

Improving Language Understanding by Generative Pre-Training

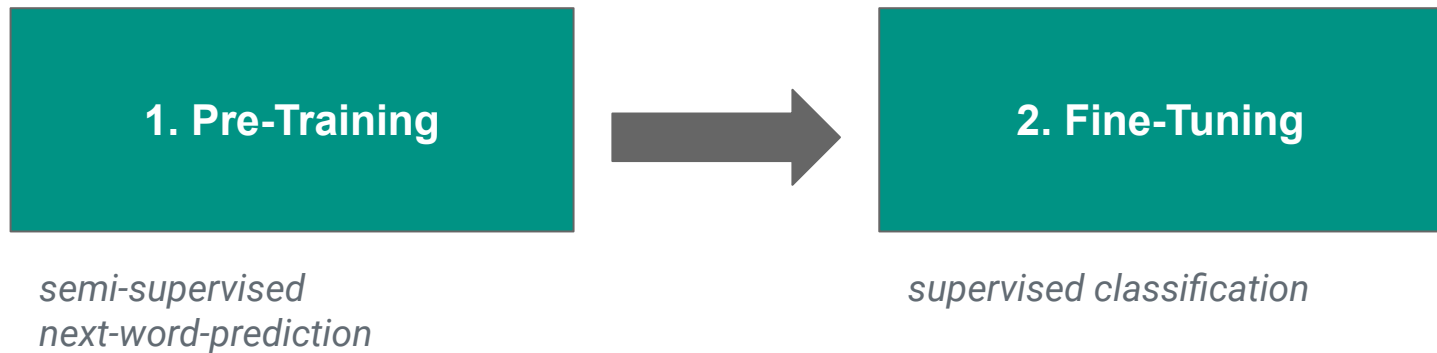
Slightly ascii-related
Paper Reading Group - II

+ Transformer = GPT, übrigens

Previously...

- Transformer Architecture is cool, because trains fast and models long dependencies
- Encoder + Decoder
- Multi-Head Self-Attention
- No Pre-training: one task = one model
- Task: Translation German/French \leftrightarrow English

Now...



Language Modelling

$$P(u_3 | u_0, u_1, u_2; \Theta)$$

Probability to predict word/ token u_3

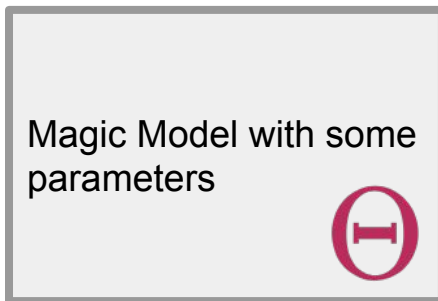
context of words before

model parameters

context words

I want a

u_0, u_1, u_2



$\begin{bmatrix} a \\ ah \\ \dots \\ cat \\ \dots \\ zoo \end{bmatrix}$

$P(u_3 | u_0, u_1, u_2; \Theta)$



next word

cat

Objective

Given an unsupervised corpus of tokens $\mathcal{U} = \{u_1, \dots, u_n\}$, we use a standard language modeling objective to maximize the following likelihood:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta) \quad (1)$$

where k is the size of the context window, and the conditional probability P is modeled using a neural network with parameters Θ . These parameters are trained using stochastic gradient descent [51].

Language Modelling

We want find Parameters Θ , such that the probability of predicting the next word for the entire data corpus is maximized:

$$P(u_3|u_0, u_1, u_2; \Theta) \cdot P(u_4|u_1, u_2, u_3; \Theta) \cdot \dots \cdot P(u_n|u_{n-3}, u_{n-2}, u_{n-1}; \Theta)$$

The Θ that maximizes that
also maximizes this

$$\begin{aligned} & \log P(u_3|u_0, u_1, u_2; \Theta) + \log P(u_4|u_1, u_2, u_3; \Theta) + \log \dots + \log P(u_n|u_{n-3}, u_{n-2}, u_{n-1}; \Theta) \\ &= \\ & \sum_i \log P(u_i|u_{i-k}, \dots, u_{i-1}; \Theta) \end{aligned}$$

Objective

Given an unsupervised corpus of tokens $\mathcal{U} = \{u_1, \dots, u_n\}$, we use a standard language modeling objective to maximize the following likelihood:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta) \quad (1)$$

where k is the size of the context window, and the conditional probability P is modeled using a neural network with parameters Θ . These parameters are trained using stochastic gradient descent [51].

