Foundations of Natural Language Processing

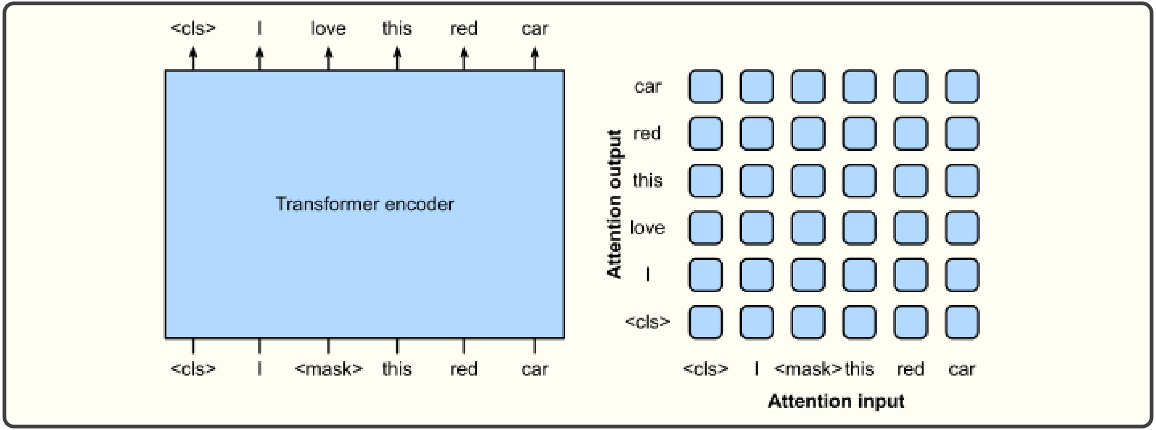
Lecture 19: Transfer Learning II

Mirella Lapata
School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

Recap

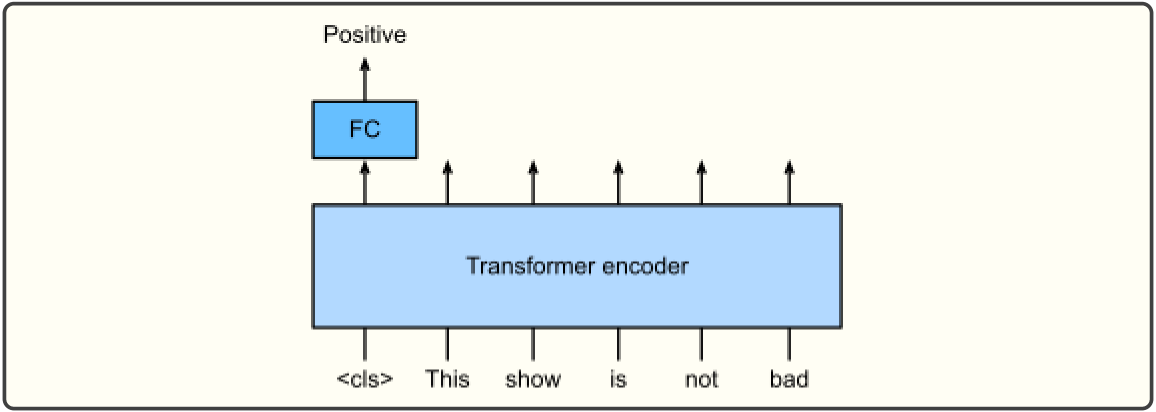
- **Last time:** we talked about BERT, a contextualized language model that uses a deep, bidirectional transformer architecture.
- BERT is *pre-trained* on unlabeld text using masking and next sentence prediction.
- It is designed designed for *finetuning* with minimal architectural modifications.
- The input uses *sentence pairs*; the output can be sentences, labels, classification decisions, depending on task.
- **Today**: we talk about T5 and GPT, two different transfer learning architectures.

Pretraining Architectures based on Transformers: BERT



- Pretraining BERT with masked language modeling (Encoder-only).
- Prediction of masked “love” token depends on all input tokens before and after “love”.
- Each token along the vertical axis attends to all input tokens along the horizontal axis.

