Improving Language Understanding by Generative Pre-Training (GPT-1)

Previously on 🤖💬 + 📚 + ☕

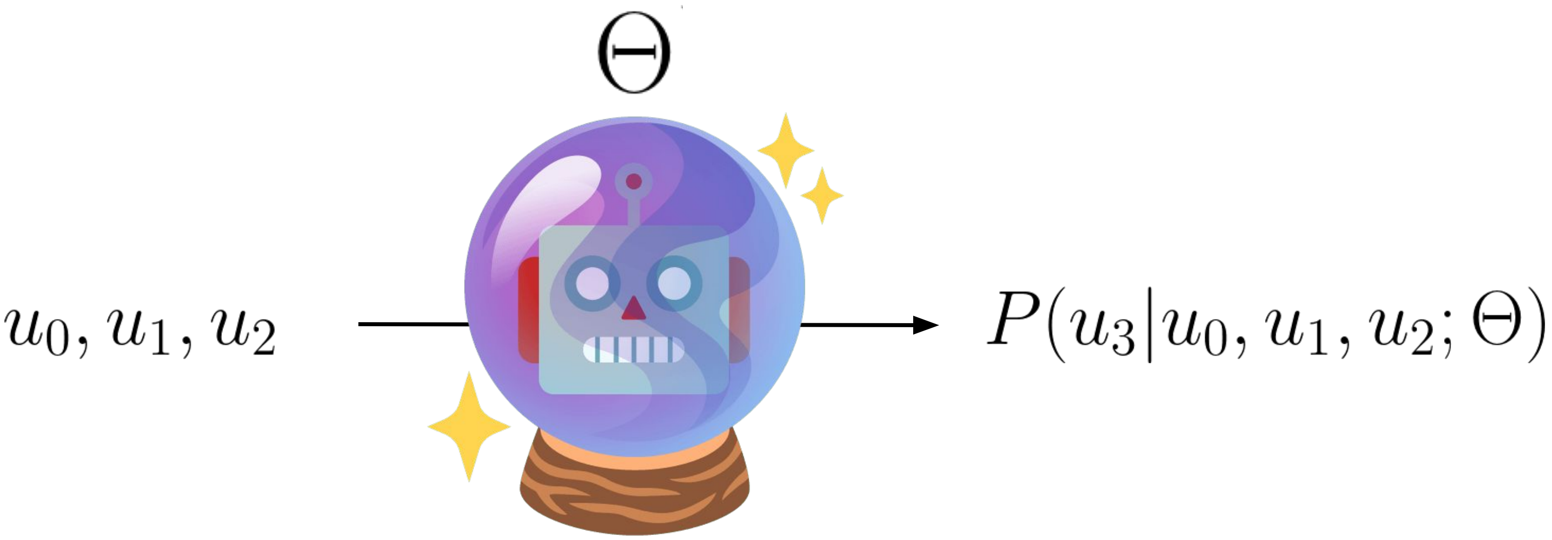
- Encoder + Decoder Transformer for Language Translation Tasks
  - single-task, no-pretraining
- Transformer Architecture
  - self-attention
  - multi-head attention
  - positional encoding
  - parallelization, causal masking
  - no recurrence anywhere 🎉

Pre-Training 📚 → Fine-Tuning 🎯

$$P(u_3|\underbrace{u_0, u_1, u_2}_{\text{previous tokens or “context”}}; \Theta)$$

next token

model parameters


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

LLM - Language Modelling

📌 We aim to maximize the probability of predicting the next word for an entire text corpus.

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

📌 next-word prediction as log-likelihood objective, given theta

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

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

$$\sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

LLM - Why log-likelihood?

📌 The log function is monotonic, meaning:

$$\arg\max P(x) = \arg\max \log P(x)$$

📌 Converts products into sums:

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)\ldots$$

$$\log P(x_1, x_2, \ldots, x_n) = \log P(x_1) + \log P(x_2|x_1) + \ldots$$

➡️ Log function maintains objective (doesn’t change which prob. is maximized)

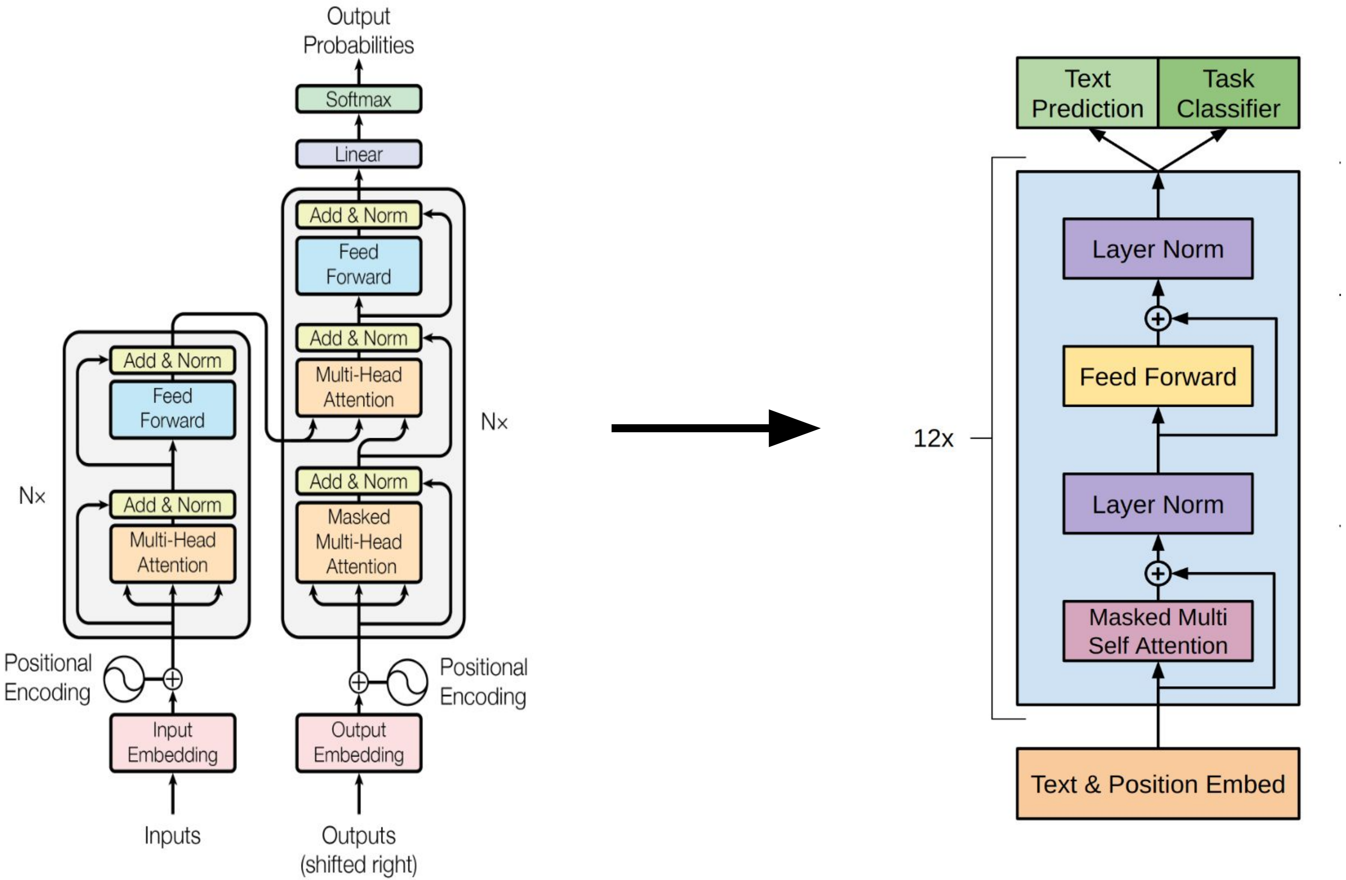
➡️ Log probs make compute simpler, more stable, numerically safe, gradients smoother