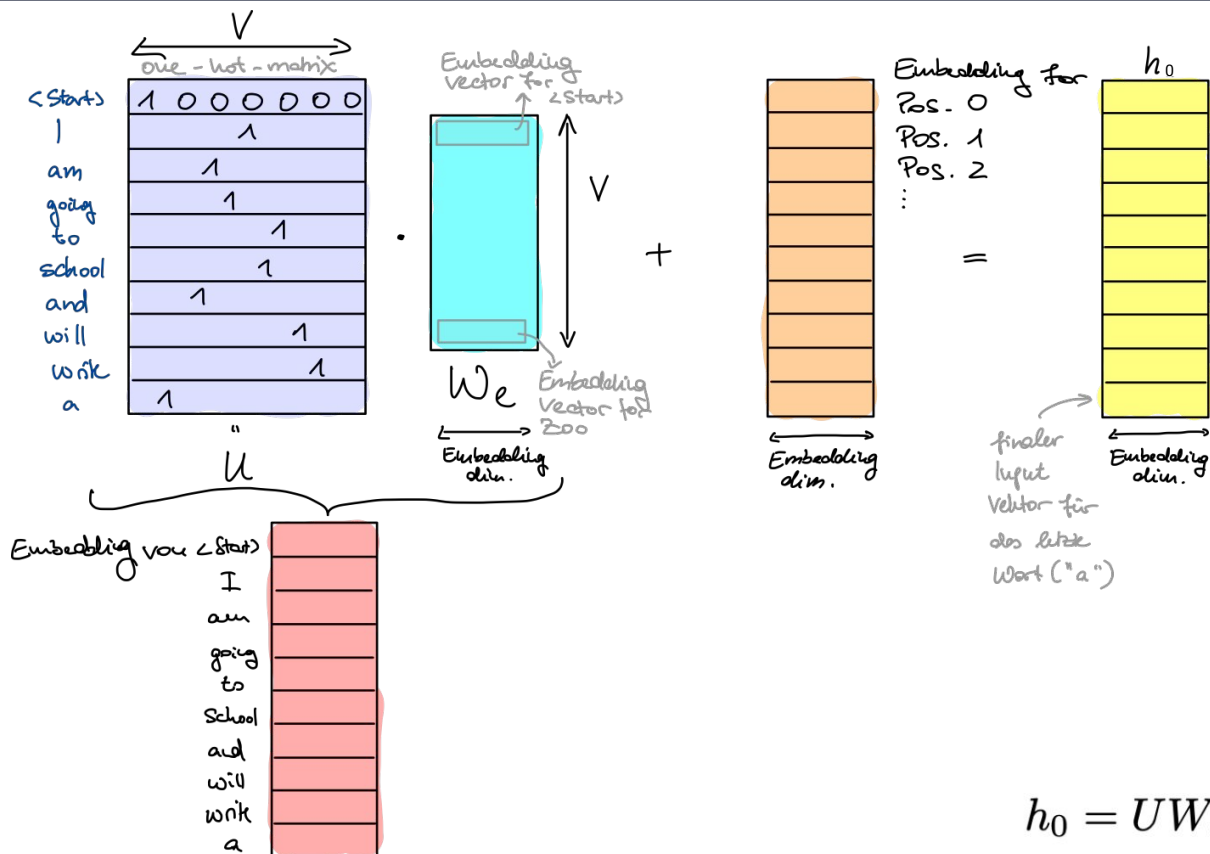
Implemented using Transformer Decoder

In our experiments, we use a multi-layer *Transformer decoder* [34] for the language model, which is a variant of the transformer [62]. This model applies a multi-headed self-attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens:

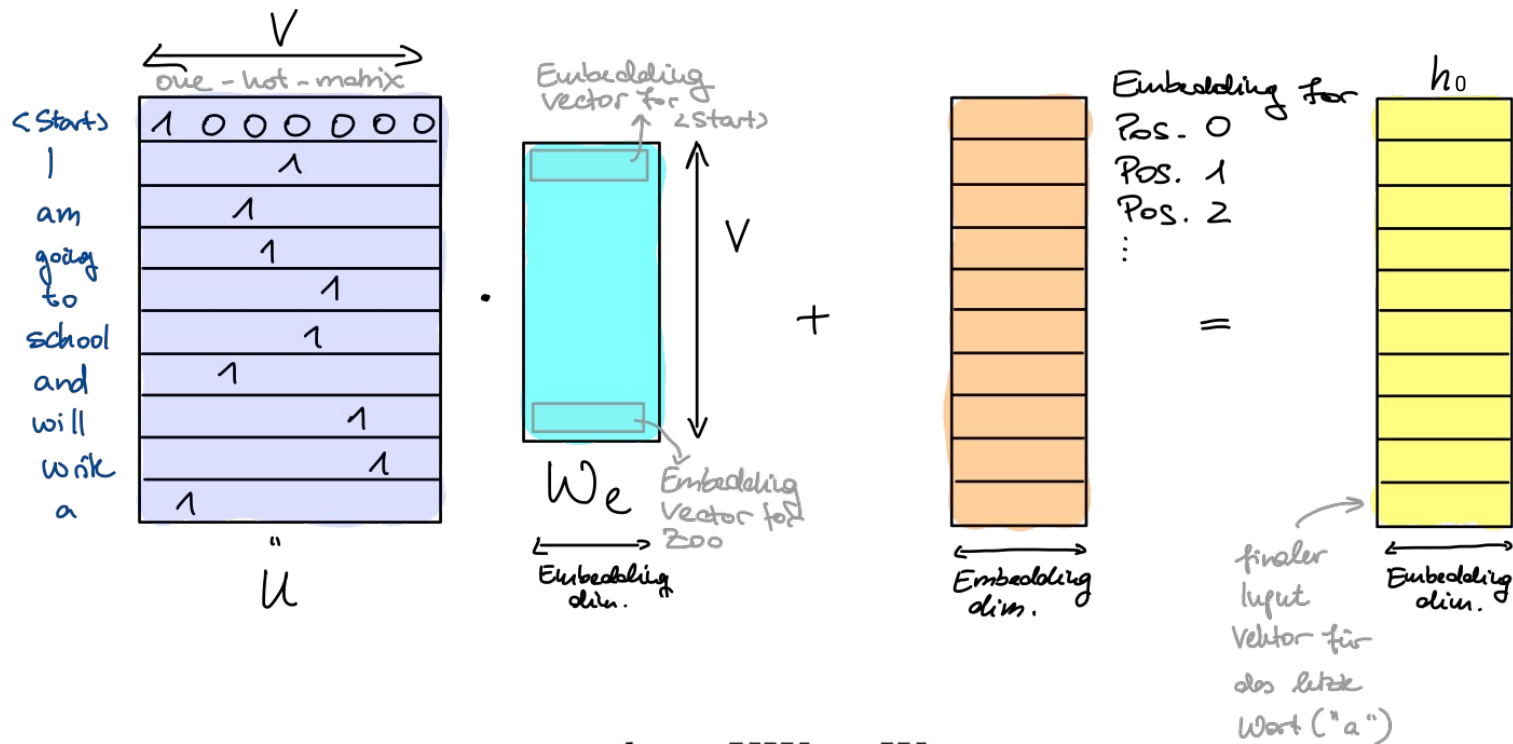
$$\begin{aligned} h_0 &= UW_e + W_p \\ h_l &= \text{transformer_block}(h_{l-1}) \forall i \in [1, n] \\ P(u) &= \text{softmax}(h_n W_e^T) \end{aligned} \tag{2}$$

where $U = (u_{-k}, \dots, u_{-1})$ is the context vector of tokens, n is the number of layers, W_e is the token embedding matrix, and W_p is the position embedding matrix.