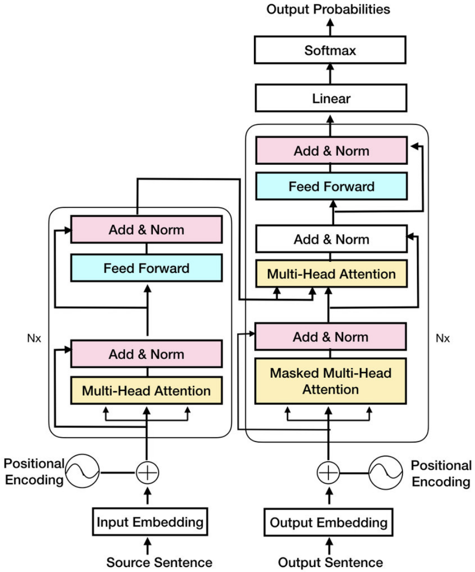
Pretraining Architectures based on Transformers: BERT



- Fine-tuning of BERT for sentiment analysis.
- encoder is a pretrained BERT: takes text sequence as input and feeds (global) “<cls>” representation into additional fully connected layer to predict the sentiment.

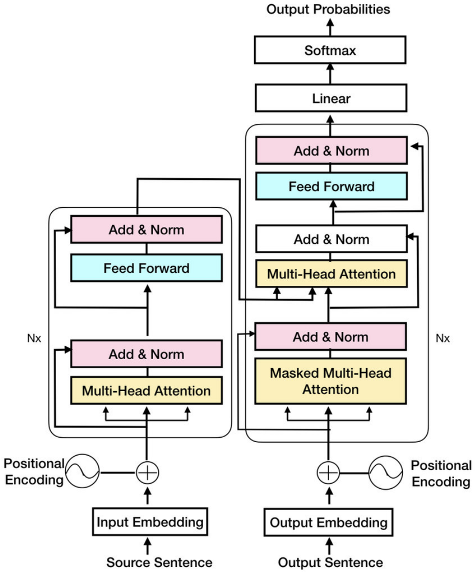
Pretraining Architectures based on Transformers: BERT

- Encoder converts sequence of input tokens into *same number* of output representations.
- Encoder-only model *cannot generate* a sequence of arbitrary length.
- Transformer architecture can be outfitted with a decoder that autoregressively predicts target sequence, token by token, conditional on both encoder output and decoder output.

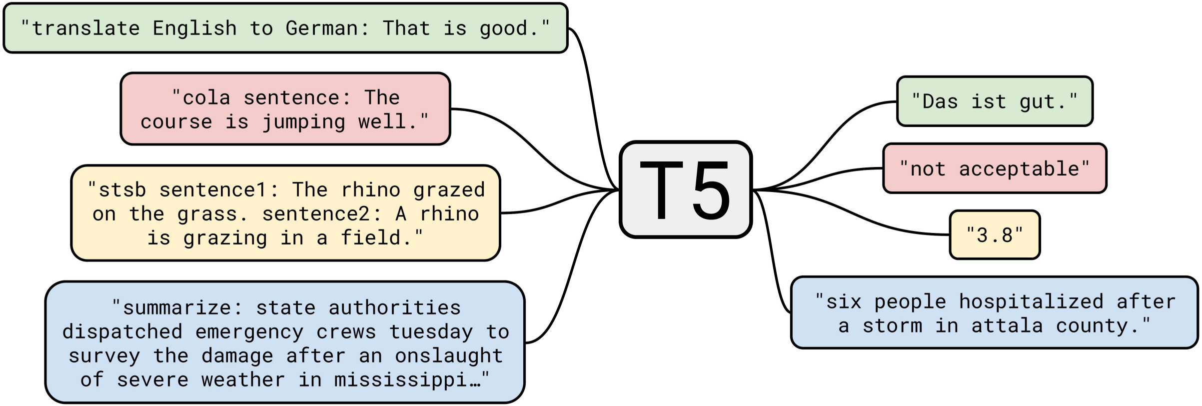


Pretraining Architectures based on Transformers: BERT

- Encoder converts sequence of input tokens into *same number* of output representations.
- Encoder-only model *cannot generate* a sequence of arbitrary length.
- Transformer architecture can be outfitted with a decoder that autoregressively predicts target sequence, token by token, conditional on both encoder output and decoder output.
- So can we then pretrain an encoder-decoder architecture?