Pre-Training Corpus 📚

BooksCorpus (GPT-1, 2018): 7.185 documents (books), 1.18 GB

dolma v1.7 (AI2 dolmo, 2024): 2.532M documents, 4.7TB

6 OOM size diff



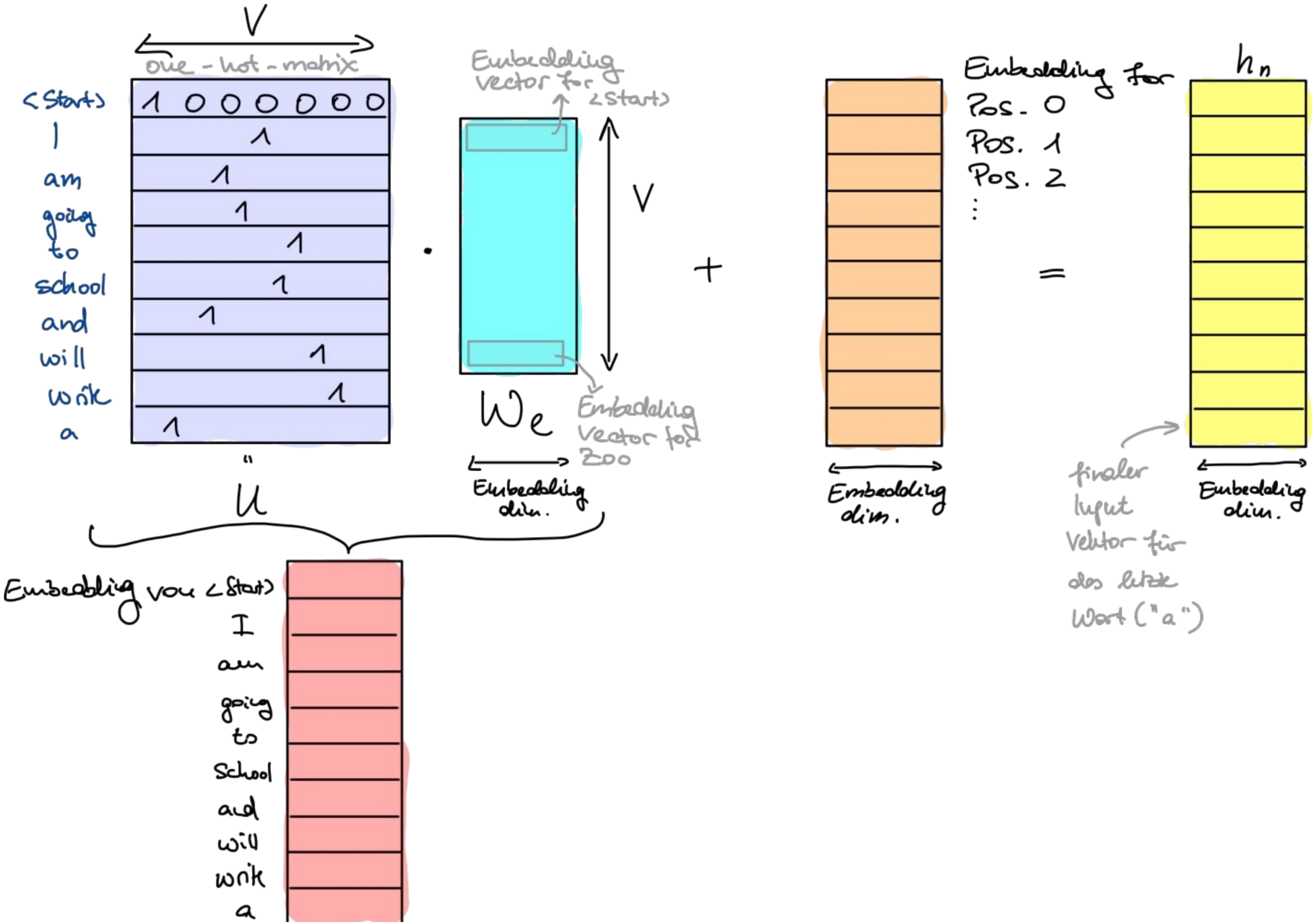
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Text Prediction
Task Classifier
Layer Norm
Feed Forward
12x
Layer Norm
Masked Multi Self Attention
Text & Position Embed

Decoder-Only Transformer

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:

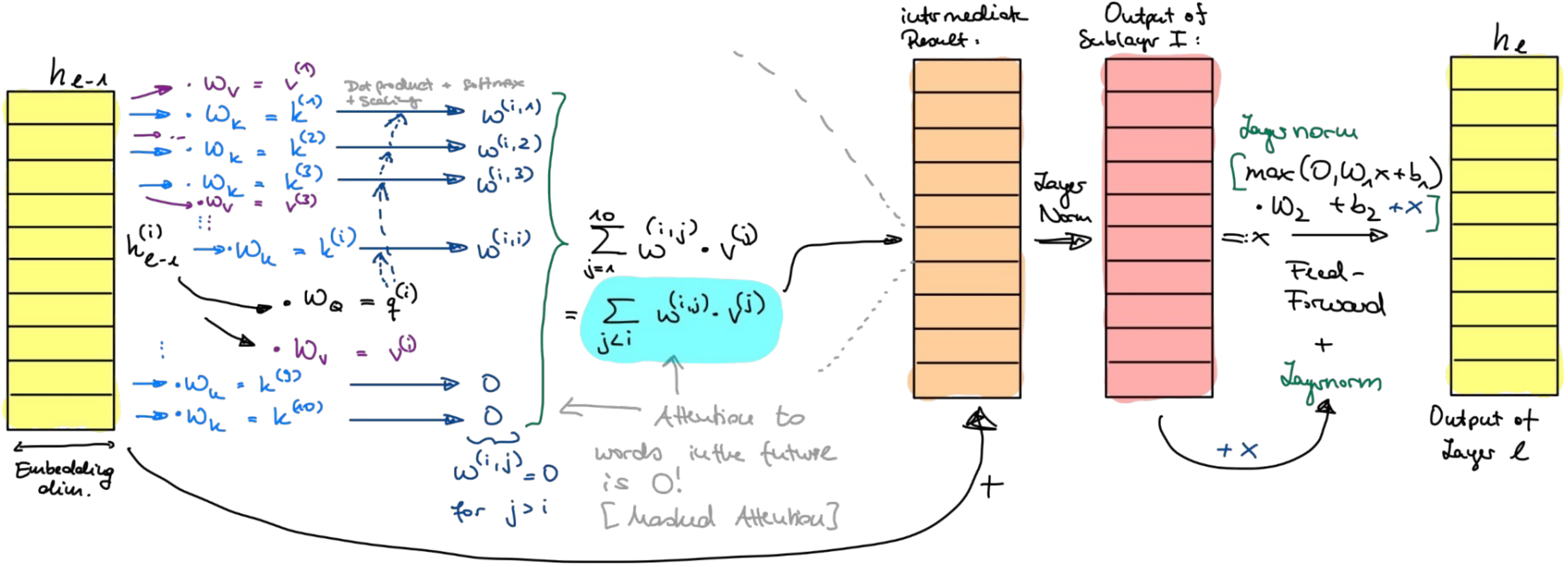
$$
\begin{aligned}
h_0 &= UW_e + W_p \\
h_l &= \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n] \\
P(u) &= \texttt{softmax}(h_n W_e^T)
\end{aligned} \tag{2}
$$