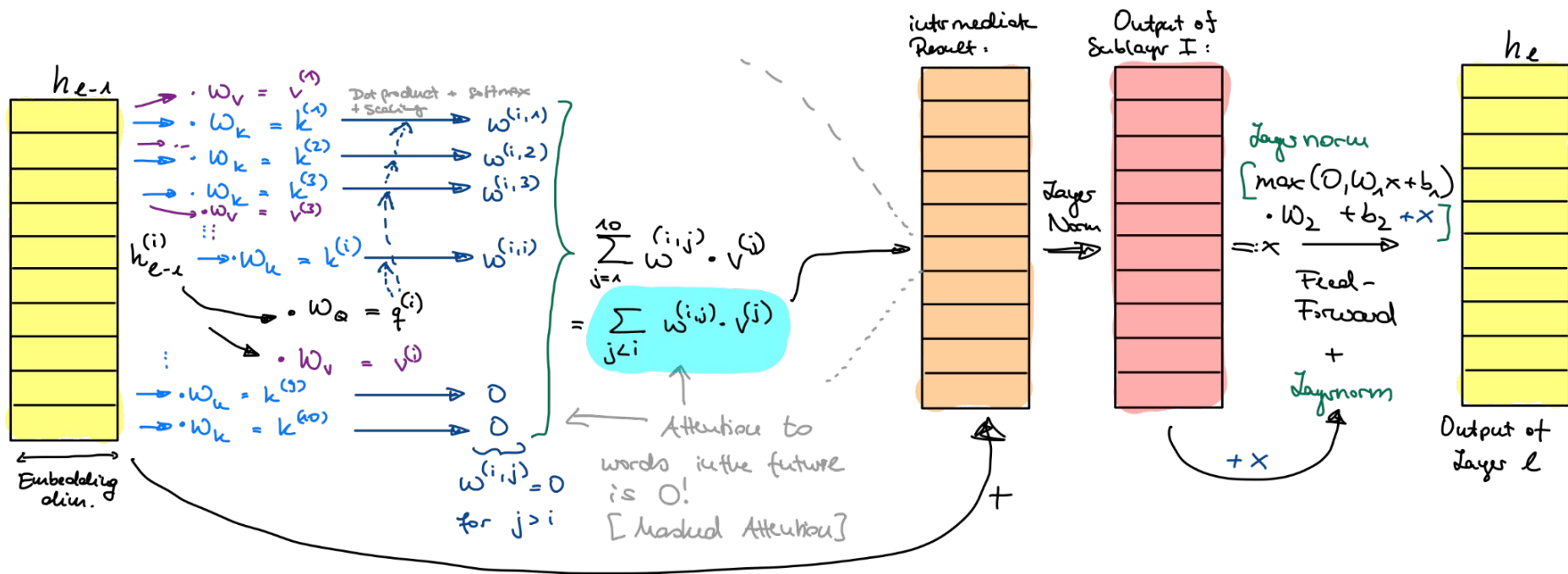
Embeddings



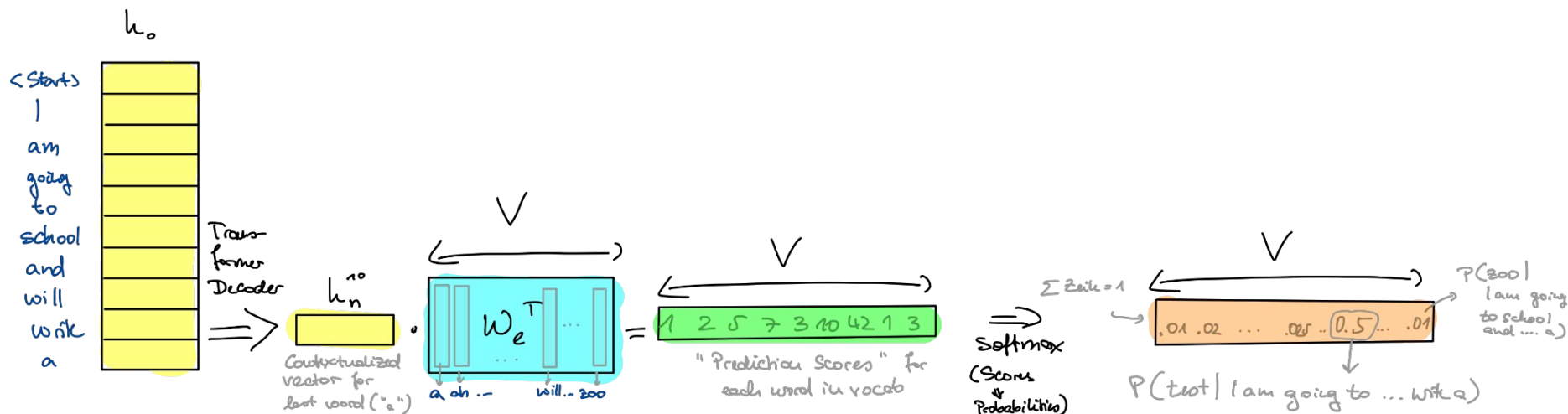
Embeddings



Input for Transformer



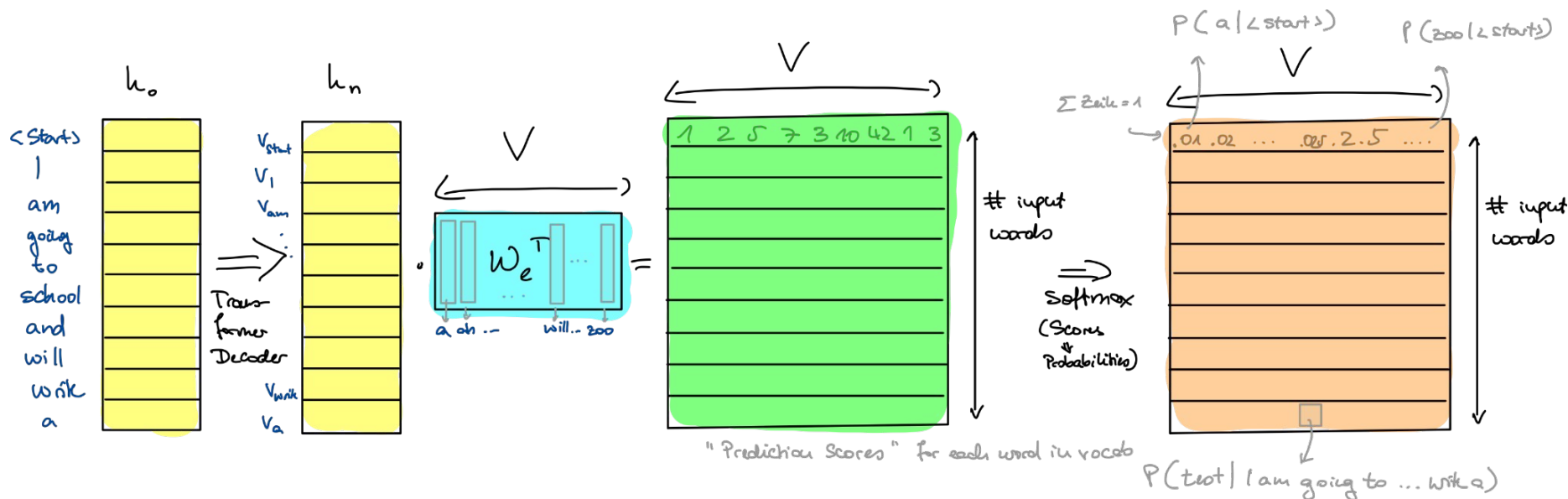
Predicting a Word from Context



$$h_l = \text{transformer_block}(h_{l-1}) \forall i \in [1, n]$$

$$P(u) = \text{softmax}(h_n W_e^T)$$

Predicting a Word from Context



Fine-Tuning

3.2 Supervised fine-tuning

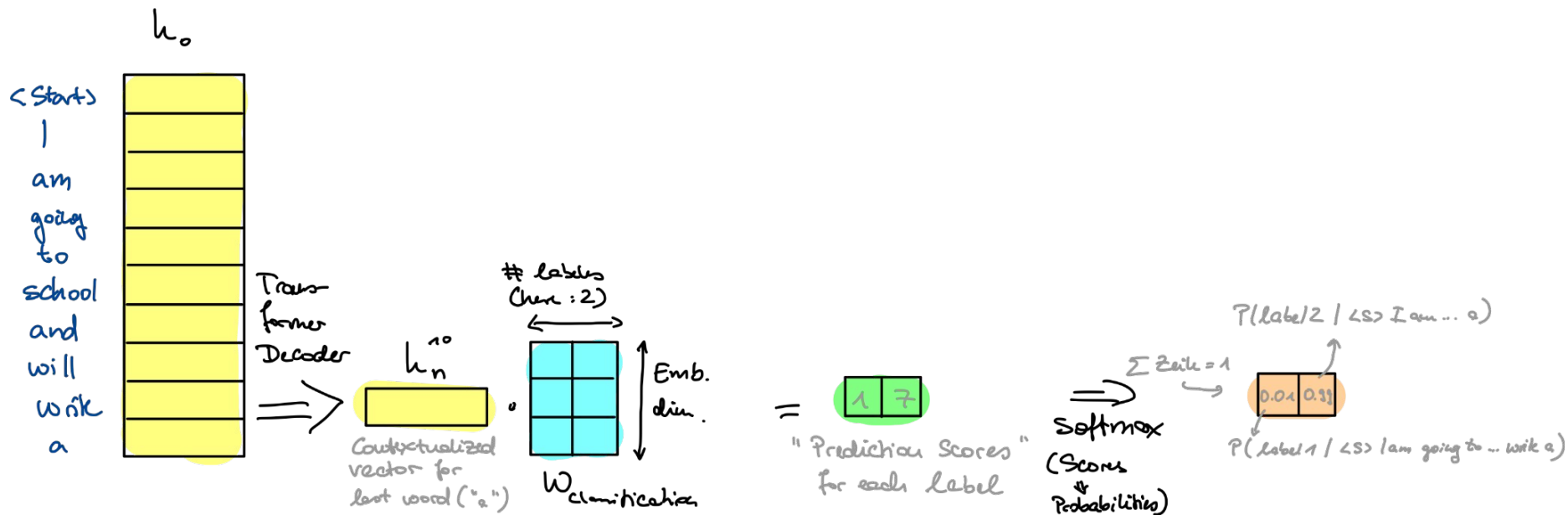
After training the model with the objective in Eq. 1, we adapt the parameters to the supervised target task. We assume a labeled dataset \mathcal{C} , where each instance consists of a sequence of input tokens, x^1, \dots, x^m , along with a label y . The inputs are passed through our pre-trained model to obtain the final transformer block's activation h_l^m , which is then fed into an added linear output layer with parameters W_y to predict y :

$$P(y|x^1, \dots, x^m) = \text{softmax}(h_l^m W_y). \quad (3)$$

This gives us the following objective to maximize:

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \dots, x^m). \quad (4)$$

Classification



Loss Combination

We additionally found that including language modeling as an auxiliary objective to the fine-tuning helped learning by (a) improving generalization of the supervised model, and (b) accelerating convergence. This is in line with prior work [50, 43], who also observed improved performance with such an auxiliary objective. Specifically, we optimize the following objective (with weight λ):

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C}) \quad (5)$$

Overall, the only extra parameters we require during fine-tuning are W_y , and embeddings for delimiter tokens (described below in Section 3.3).