T5: Text-to-Text Transfer Transformer



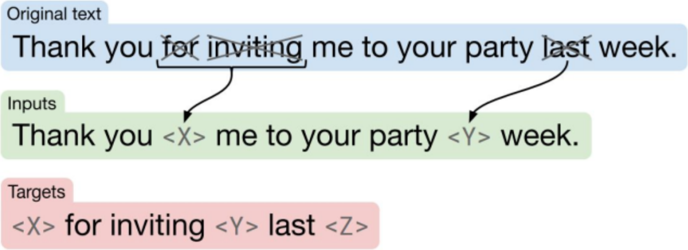
Main claim: All text processing tasks can be represented in a common format that takes in text and produces text as output. All NLP is text to text!

Baseline T5 Model

- Encoder-decoder Transformer.
- Architecture follows “Attention Is All You Need” (Vaswani et al., 2017).
- **Baseline size:** two stacks of size $\text{BERT}_{\text{BASE}}$
- Encoder and decoder consist of 12 blocks.
- **Relative position** embeddings instead of a fixed embedding for each position.
- Relative position embeddings produce different learned embedding according to the offset between the “key” and “query” compared in the self-attention mechanism.
- Baseline T5 has about **220** million parameters ($\times$ 2 the parameters of $\text{BERT}_{\text{BASE}}$).

What is the learning objective for T5?

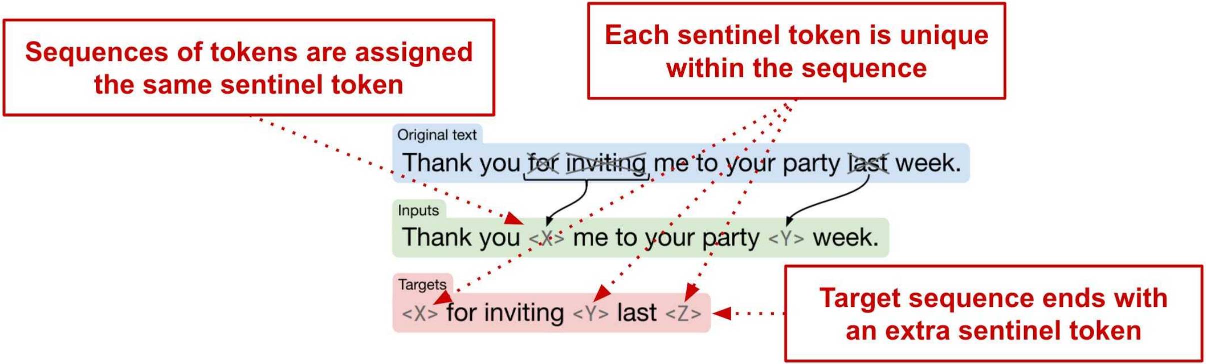
Denoising objective: also called “masked language modeling”. Model is trained to predict missing (corrupted) tokens in the input.



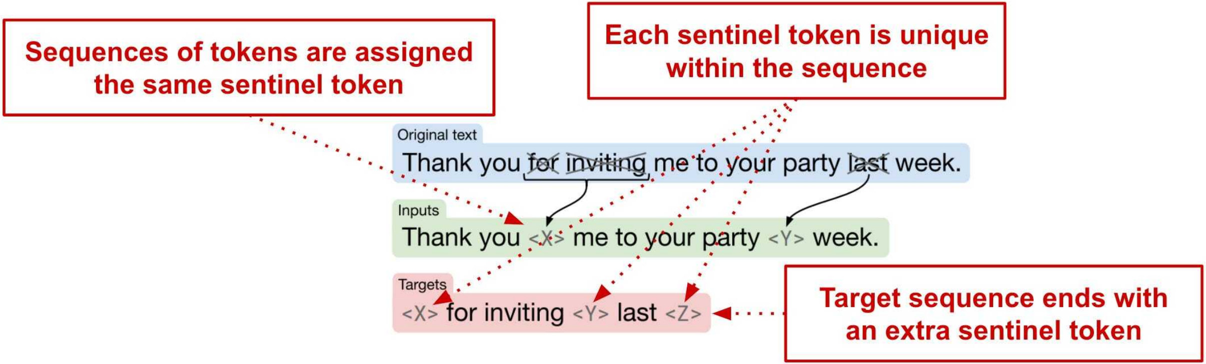
- Randomly sample and drop out 15% of tokens in input.
- Replace **consecutive spans** by **sentinel** tokens.
- Mask consecutive spans and only predict dorpped-out tokens (computational cost).