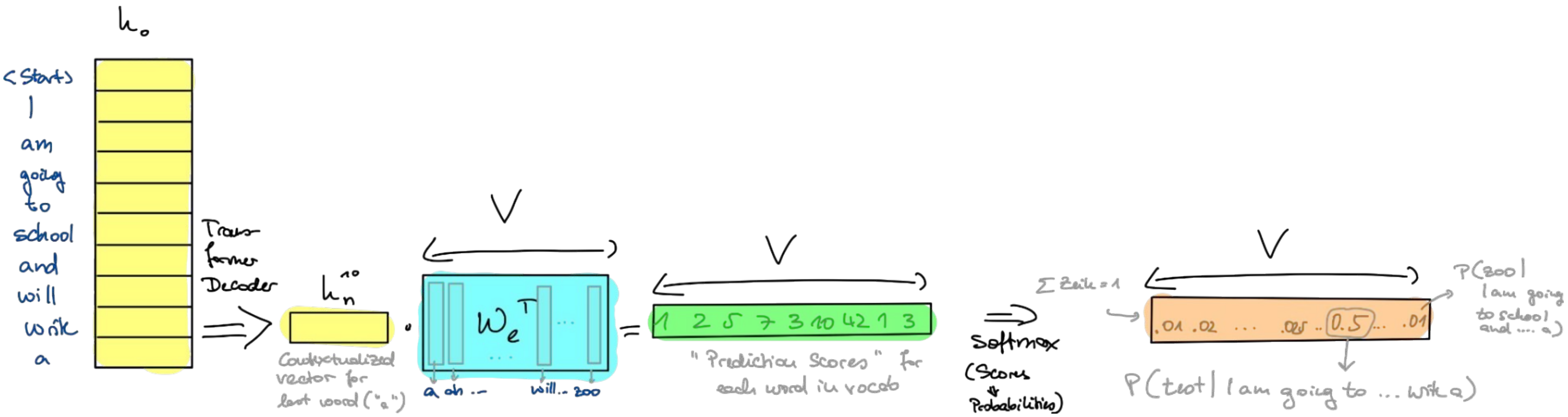
where $U = (u_{-k}, \ldots, u_{-1})$ is the context vector of tokens, n is the number of layers, W_e is the token embedding matrix, and W_p is the position embedding matrix.



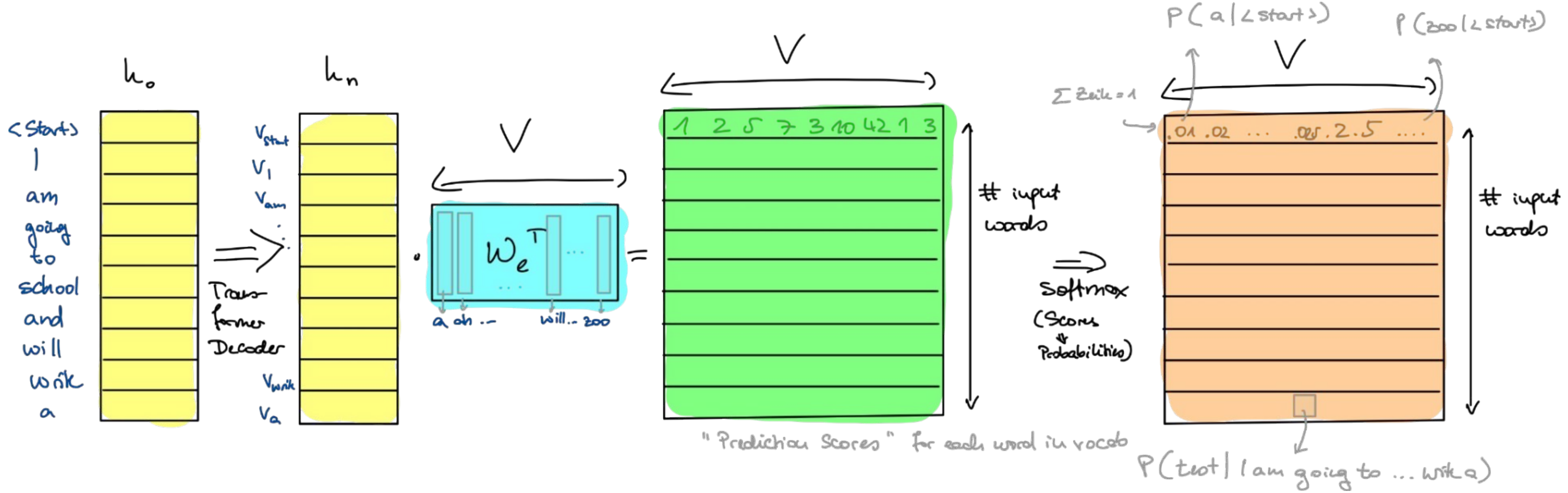
V
one - hot - matrix
<Start>
I
am
going
to
school
and
will
work
a
1 0 0 0 0 0 0
U
Embedding vector for <Start>
V
W_e
Embedding vector for Zoo
Embedding dim.
Embedding von <Start>
I
am
going
to
School
and
will
work
a
+
Embedding for
Pos. 0
Pos. 1
Pos. 2
=
h_n
Embedding dim.
finaler Input Vektor für das letzte Wort ("a")
Embedding dim.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$





h_0
<Start>
I
am
going
to
school
and
will
work
a
Trans-former Decoder
h_n^10
Contextualized vector for last word ("a")
V
W_e^T
a ah ...
will .. zoo
V
1 2 5 7 3 10 42 1 3
" Prediction Scores " for each word in vocab
Softmax
(Scores → Probabilities)
Σ z_ik = 1
V
.01 .02 ... .025 .. 0.5 ... .01
P(zoo | I am going to school and .... a)
P(tent | I am going to ... with a)



h_0
h_n
<start>
I
am
going
to
school
and
will
write
a
Transformer Decoder
V_{start}
V_I
V_{am}
V_{write}
V_a
V
W_e^T
a ah ...
will... zoo
1 2 5 7 3 10 42 1 3
input words
"Prediction Scores" for each word in vocab
Softmax
(Scores → Probabilities)
Σ Zeile = 1
.01 .02 ... .05 .2 .5 ....
P(a | <start>)
P(zoo | <start>)
P(test | I am going to ... write a)

Fine-Tuning Setup

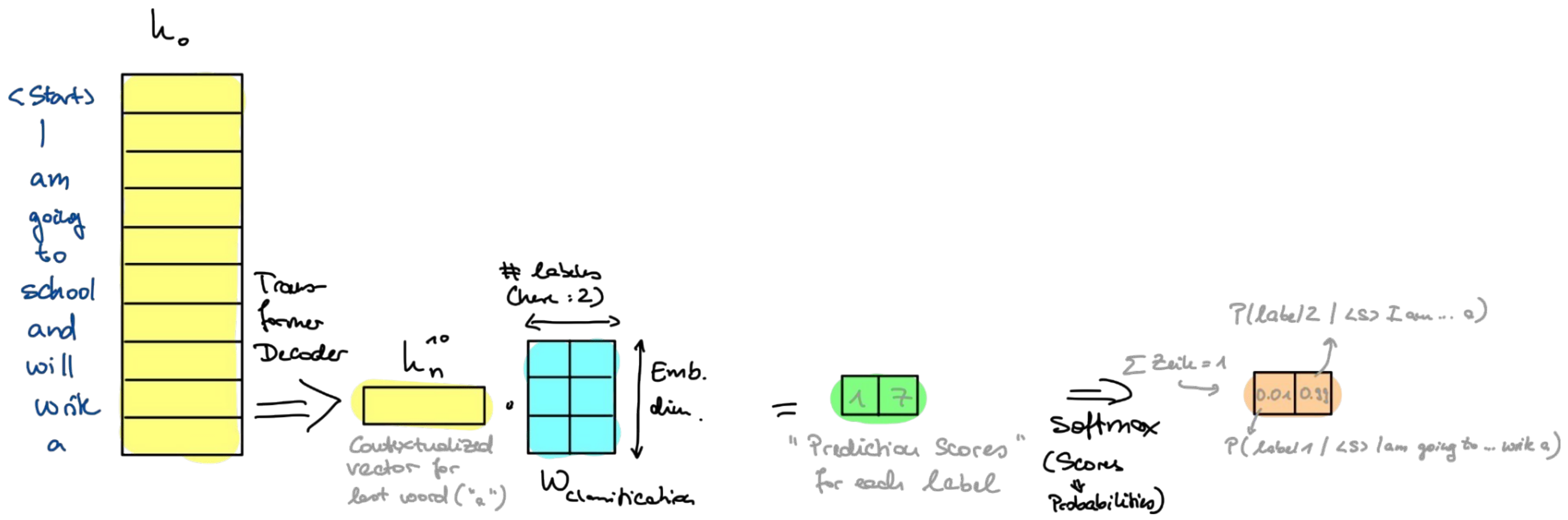
3.2 Supervised fine-tuning

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, $x^1, \ldots, x^m$, 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:

$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y). \tag{3}$$

This gives us the following objective to maximize:

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



h0
<Start>
I
am
going
to
school
and
will
work
a
Transformer Decoder
h^10_n
Contextualized vector for last word ("a")
labels (here: 2)
Emb. dim.
W classification
1 7
" Prediction Scores " for each label
Softmax
(Scores → Probabilities)
Σ Zeile = 1
0.01 0.99
P(label 2 | <s> I am ... a)
P(label 1 | <s> I am going to ... work a)

Fine-Tuning Overview

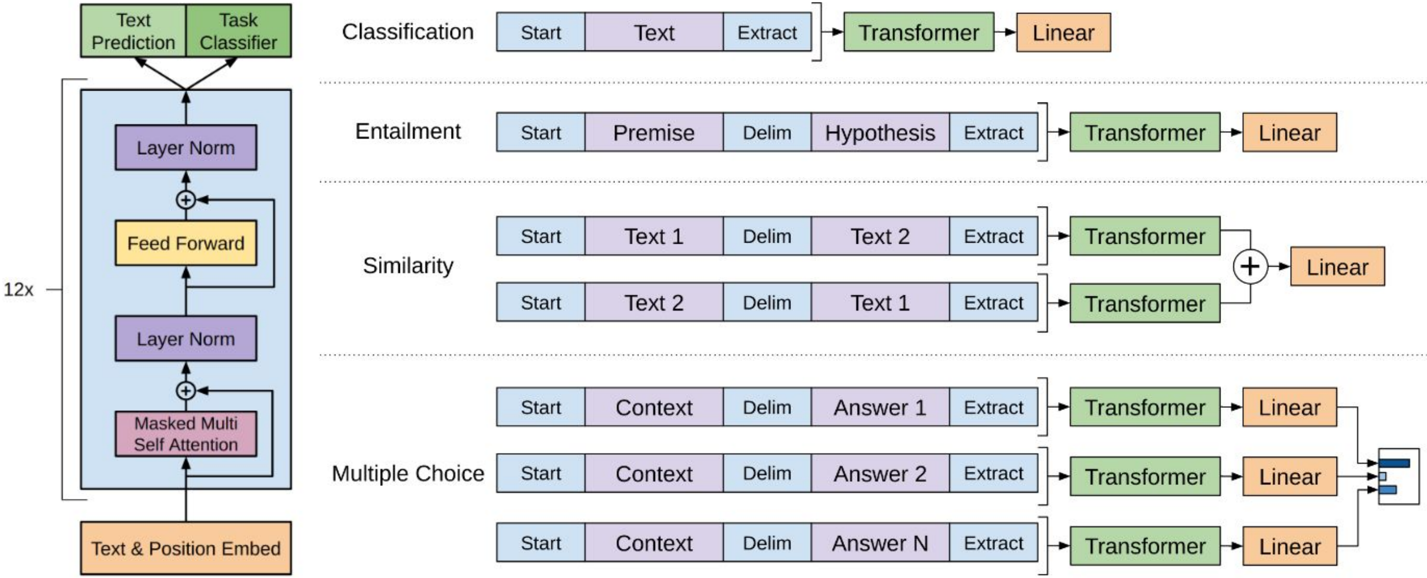


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Combined Loss (Task Specific + Weighted LM Loss)

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):

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

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

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]

Sentiment Analysis (Classification)

📜 Sentence 1 (✅ positive):
"This burger was so good, I almost proposed to the chef."

📜 Sentence 2 (❌ negative):
"This hotel room was so small, I had to step outside just to change my mind."

📜 Sentence 3 (🤔 neutral):
"The chair exists. It does chair things. Nothing more, nothing less." 🪑

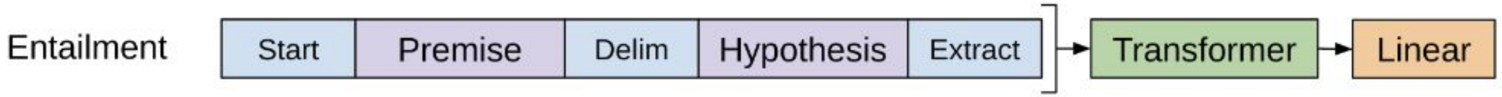
Text Entailment $<p>\$<h>$

Premise:

📜 "The astronaut stepped out of the spacecraft onto the moon."

Hypothesis:

1. ✅ Entailment (logically follows 📜) → "The astronaut is on the moon."
2. ❌ Contradiction (contradicts 📜) → "The astronaut remained inside the spacecraft."
3. 🤔 Neutral (plausible but not supported by 📜) → "The astronaut was sent on a mission to Mars."

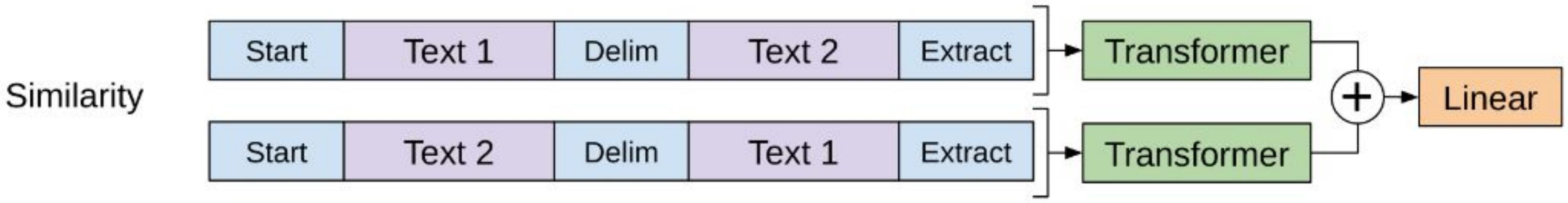


Sentence Similarity

Task: Determine how similar two sentences are in meaning [sim. index]

📜 Sentence 1: "The cat is sleeping on the sofa."
📜 Sentence 2:

1. ✅ High Similarity → "A cat is resting on the couch."
2. 🤔 Moderate Similarity → "A dog is lying on the carpet."
3. ❌ Low Similarity → "She is reading a book in the library."



RACE-Style Question Answering $[z; q; \$; a_k]$

Passage (Context):

📜 "Marie Curie was a scientist known for her research on radioactivity. She discovered two elements, polonium and radium, and won two Nobel Prizes. Her work laid the foundation for modern medical treatments like radiation therapy."

❓ Question:

"What is one major contribution of Marie Curie?"

