Common types of Transformers:

**Transformer encoders** (e.g., BERT)

- Useful if we're interested in using the final layer token representations to solve downstream NLP tasks

**Transformer decoder** ("Transformer LM", GPT-2)

- Useful to generate text, trained via LM objective

\[ C_1, C_2, C_3, C_4 \]
Transformer encoder/decoder model (e.g., T5):

- Useful for conditional text generation

Prefix LM:

- Useful for conditional text generation
  alternative to encoder/decoder approach w/ just 1 model instead of 2
Using neural LMs for transfer learning:

Let's consider sentiment analysis.

\[ \text{this movie is great} \]

all params of RNN are trained from scratch on labeled sentiment data

Issues:

- we are forcing the RNN to learn composition and world knowledge just from a tiny labeled dataset

- what if we repurpose a large-scale neural LM to solve this task?

**ELMo**: embeddings from LM
RNN LM:

- Language models can be trained on huge datasets like labels come for free.
- Word embeds come from hidden states of language model, kept frozen during training.
- Softmax predicts positive.