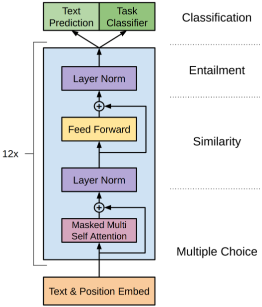
Using the Transformer

GPT (Radford et al., 2018)

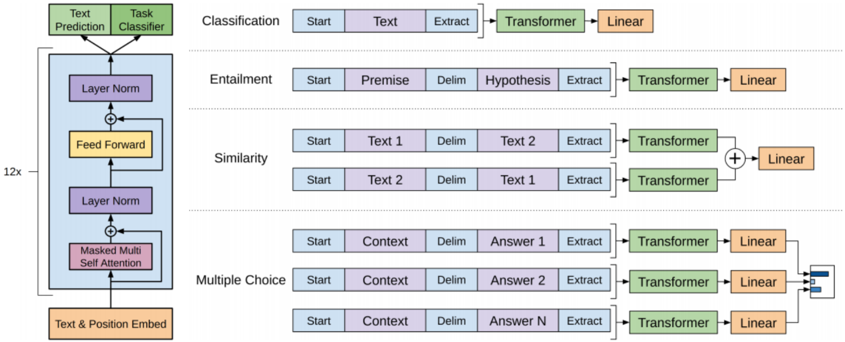


Learning goals

- Understand the use of the transformer decoder in this model
- Understand the input modifications and how this is useful

GPT

RADFORD ET AL., 2018



Source: *Radford et al., 2018*

ARCHITECTURAL DETAILS

- Transformer *decoder* as backbone of the architecture
  - 12-layer-decoder with masked self-attention mechanism
  - Hidden dimension $H = 768$, $A = 12$ Attention heads
  - BPE vocabulary w/ 40k merges
  - Learned positional embeddings (as opposed to fixed, sinusoidal ones in the original Transformer)

$$h_0 = UW_e + W_p$$
$$h_l = Trafo(h_{l-1}) \forall l \in [1, n]$$
$$P(u) = softmax(h_n W_e^\top)$$

PRE-TRAINING GPT

- Standard LM objective

$$L_1(\{u_1, \ldots, u_n\}) = \sum_i \log(P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta))$$

  where $\{u_1, \ldots, u_n\}$ is an *unlabeled* sequence of tokens
- *Resource:* BooksCorpus
  - > 7k unpublished books from various genres
  - contains long texts and thus allows learning long range dependencies

FINE-TUNING GPT

- Linear output layer with softmax activation on top
- Auxiliary language modeling objective during fine-tuning
  $\rightarrow$ Improves generalization
  $\rightarrow$ Accelerates convegence
- Task-specific input transformations
  - *Entailment:*
    Concatenation of premise (p) & hypothesis (h): $[p; \$; h]$
  - *Similarity:* Use both orderings and concatenate resulting representations: $[s_1; \$; s_2]$ and $[s_2; \$; s_1]$
  - *Q&A and Commensense Reasoning:*
    Concatenate context (z), question (q) and each possible answer (a_k): $[z; q; \$, a_k]$
- Fine-tuning is rather quick, 3 epochs were sufficient

FINE-TUNING GPT

- Additional objective:

$$L_2(\{x^1,\ldots,x^m\}) = \sum_{x,y} \log(P(y|x^1,\ldots,x^m))$$

  where

  - $P(y|x^1,\ldots,x^m) = \mathit{softmax}(h_l^m W_y)$ and
  - $\{x^1,\ldots,x^m\}$ is a *labeled* sequence of tokens

- Combining both objectives:

$$L_3(\{x^1,\ldots,x^m\}) = L_2(\{x^1,\ldots,x^m\}) + \lambda \cdot L_1(\{x^1,\ldots,x^m\})$$

SOTA RESULTS

Performance on different benchmarks:

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

Source: Radford et al. (2018)