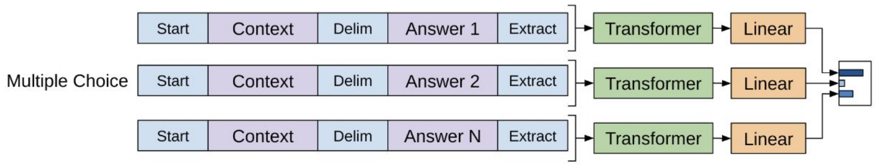
🔢 Answer Choices:

a) She discovered X-rays.

b) <u>She discovered polonium and radium.</u> ✅

c) She invented the microscope.

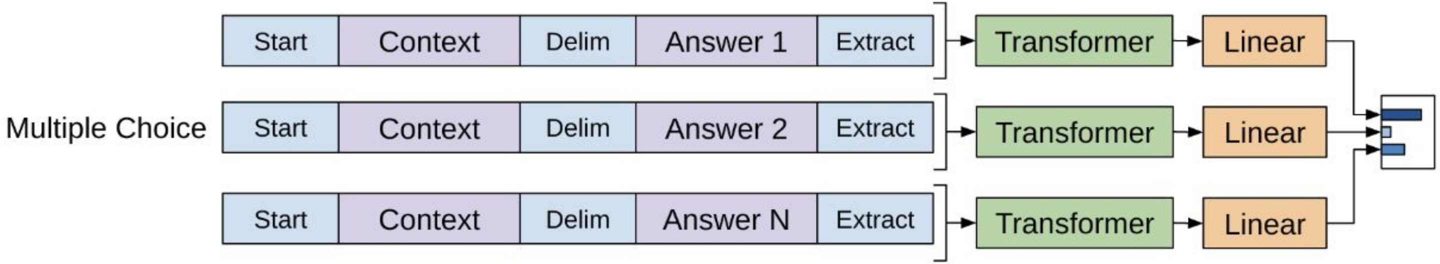
d) She was the first female astronaut.



Story Cloze Test (Commonsense Completion)

Story Start:
📜"Emma was excited about her birthday party. She decorated the house, prepared snacks, and invited all her friends. When the time came, she..."

Which sentence best completes the story?
a) <u>"...welcomed her friends with a big smile."</u> ✅
b) "...went to bed early and slept through the night." ❌



GLUE Benchmark

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

Fine-Tuning Overview

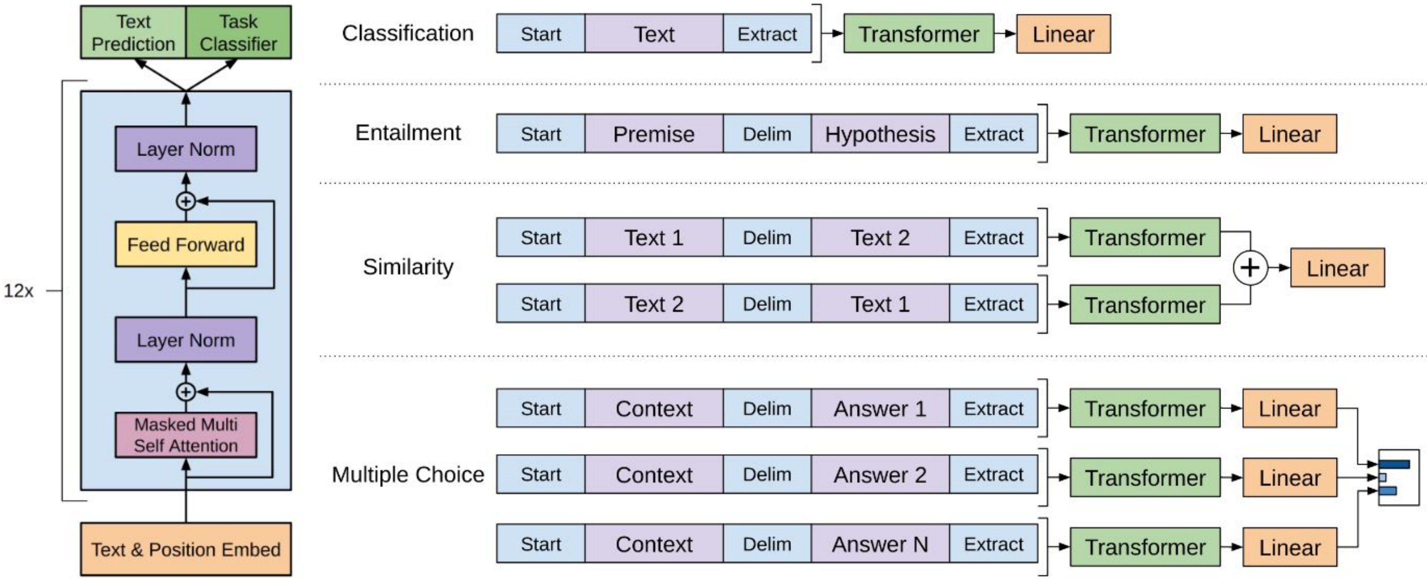
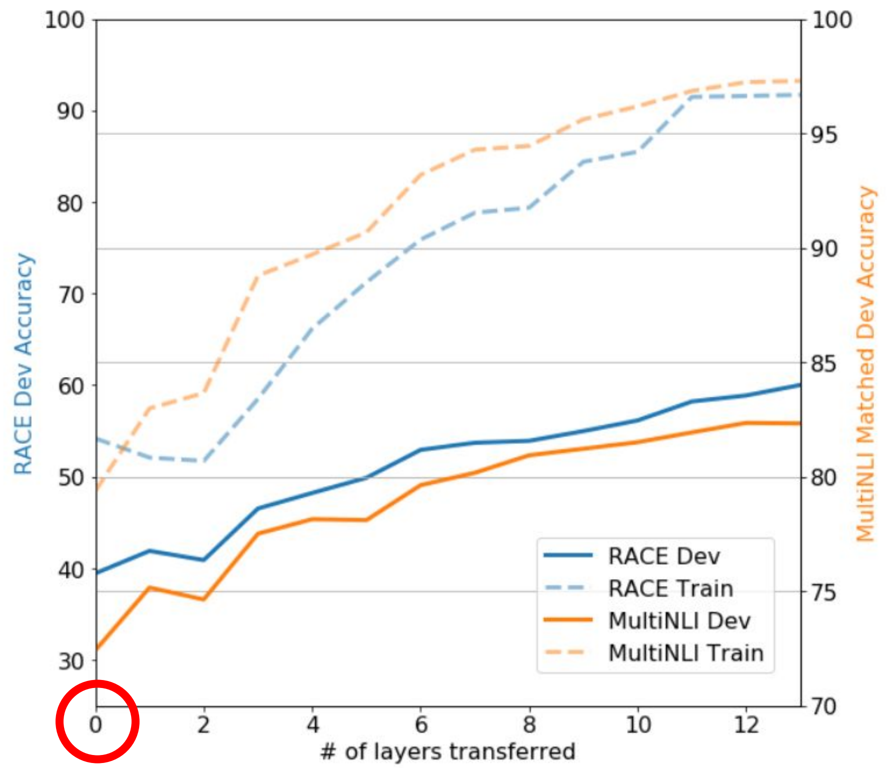
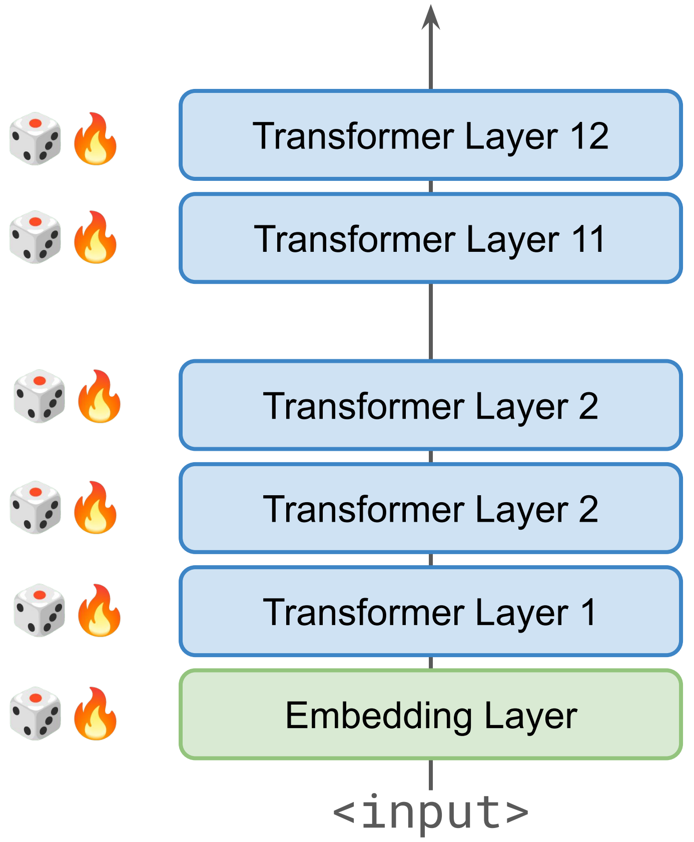