Overview – Finetuning

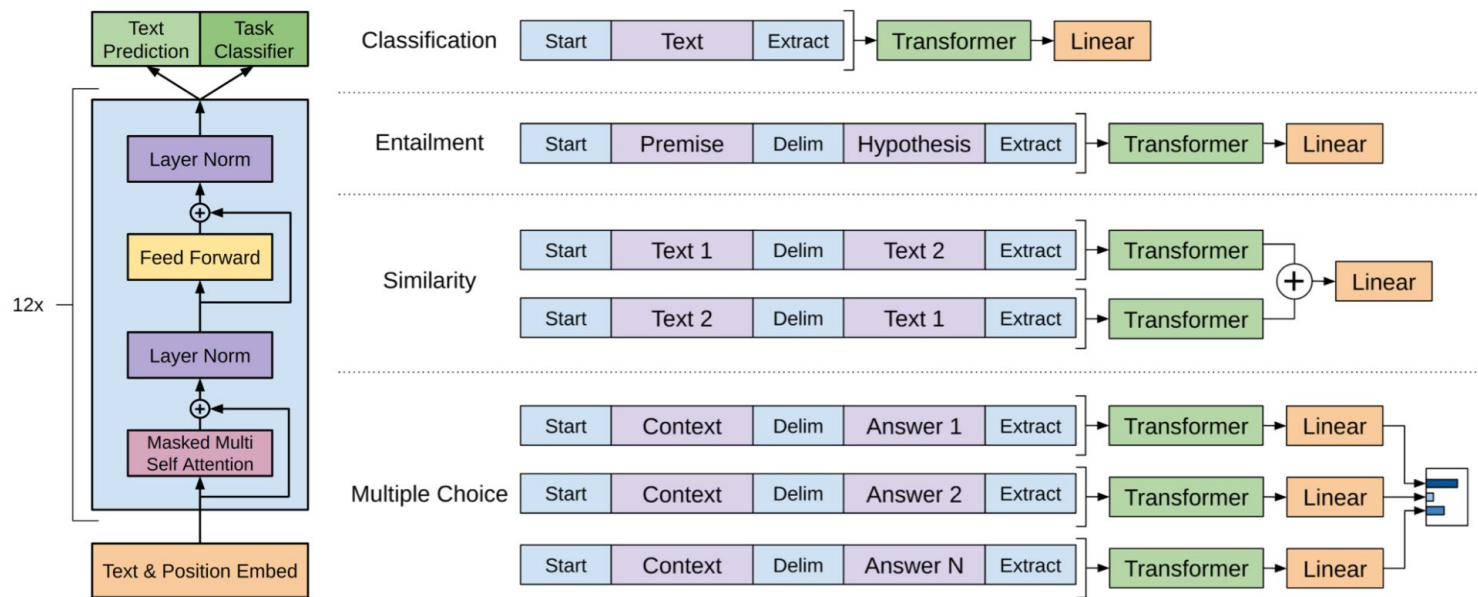


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Fine-tuning Tasks

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

Natural Language Inference (SNLI)

premise (string)	hypothesis (string)	label (class label)
"This church choir sings to the masses as they sing joyous songs from the boo..."	"The church has cracks in the ceiling."	1 (neutral)
"This church choir sings to the masses as they sing joyous songs from the boo..."	"The church is filled with song."	0 (entailment)
"This church choir sings to the masses as they sing joyous songs from the boo..."	"A choir singing at a baseball game."	2 (contradiction)
"A woman with a green headscarf, blue shirt and a very big grin."	"The woman is young."	1 (neutral)
"A woman with a green headscarf, blue shirt and a very big grin."	"The woman is very happy."	0 (entailment)

Question Answering (RACE)

Story

The rain had continued for a week and the flood had created a big river which were running by Nancy Brown's farm. **As she tried to gather her cows to a higher ground, she slipped and hit her head on a fallen tree trunk.** The fall made her unconscious for a moment or two. When she came to, Lizzie, one of her oldest and favorite cows, was licking her face. At that time, the water level on the farm was still rising. Nancy gathered all her strength to get up and began walking slowly with Lizzie. The rain had become much heavier, and the water in the field was now waist high. Nancy's pace got slower and slower because she felt a great pain in her head. Finally, all she could do was to throw her arm around Lizzie's neck and try to hang on. About 20 minutes later, Lizzie managed to pull herself and Nancy out of the rising water and onto a bit of high land, which seemed like a small island in the middle of a lake of white water. Even though it was about noon, the sky was so dark and the rain and lightning was so bad that it took rescuers more than two hours to discover Nancy. A man from a helicopter lowered a rope, but Nancy couldn't catch it. A moment later, two men landed on the small island from a ladder in the helicopter. They raised her into the helicopter and took her to the school gym, where the Red Cross had set up an emergency shelter. When the flood disappeared two days later, Nancy immediately went back to the "island." Lizzie was gone. She was one of 19 cows that Nancy had lost in the flood. "I owe my life to her," said Nancy with tears.

Question

"What did Nancy try to do before she fell over?"

Answer Options

["Measure the depth of the river",
"Look for a fallen tree trunk",
"Protect her cows from being drowned",
"Run away from the flooded farm"]

Sentence Similarity (STS)

split (string)	sentence1 (string)	sentence2 (string)	score (float64)
"train"	"But other sources close to the sale said Vivendi was keeping the door open to further bids and hoped to see bidders interested in individual assets team up."	"But other sources close to the sale said Vivendi was keeping the door open for further bids in the next day or two."	4
"train"	"Micron has declared its first quarterly profit for three years."	"Micron's numbers also marked the first quarterly profit in three..."	3.75
"train"	"The fines are part of failed Republican efforts to force or..."	"Perry said he backs the Senate's efforts, including the fines, to..."	2.8
"train"	"The American Anglican Council, which represents Episcopalian..."	"The American Anglican Council, which represents Episcopalian..."	3.4

Classification (SST2 und CoLA)

label (class label)	sentence (string)
1 (positive)	"A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror films"
0 (negative)	"Apparently reassembled from the cutting-room floor of any given daytime soap."

"These cars drive easily."	0 (acceptable)
"He washed yourself."	1 (unacceptable)
"Bill's wine from France and Ted's from California cannot be compared."	0 (acceptable)
"What the water did to the bottle was fill it."	1 (unacceptable)

General Language Understanding (GLUE)

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = Ungrammatical	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = .93056 (Very Positive)	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = A Paraphrase	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = 4.6 (Very Similar)	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = Not Similar	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = Contradiction	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = Answerable	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = Entailed	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = Incorrect Referent	Accuracy

