Raw Data

Decoding LMs

Cornell CS 5740: Natural Language Processing
Yoav Artzi, Spring 2023

What Can We Do with LMs?

- Given a sequence $\bar{x}$ compute the probability of the sequence
  - For example, for an autoregressive model[1]
- $$p(\bar{x}) = \prod_{i=1}^{N} p(x_i \mid x_1, \ldots, x_{i-1})$$
- Given a prefix, autoregressive sequence generation (i.e., decoding)
  - The prefix can be empty (sort of: always includes a start token)
  - This prefix is called a **prompt**

[1] predict future behavior based on past behavior data

Greedy vs. Sampling

- Sampling:

$$x_i \sim p(x_i \mid x_1, \ldots, x_{i-1}) \text{ until } x_i = \text{STOP}$$

- Greedy (i.e., $\arg\max$):

$$x_i = \arg\max_{x_1 \in \mathcal{V}} p(x_i \mid x_1, \ldots, x_{i-1}) \text{ until } x_i = \text{STOP}$$

  - How many different strings can we generate this way?

Adjusting Distribution Temperature

- Let's say we want something between sampling and greedy
  - Not fully deterministic
  - But to control how focused on the top of the distribution with high likelihood
- Add a temperature parameter to the $\mathrm{softmax}$
  - Given $\mathbf{h}$ is the vector with logits, and $T \in \mathbb{R}$ in the temperature

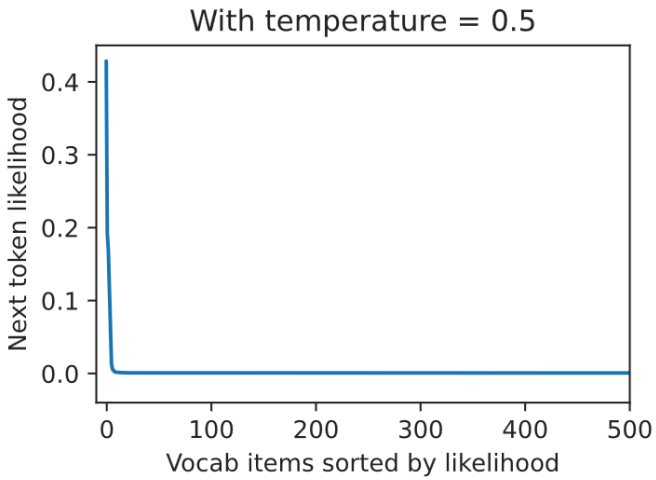
$$p_T(x_i = w) = \frac{\exp(h_w/T)}{\sum_{w'} \exp(h_{w'}/T)}$$

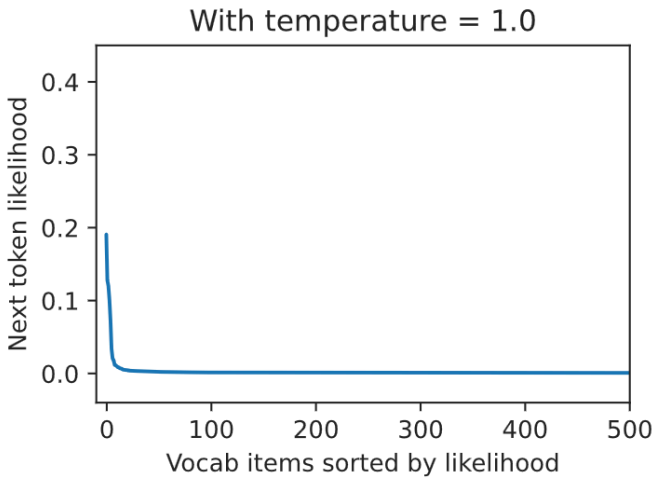
Adjusting Distribution Temperature

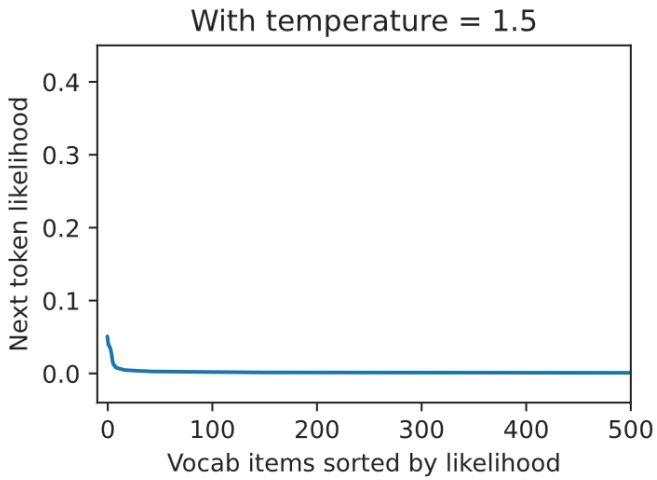
- Add a temperature parameter to the $\mathrm{softmax}$
  - Given $\mathbf{h}$ is the vector with logits, and $T \in \mathbb{R}$ in the temperature

$$p_T(x_i = w) = \frac{\exp(h_w/T)}{\sum_{w'} \exp(h_{w'}/T)}$$

  - What happens with $T = 1$? $T = 0$ (or almost)? $T \in [0,1)$? $T > 1$?