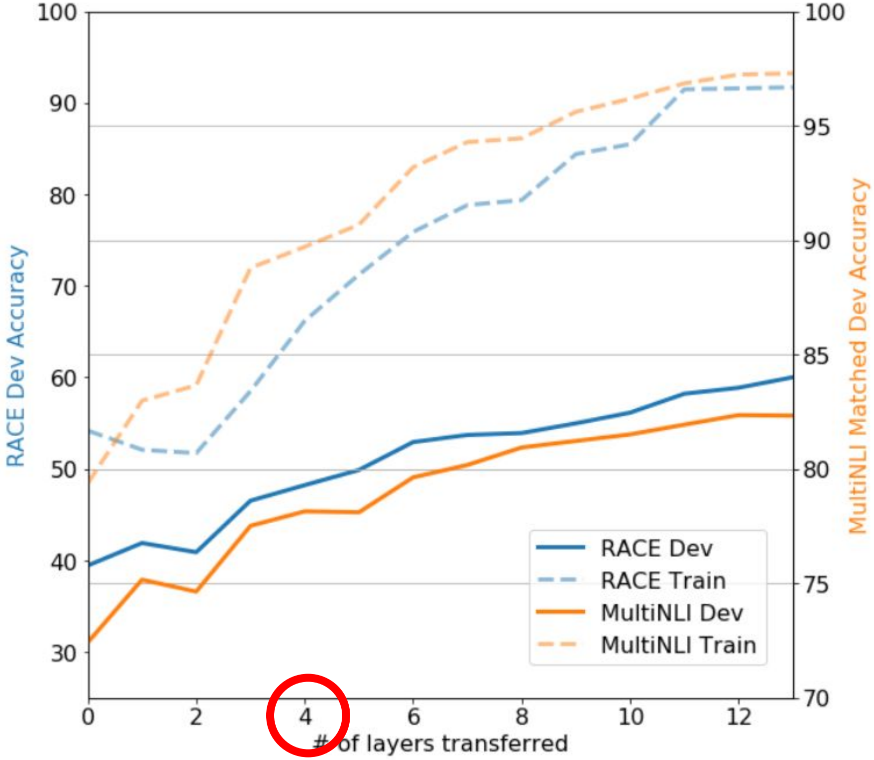
Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

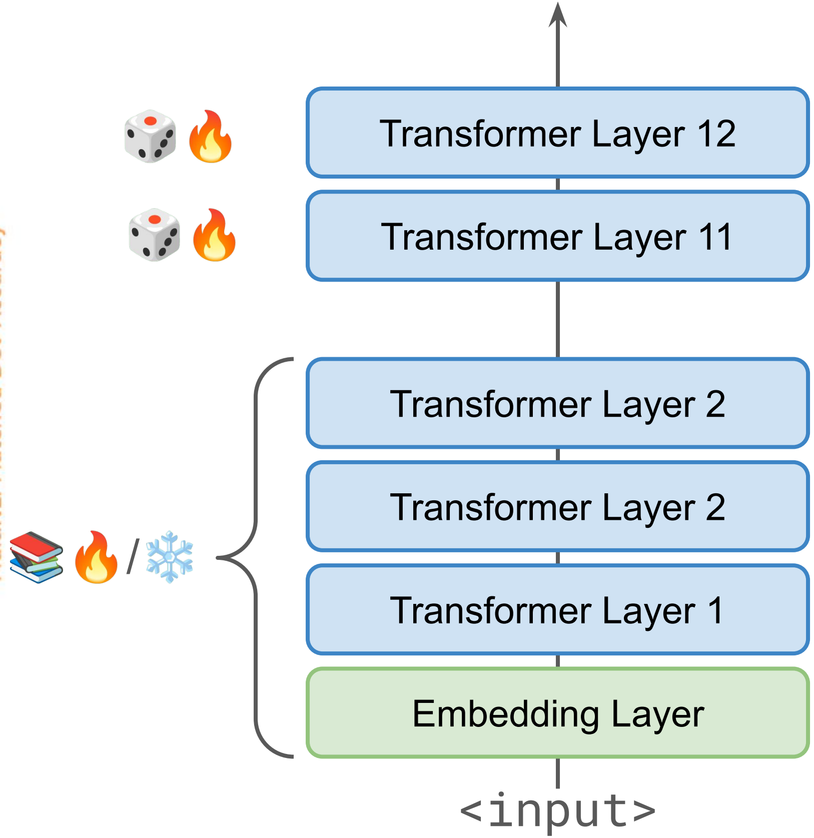
Why bother w/ Pre-Training?