A sentinel is a soldier or guard whose job is to stand and keep watch. Synonyms: guard, lookout, warden.

What is the learning objective for T5?

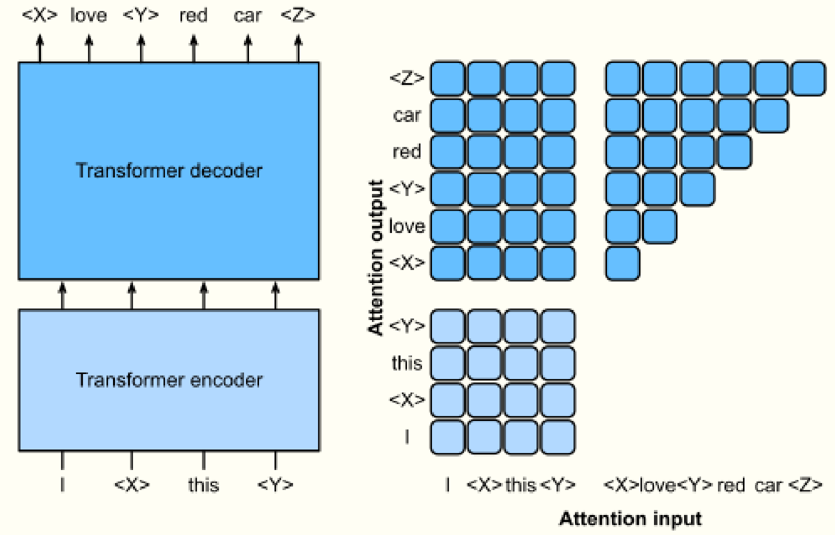


What is the learning objective for T5?



Why are sentinel tokens useful? Why can we not use BERT's `[MASK]`?

Pretraining T5



What is T5 pretrained on?

- **Common Crawl** is a publicly-available web archive that provides “web extracted text” by removing markup and other non-text content from the scraped HTML files.
- Use heuristics to clean it.
- Discard pages with fewer than 3 sentences, and lines with less than 5 words.
- Discard lines that do not end with punctuation mark.
- Remove obscene words, lines with word Javascript.
- Remove code lines, deduplicate.

The **Colossal Clean Crawled Corpus** (or C4 for short) contains 365M documents and 156B tokens.

So what are the tasks is T5 fine-tuned on?

GLUE and **SuperGLUE** comprise a collection of **text classification** tasks meant to test general language understanding abilities, and a few **generation tasks**:

- Sentence acceptability judgment
- Sentiment analysis
- Paraphrasing/sentence similarity
- Natural language inference
- Coreference resolution
- Sentence completion
- Word sense disambiguation
- SQuAD Question answering
- CNN/Daily Mail abstractive summarization
- English to German, French, and Romanian translation

What do these tasks look like?

Datasets are preprocessed to fit a **unified text format**.

CoLA

Original:	John made Bill master of himself.	1
Processed:	cola sentence: John made Bill master of himself.	acceptable

STSB

Original:	Sentence 1: Representatives for Puretunes could not immediately be reached for comment Wednesday. Sentence 2: Puretunes representatives could not be located Thursday to comment on the suit.	3.25
Processed:	stsb sentence1: Representatives for Puretunes could not immediately be reached for comment Wednesday. sentence2: Puretunes representatives could not be located Thursday to comment on the suit.	"3.2"

What do these tasks look like?

Datasets are preprocessed to fit a **unified text format**.

