GPT, GPT-2, GPT-3

GPT: **Improving Language Understanding by Generative Pre-Training** (2018) 4.4k

GPT-2: **Language Models are Unsupervised Multitask Learners** (2019) 4.4k

GPT-3: **Language Models are Few-Shot Learners** (2020) 7k

GPT-3与BERT都基于transformer，GPT-3效果更好，但BERT更小更容易复现，所以在学术界的影响力更大

GPT

generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task

1. Unsupervised pre-training

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

$L_1(U) = \Sigma_i logP(u_i|u_{i-k}, ..., u_{i-1}; \theta)$

where k is the size of the context window, and the conditional probability P is modeled using a neural network with parameters θ.

Use a multi-layer Transformer decoder for the language mode.

2. Supervised fine-tuning

$x_1, ..., x_m$, along with a label y

$P(y|x_1, ..., x_m) = \text{softmax}(h_l^m|W_y)$

objective to maximize: $L_2(C) = \Sigma_{(x,y)} logP(y|x^1, ..., x^m)$

把之前的语言模型同时作为objective效果更好：$L_3(C) = L_2(C) + \lambda L_1(C)$

3. Task-specific input transformations

下面需要把task转化为$x_1, ..., x_m$, along with a label y

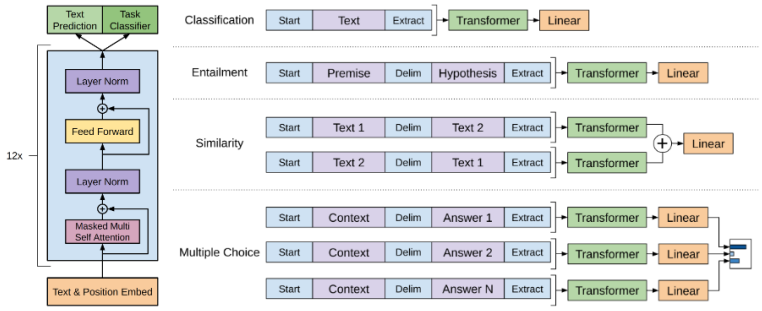

Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

GPT-2

更大的数据集，更大的模型

卖点：zero-shot，下游任务不用labels

GPT-3

更更大的模型，不在下游任务上做fine-tuning

zero-shot → few-shot

卖点：能生成高质量文本