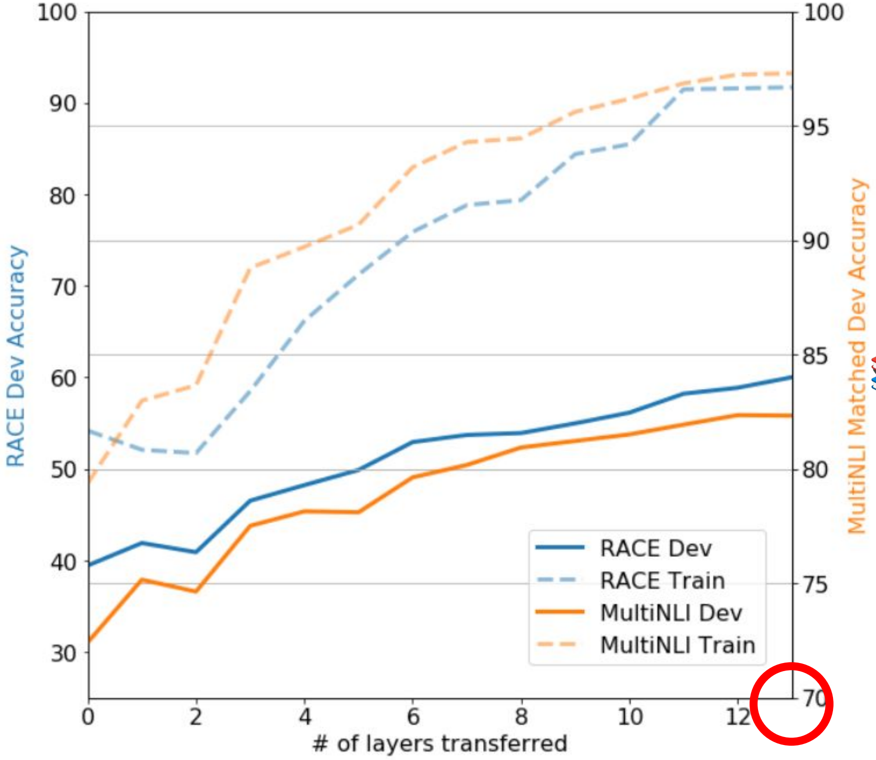


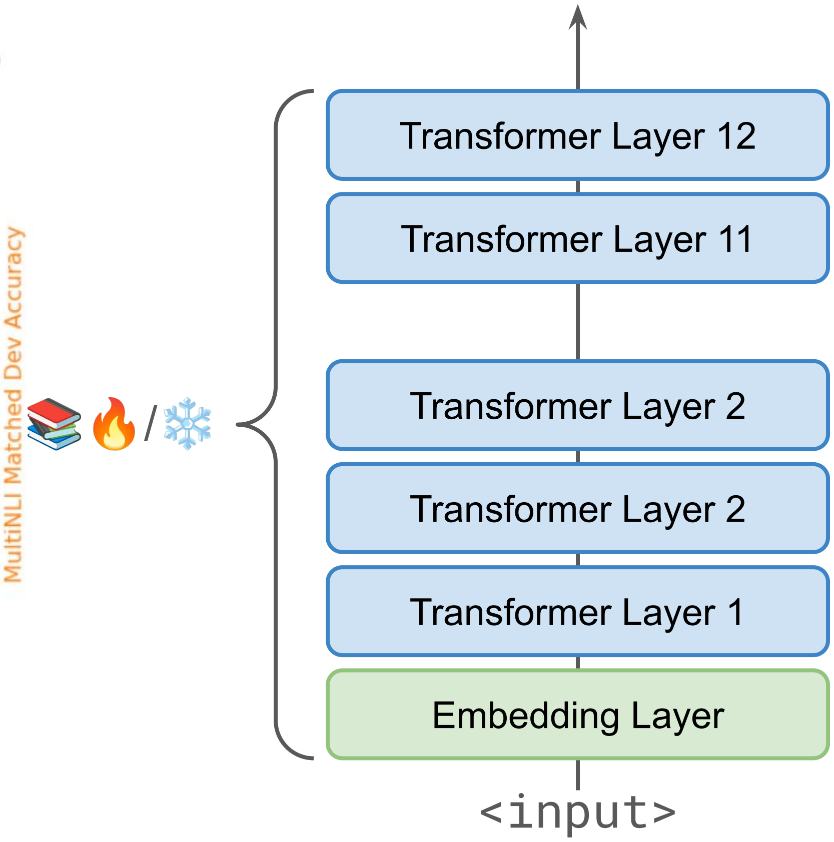
Why bother w/ Pre-Training?





Why bother w/ Pre-Training?

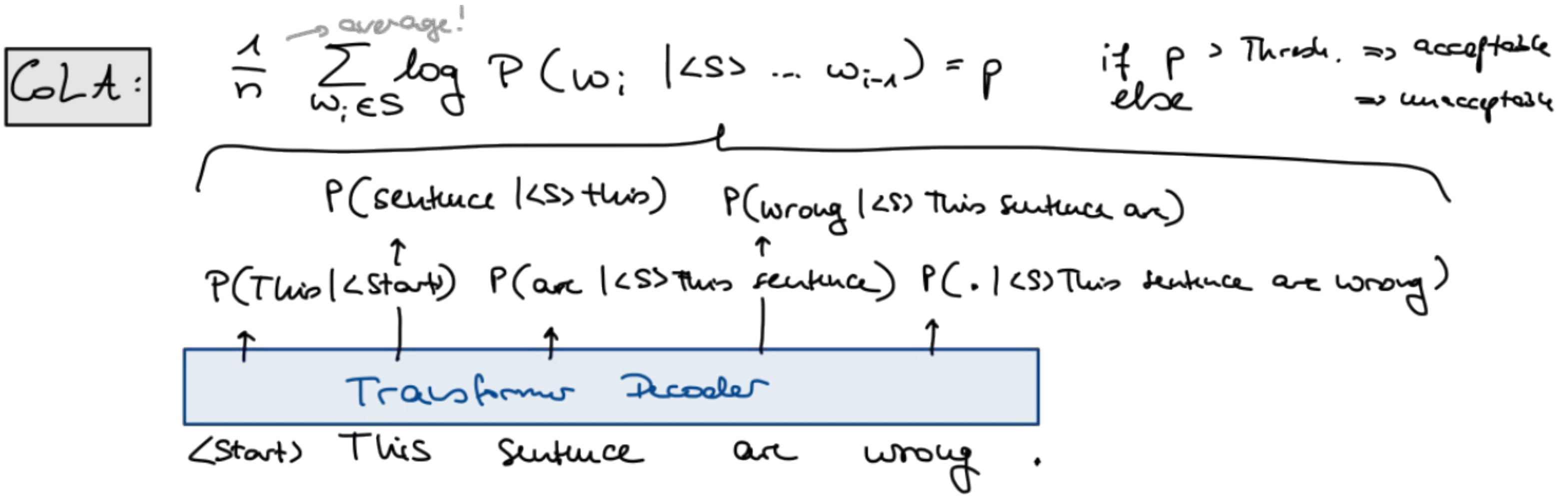




“Zero-Shot” or
Let’s skip Fine-Tuning💅

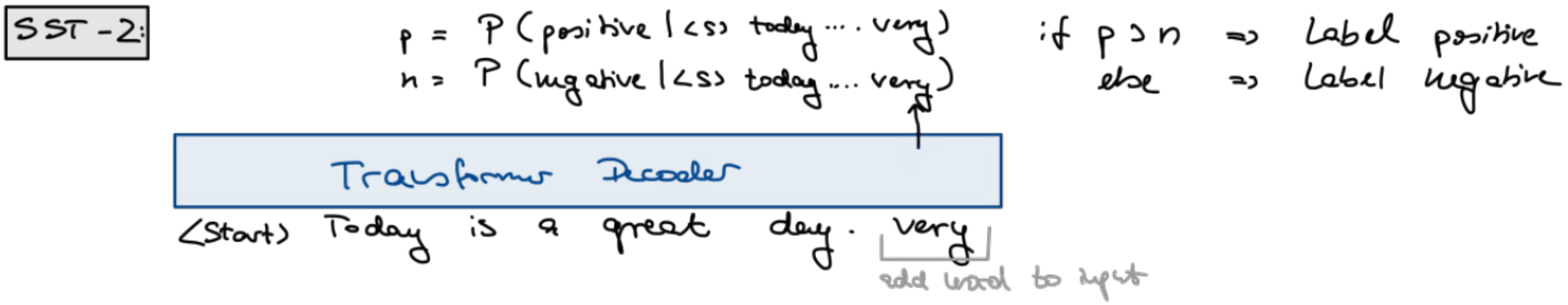
Zero-Shot CoLA

📌 Compute the average log-probability for each sentence; if above a threshold, mark as grammatical.



Zero-Shot SST-2

📌 Append "very", restrict output to "positive" or "negative", and choose the higher log-prob. word.



Zero-Shot RACE

📌 Select the answer with the highest log-probability given the document and question.

$\boxed{\text{RACE:}}$

$$a_1 = P(\text{answer tokens}_1 \mid \text{doc tokens}, \text{quest. tokens}) \cdot \frac{1}{|\text{answer tokens}_1|}$$

average

$$a_2 = P(\text{answer tokens}_2 \mid \text{doc tokens}, \text{quest. tokens}) \cdot \frac{1}{|\text{answer tokens}_2|}$$

if $a_1 > a_2$: Predict a_1

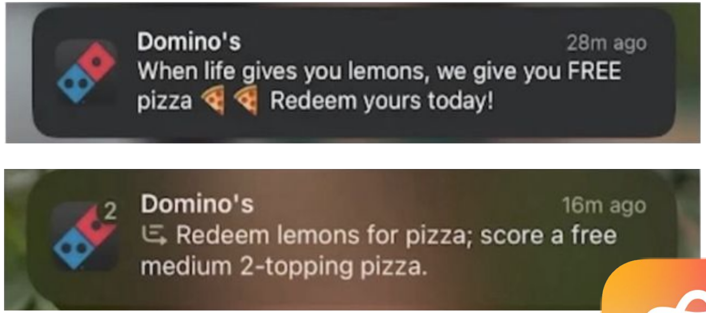
else : Predict a_2.