Machine Translation

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

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

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

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

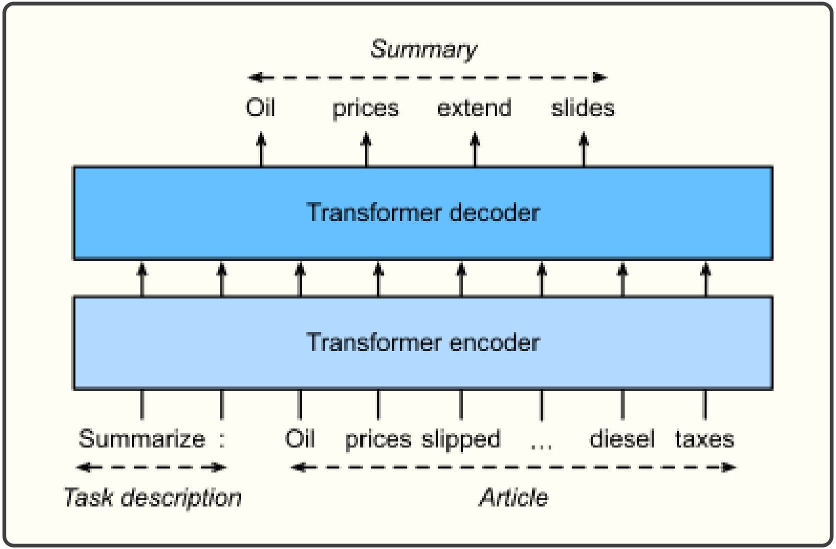
What do these tasks look like?

Datasets are preprocessed to fit a **unified text format.**

QNLI

Original:	Question: Where did Jebe die? Sentence: Genghis Khan recalled Subutai back to Mongolia soon afterwards, and Jebe died on the road back to Samarkand.	1
Processed:	qnli question: Where did Jebe die? sentence: Genghis Khan recalled Subutai back to Mongolia soon afterwards, and Jebe died on the road back to Samarkand.	entailment

Finetuning T5



The T5 Model family

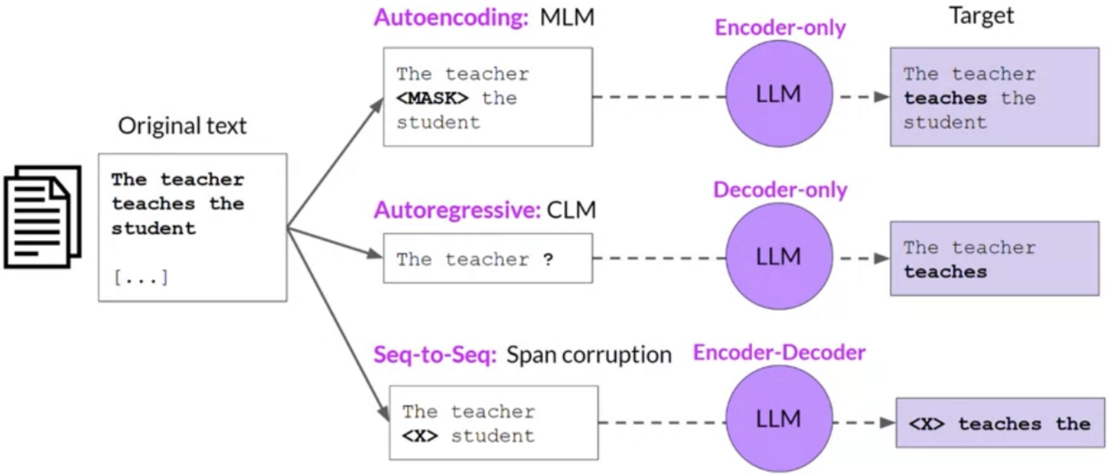
Name	d_{model}	d_{ff}	d_{kv}	Attention Heads	Encoder Layers	Decoder Layers	Size
Small	512	2,048	64	8	6	6	60M
Base	768	3,072	64	12	12	12	220M
Large	1,024	4,096	64	16	24	24	770M
3B	1,024	16,384	128	32	24	24	2.8B
11B	1,024	65,536	128	128	24	24	11B