Overview – Finetuning

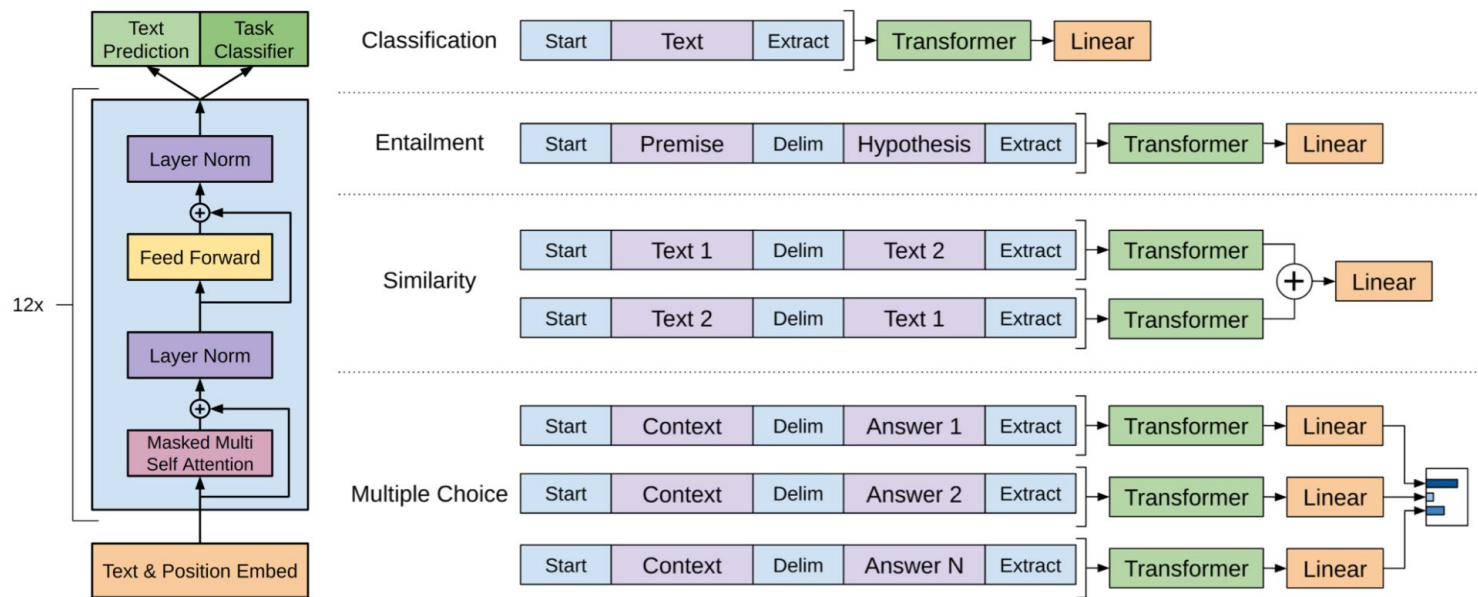
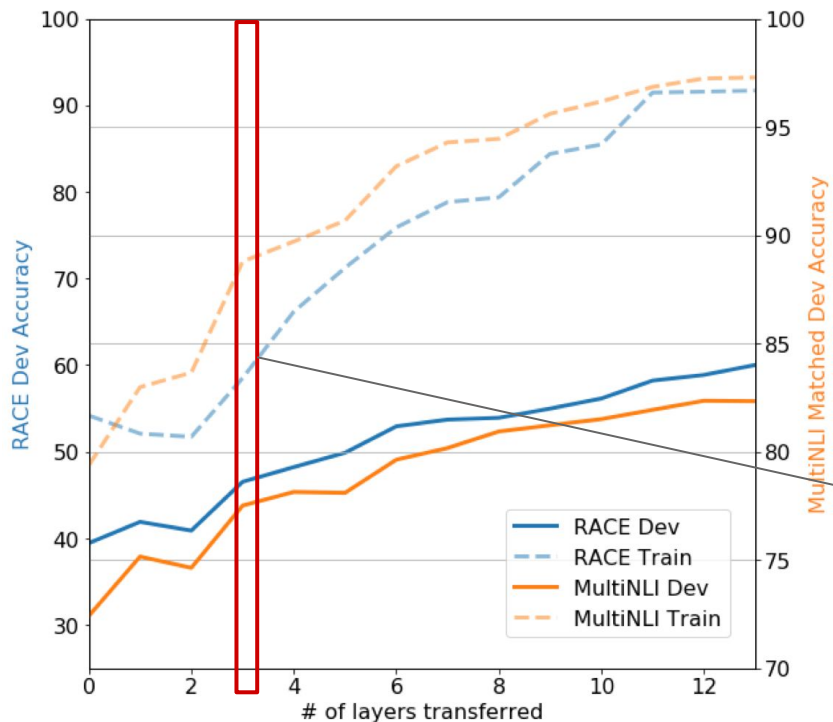
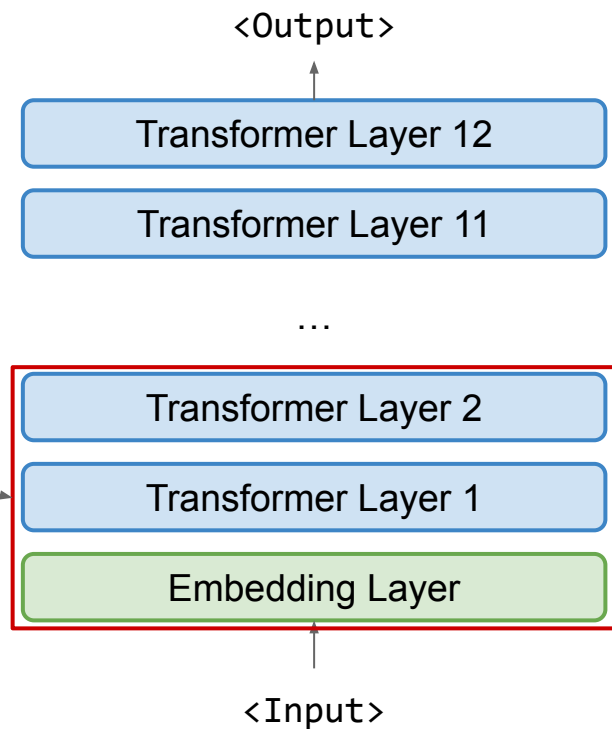