Winograd Schema?

📜 Sentence:
"The trophy doesn't fit in the suitcase because it is too big."

Question: What does "it" refer to?

a. ✅ trophy
b. ❌ suitcase



Zero-Shot Winograd Schema

📌 Replace the pronoun with each referent and choose the one with the highest resulting log-probability.

$\boxed{\text{DPRD:}}$ Sentence: Lions eat zebras, because they are predators.

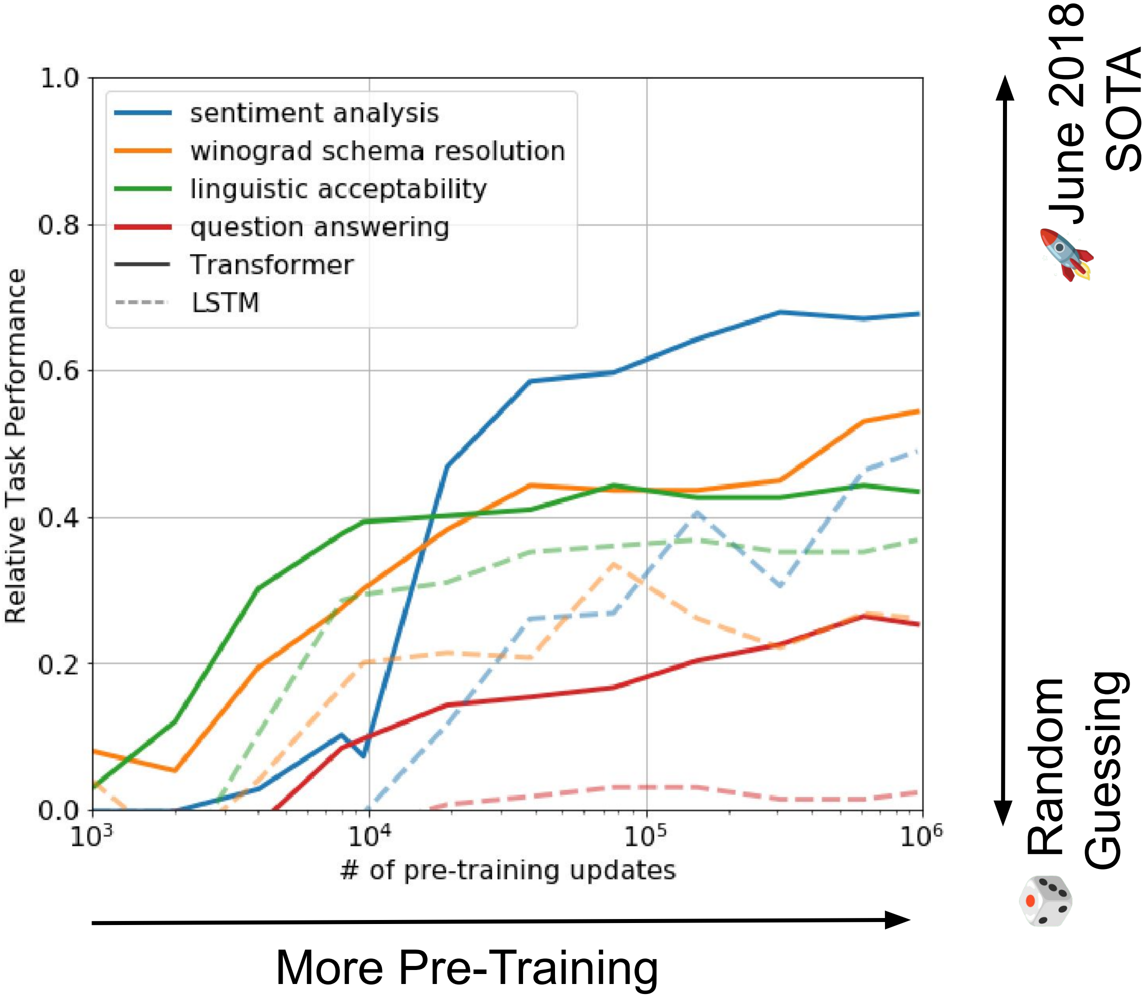
what refers they to?

$p_1 = P($ Lions eat zebras, because lions are predators. $)$

$p_2 = P($ Lions eat zebras, because zebras are predators. $)$

if $p_1 > p_2 \Rightarrow$ it = Lions

else $\Rightarrow$ it = zebras


sentiment analysis
winograd schema resolution
linguistic acceptability
question answering
Transformer
LSTM
Relative Task Performance
# of pre-training updates
More Pre-Training
June 2018 SOTA
Random Guessing

Take Away Messages

Pretraining helps the model perform reasonably well from the start.

More pretraining improves performance over time.

LSTM performance is highly inconsistent (see orange, Winograd Schema).

LSTM underperforms compared to Transformers, sometimes drastically (see red, QA).

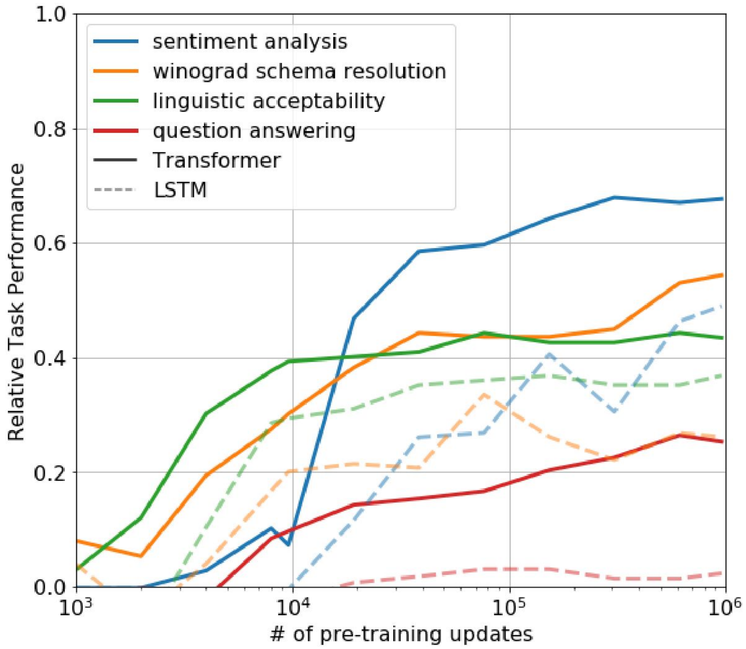
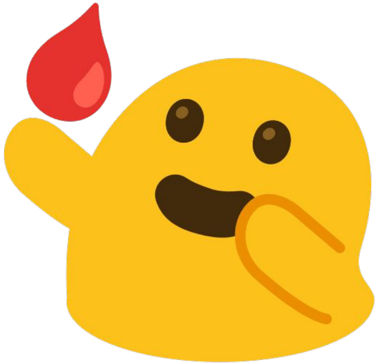


Table 5: Analysis of various model ablations on different tasks. Avg. score is a unweighted average of all the results. (*mc*= Mathews correlation, *acc*=Accuracy, *pc*=Pearson correlation)

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

- SSL NTP pre-training is surprisingly capable
- Scaling pre-training increases downstream performance
- Adaptive task formulation instead of task adaptive architecture

TW: Blood
Shameless Plug



Research Project Topic [INF-PM-FPA, CMS-PRO]