A sample of model performance

Model	GLUE	SST-2	MRPC	STS-B	MNLI-m	MNLI-mm	SQuAD	SuperGLUE	BoolQ
	Avg	Acc	F1	ρ	Acc	Acc	F1	Acc	Avg
Previous best	89.4	97.1	93.6	92.3	91.3	91.0	95.5	84.6	87.1
T5-Small	77.4	91.8	89.7	85.0	82.4	82.3	87.24	63.3	76.4
T5-Base	82.7	95.2	90.7	88.6	87.1	86.2	92.08	76.2	81.4
T5-Large	86.4	96.3	92.4	89.2	89.9	89.6	93.79	82.3	85.4
T5-3B	88.5	97.4	92.5	89.8	91.4	91.2	94.95	86.4	89.9
T5-11B	**90.3**	**97.5**	**92.8**	**92.8**	**92.2**	**91.9**	**96.22**	**88.9**	91.2

- Better than previous best results across the board.
- Larger models perform better.

Model architectures and pretraining objectives



MLM: Masked Language Model, CLM: Causal Language Model

The GPT Family

Improving Language Understanding by Generative Pre-Training

Alec Radford Karthik Narasimhan Tim Salimans Ilya Sutskever

GPT (2018)
117 million parameters

Language Models are Unsupervised Multitask Learners

Alec Radford [* 1] Jeffrey Wu [* 1] Rewon Child [1] David Luan [1] Dario Amodei [** 1] Ilya Sutskever [** 1]

