

Decoding Strategies

What is Decoding?

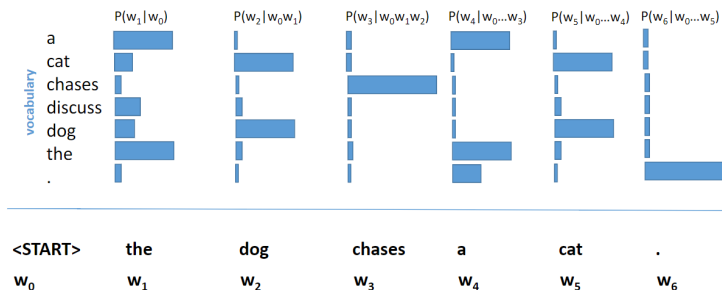
Learning goals

- Get to know the concept of decoding in NLP
- Learn about different decoding strategies

REMINDER: ARLM


- In Autoregressive Language Modeling (ARLM) the model predicts the next token given the previous tokens
- Given the context a language model produces a probability distribution over all the tokens in the vocabulary
- The context is the prompt given to the model plus the already generated tokens
- The way we then choose the next token from that probability distribution to generate natural text is called a decoding strategy

DECODING EXAMPLE



- At each timestep the model produces a probability distribution
- A decoding strategy determines how to choose the next token from that distribution, that token is then added to the context
- Generation stops based on stopping criteria (*see: next slide*)

STOPPING CRITERIA FOR TEXT GENERATIONS

- **<EOS> Token:** When this token is generated the model stops
- **Maximum Length:** A predefined maximum length can be set for the generated text. When the text reaches this length, generation stops to prevent excessively long outputs
- **Maximum Time:** A predefined maximum time for generation can be set. After this time has been reached, generation stops
- **Other Criteria:** There are more stopping criteria implemented in huggingface 

DIFFERENT DECODING STRATEGIES

The previous slide is an example for greedy search. The generated token is the one with the maximum probability at the current timestep. Various other strategies are going to be covered in this chapter:

- Deterministic Methods
 - Greedy Search
 - Beam Search
- Sampling Methods
 - Top-p
 - Top-n
 - Contrastive Search
 - Contrastive Decoding
- Decoding Hyperparameters
 - Temperature
 - ...