Plots from Daphne Ippolito / Chenyan Xiong, CMU LLMs course http://cmu-llms.org/

Other Decoding Techniques

- Top-k sampling
  - Drop everything but top-k tokens in the probability distribution, and re-normalize
- Nucleus sampling (Holtzman et al. 2020)
  - Drop everything but the top tokens that their probability sums to a specified value (e.g., 0.9) and re-normalize

Decoding

- Various decoding techniques: greedy, sampling, temperature-based, top-k, nucleus
- Most common: temperature-based
- Which are guaranteed to give you the optimal output? Will $\arg\max$ give you the optimal output?

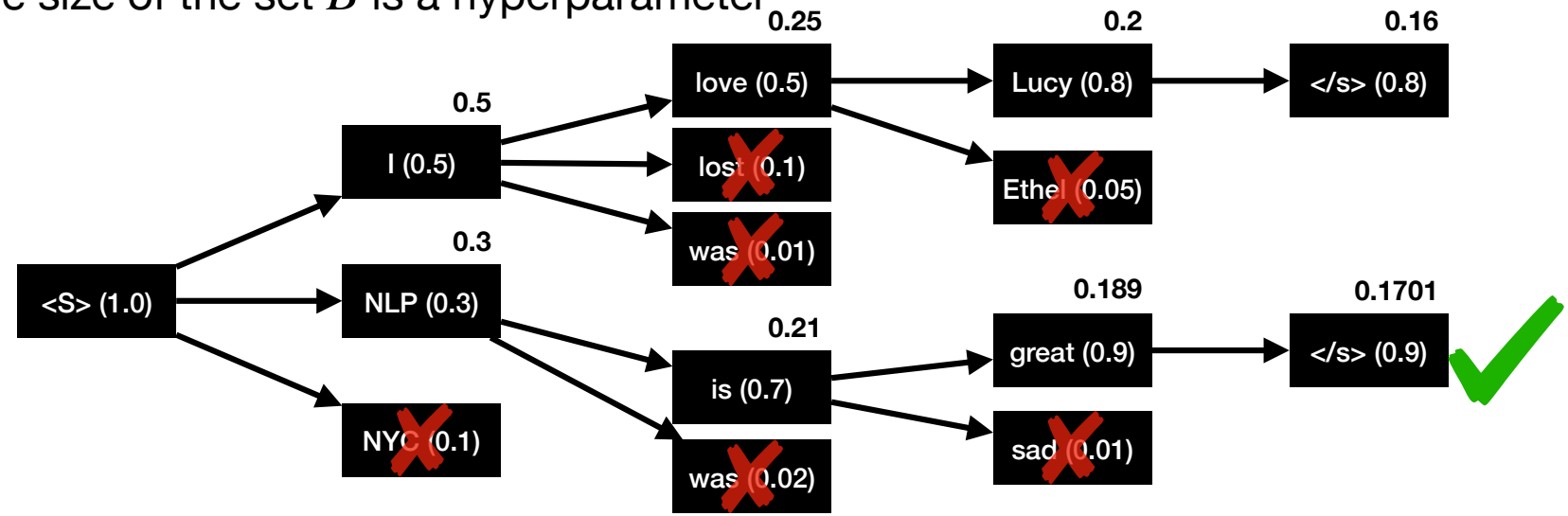
Decoding

- Various decoding techniques: greedy, sampling, temperature-based, top-k, nucleus
- Most common: temperature-based
- Which are guaranteed to give you the optimal output? Will $\arg\max$ give you the optimal output?

Output 1	0.2 I	0.5 love	0.1 Lucy	**0.01**
Output 2	0.2 I	0.1 hate	0.99 Lucy	**0.0198**

Beam Search

- Sampling techniques are not optimal
  - Following a single hypothesis is just not sufficient, but enumerating all is intractable
- Beam search is middle ground
  - Follow a set of hypothesis, always keeping the top ones
  - The size of the set B is a hyperparameter



Beam Search

- Sampling techniques are not optimal
  - Following a single hypothesis is just not sufficient, but enumerating all is intractable
- Beam search is middle ground
  - Follow a set of hypothesis, always keeping the top ones
  - The size of the set B is a hyperparameter
  - It's an approximation method
  - What happens with $B = 1$? $B = \infty$?
  - What is the cost of beam search compared to the sampling techniques we saw?
  - Can you combine sampling techniques with beam search?

Acknowledgements

- These slides are based on slides from:
  - Berkeley's NLP class by Alane Suhr and Dan Klein
  - CMU's LLM class by Daphne Ippolito and Chenyan Xiong