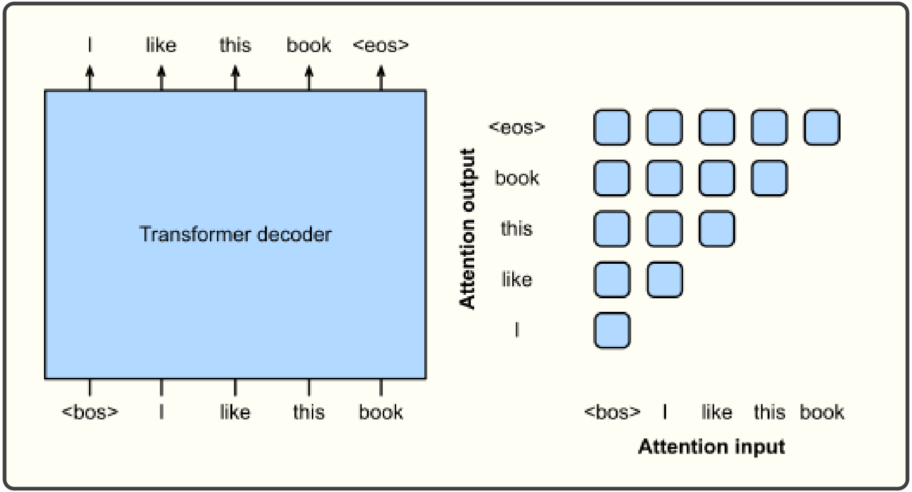
GPT-2 (2019)
1.5 billion parameters

Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

GPT-3 (2020)
175 billion parameters
NeurIPS 2020 best paper

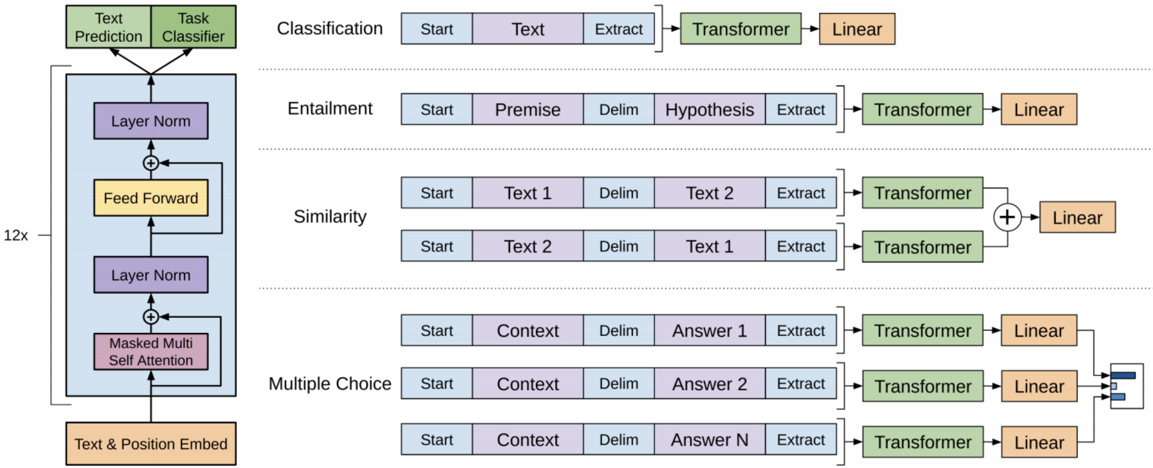
Pretraining with GPT and GPT-2



The anatomy of a GPT model

- Autoregressive model that predicts next token given tokens so far (either predicted or given as part of input).
- As it processes each subword, it masks the “future” words and conditions on (i.e., attends to) previous words.
- Consists only of decoder transformer blocks (contrast with BERT which consists only of encoders).
- There is no encoder-decoder cross-attention.

The first GPT model (sometimes called GPT-1)



Pretrained on the BooksCorpus (7,000 unique unpublished books); also shows results on fine-tuning for end tasks, where inputs and outputs are converted to text.

GPT-2 is identical to GPT-1, but

- Layer norm moved to the input of each sub-block
- Vocabulary extended to 50,257 tokens and context size increased from 512 to 1,024
- Trained on 8 million docs from the web (Common Crawl), minus Wikipedia
- Play with it here: `https://huggingface.co/gpt2`
- You can visualize attention heads here: `https://huggingface.co/spaces/exbert-project/exbert`

GPT2: Model Sizes

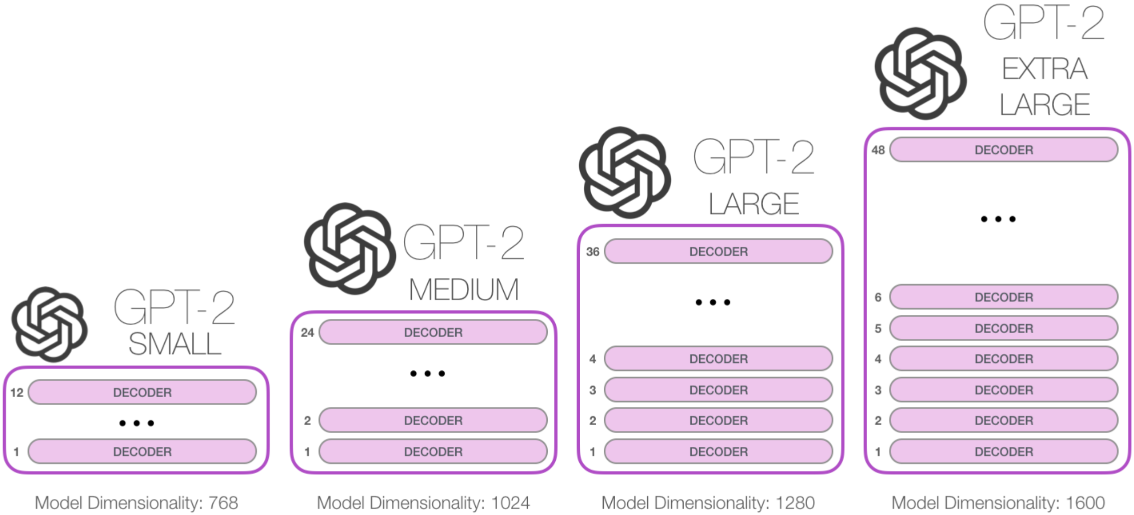


Image from `http://jalammar.github.io/illustrated-gpt2/`

Summary

- Pretraining architectures: BERT, T5, GPT
- Encoder-only, Encoder-Decoder, Decoder-only architectures
- General trend: as model grow *bigger* their *performance improves*.
- General trend: most NLP tasks can be *reformatted as generation tasks*.
- **Next time:** GPT-3, prompting, and in-context learning.