Today:
- "trained from scratch" vs. "fine-tuned"
- text-to-text transfer learning
- decoding algorithms to generate text

BERT for downstream tasks:
model: deep Transformer

What are the?
Model params:
- 3 for each head
- Wq, Wk, Wv for each block
- 2 feedforward matrix for each block
- word embeddings (maybe 50k)
- softmax matrix W₀

token embedding
Step 1: Randomly init all params

Step 2: Pretrain BERT

Random BERT

Transformer block

[CLS] the students opened their [MASK]
Step 3: fine-tune BERT

[CLS] this movie was awesome

limitations of BERT:

- What tasks can BERT not solve?
  - text generation
  - translation, summarization

Can we develop self-supervised pretraining obj.
that covers all types of NLP tasks?
- classification, seq. labeling, generation

TS paper: reformulate every NLP task as a generation problem, "text-to-text"
Sequence-to-sequence:

Input words: \( e_1, e_2, e_3, e_4 \)

Transformer encoder \( \xrightarrow{\text{cross attn}} \) Transformer decoder

Output words: \( f_1, f_2, f_3, f_4 \)

Input/output for TS:

- Missing spans
- Backprop through encoder/decoder

Random TS encoder \( \xrightarrow{\text{random TS decoder}} \) Random TS decoder

Corrupted seq.

Contradiction \( \langle \text{eos} \rangle \)

Pretrained TS encoder \( \xrightarrow{\text{pretrained TS decoder}} \)

Finetuned TS for entailment

Pretrained TS encoder \( \xrightarrow{\text{pretrained TS decoder}} \)

Fine-tune TS for machine translation

The dog walks \( \langle \text{SEQ} \rangle \)

The dog is sleeping

English sentence

French sentence
decoder-only version of T5:

thanks for inviting me to your party! for inviting less
marked self-attention, no cross attention