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

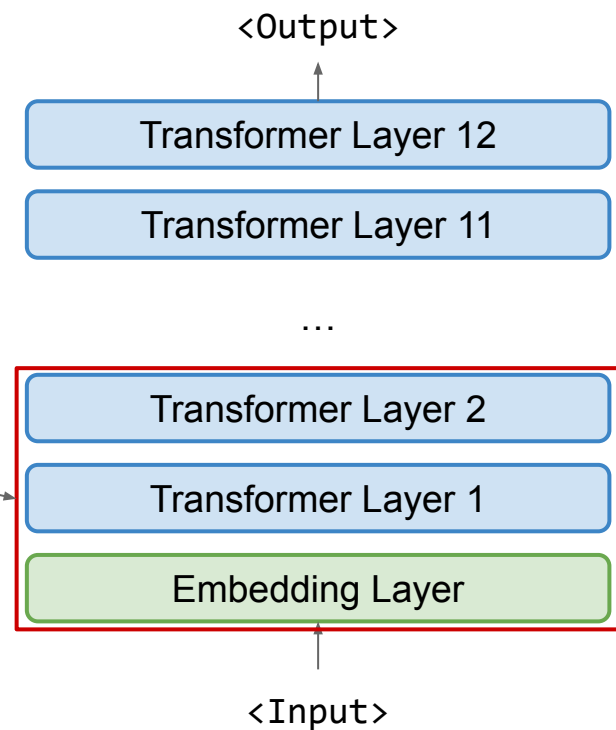
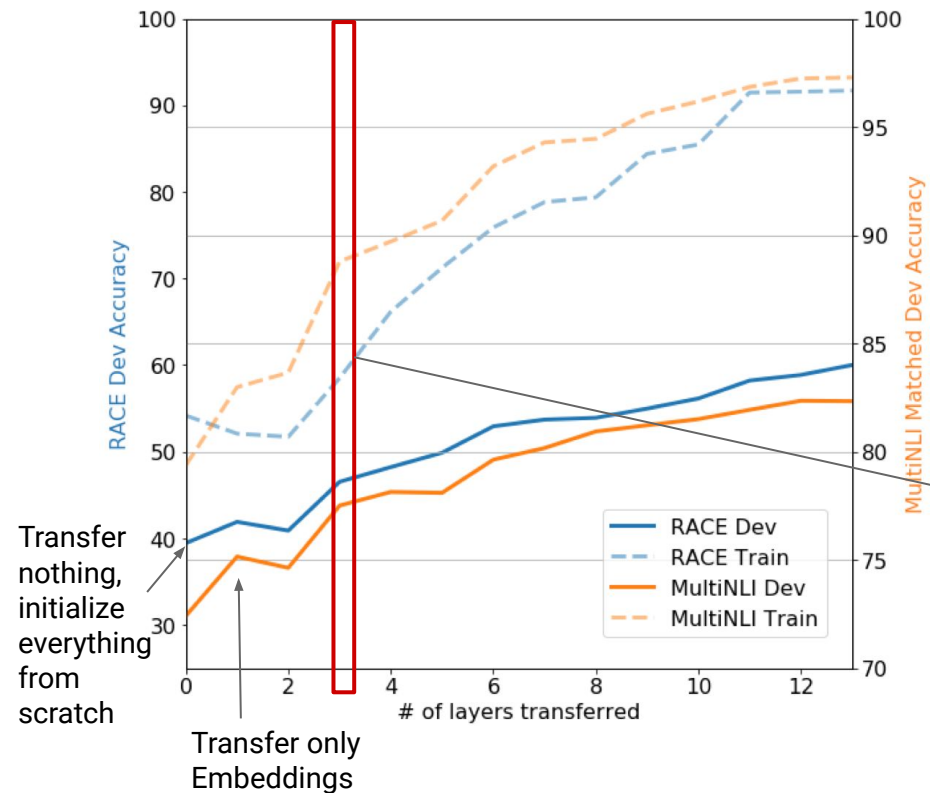
Why Pre-Training?



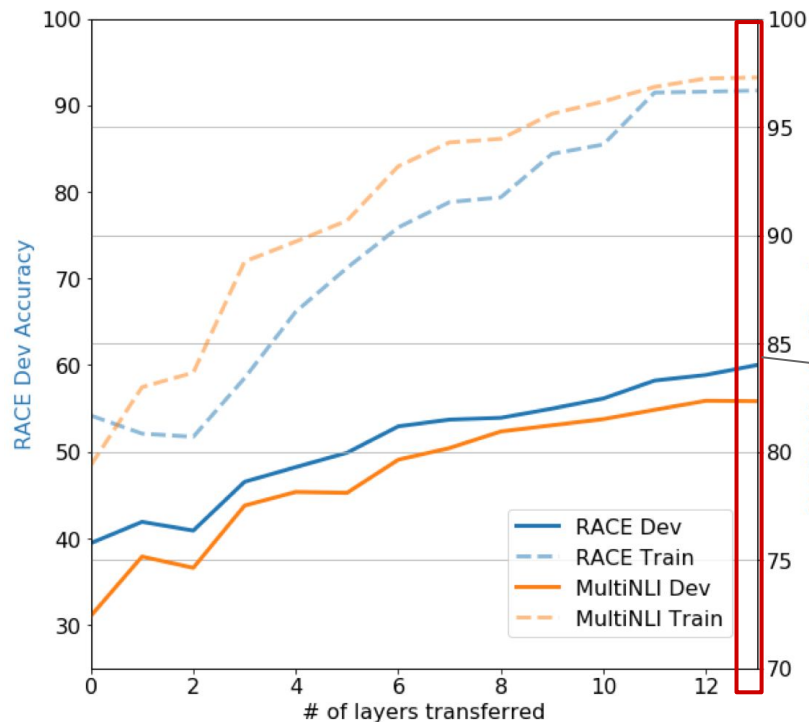
Transfer
Embedding
Layer and
TLayer 1
and 2
=> 3 Layers
transferred



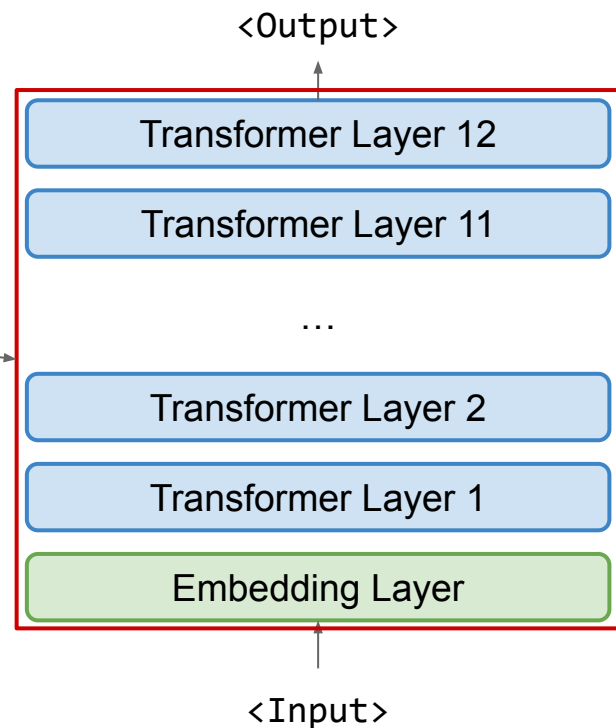
Why Pre-Training?



