text{BASE}}$-6L-PKD*	66.5M	11.3B	-	-	92.0	85.0	-	70.7	81.5/81.0	89.0	65.5	-
$\text{BERT}_{\text{BASE}}$-4L-PKD†*	52.2M	7.6B	-	24.8	89.4	82.6	79.8	70.2	79.9/79.3	85.1	62.3	-
$\text{BERT}_{\text{BASE}}$-3L-PKD*	45.3M	5.7B	-	-	87.5	80.7	-	68.1	76.7/76.3	84.7	58.2	-
$\text{DistilBERT}_{\text{BASE}}$-6L†	62.2M	11.3B	-	-	92.0	85.0		70.7	81.5/81.0	89.0	65.5	-
$\text{DistilBERT}_{\text{BASE}}$-4L†	52.2M	7.6B	-	32.8	91.4	82.4	76.1	68.5	78.9/78.0	85.2	54.1	-
TinyBERT*	14.5M	1.2B	-	43.3	92.6	86.4	79.9	**71.3**	82.5/81.8	87.7	62.9	75.4
$\text{MobileBERT}_{\text{TINY}}$	15.1M	3.1B	40 ms	46.7	91.7	87.9	80.1	68.9	81.5/81.6	89.5	65.1	75.8
MobileBERT	25.3M	5.7B	62 ms	50.5	92.8	88.8	84.4	70.2	83.3/82.6	90.6	66.2	77.7
MobileBERT w/o OPT	25.3M	5.7B	192 ms	51.1	92.6	88.8	84.8	70.5	84.3/**83.4**	**91.6**	**70.4**	**78.5**