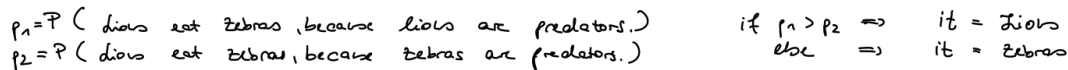
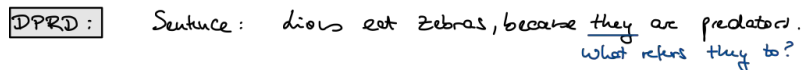
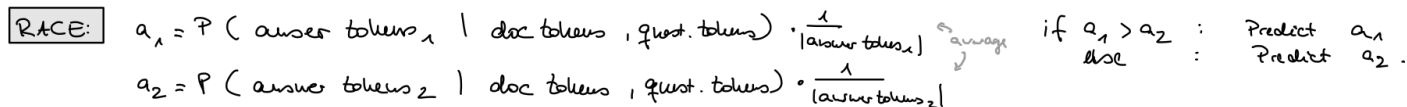
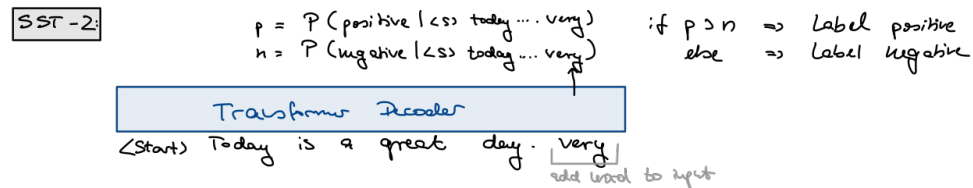
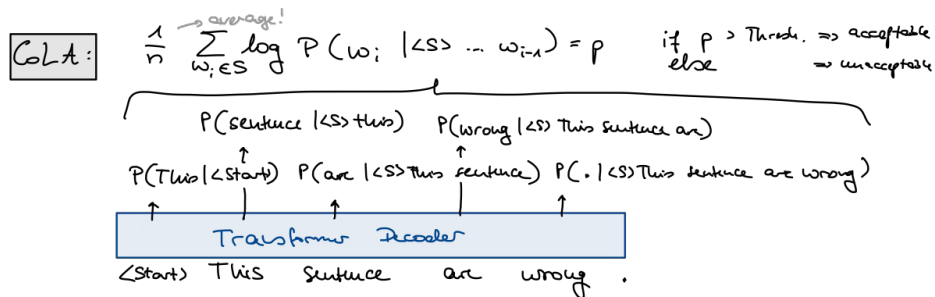
Why Pre-Training?



Best:
Transfer
Embedding
s + 12
TLayers
=> 13
Layers
transferred




Zero-Shot = "without training"



Winograd Schema (DPRD)


Does “they” refer to lions or zebras?

Lions eat zebras because **they** are predators.

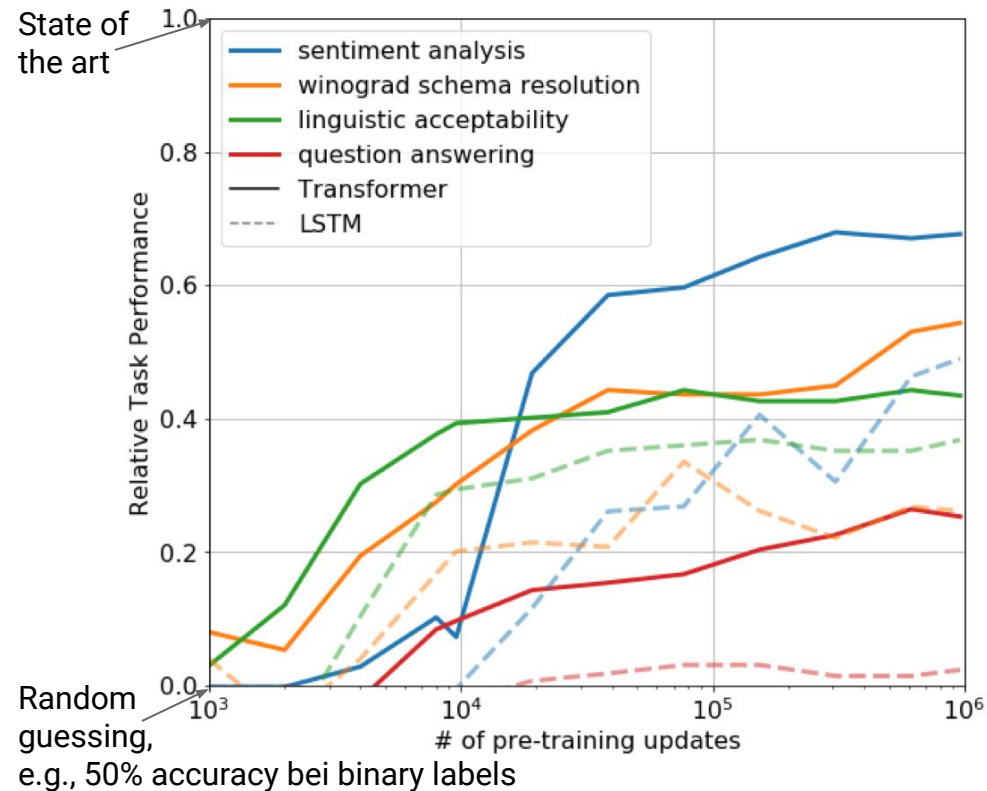
The diagram shows two dashed blue arrows originating from the word 'they' in the sentence. One arrow points to the word 'Lions' and the other points to the word 'zebras', illustrating the ambiguity of the pronoun.

Does “it” refer to knife or flesh?

The knife sliced through the flesh because **it** was sharp.

The diagram shows two dashed blue arrows originating from the word 'it' in the sentence. One arrow points to the word 'knife' and the other points to the word 'flesh', illustrating the ambiguity of the pronoun.

Zero-Shot = “without training”



Take Away Messages:

- Pre-training enables the model to solve the task already quite okay.
- with more pre-training, performance gets better
- LSTM performance varies a lot (see orange, winograd schema)
- LSTM is always worse, sometimes much much worse (see red, QA)

Ablation and GLUE Data Sizes

Table 5: Analysis of various model ablations on different tasks. Avg. score is a unweighted average of all the results. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STS-B (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

Take Away Messages:

- Pre-Trained Transformer always better than LSTM
- Auxiliary Language Modelling helps *sometimes*

Corpus	Train	Test
CoLA	8.5k	1k
SST-2	67k	1.8k
MRPC	3.7k	1.7k
STS-B	7k	1.4k
QQP	364k	391k
MNLI	393k	20k
QNLI	105k	5.4k
RTE	2.5k	3k
WNLI	634	146

Pre-Trained Transformers
are the best! 🎉

Link to Paper

https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

	architecture	parameter count	training data
GPT-1	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	0.12 billion	BookCorpus : ^[38] 4.5 GB of text, from 7000 unpublished books of various genres.
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.
GPT-3	GPT-2, but with modification to allow larger scaling.	175 billion	570 GB plaintext, 0.4 trillion tokens. Mostly CommonCrawl, WebText, English Wikipedia, and two books corpora (Books1 and Books2).

Next:

GPT-2: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

GPT-3: <https://arxiv.org/pdf/2005.14165.pdf>