Decoding Strategies

Stochastic Decoding

Learning goals

- Get to know different stochastic decoding strategies
- Learn about sampling with temperature, top-k sampling and top-p (nucleus) sampling

SAMPLING MOTIVATION

- **Diversity in Output**
  - *Creativity and Variation*: Sampling methods produce varied outputs for the same input, useful in creative applications like story generation and dialogue systems.
  - *Avoiding Repetition*: These methods are less likely to generate repetitive loops compared to deterministic methods.
- **Technical Advantages**
  - *Reduced Computational Cost*: Sampling methods are often cheaper than beam search, particularly for long sequences.
  - *Scalability*: They are easier to scale for large models and datasets due to simpler implementation and reduced computational demands.

SAMPLING (WITH TEMPERATURE) (1)

The next token is selected randomly based on its conditional probability distribution. To control the randomness of the output sequence, a temperature parameter can be applied to the softmax function

$$\sigma\left(z_i\right) = \frac{e^{\frac{z_i}{temp}}}{\sum_{j=1}^{N} e^{\frac{z_j}{temp}}}$$

- $temp \rightarrow \infty$: Output distribution $\approx$ Uniform distribution
- $temp \rightarrow 0$: Output distribution $\approx$ Point mass (Greedy search)

SAMPLING (WITH TEMPERATURE) (2)

Prompt: "Once upon a time"

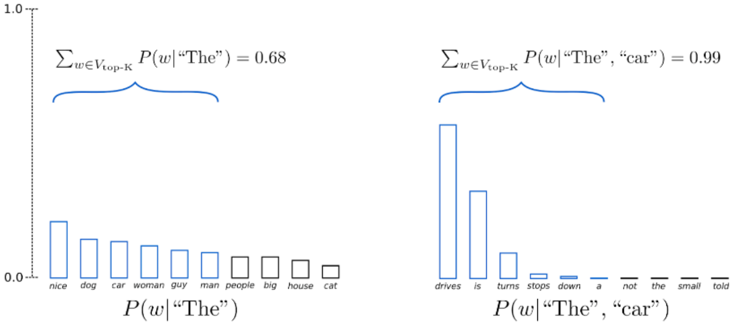
- Sampling with low temperature: *", during the Second World War, during the final months for his three most talented young players, the coach, Harry Gregg said this"*
- Sampling with high temperature: *"— well. Nowhere you call back my call, not on time; never the two on account my four. Do not come." This old woman — you might have liked, she herself — she did smile."*

The generated stories are diverse but sometimes very erratic.

$\Rightarrow$ Sample from the top-*k* tokens

TOP-*K* SAMPLING

In Top-*k* sampling, the *k* most likely next tokens are filtered, and the probability mass is redistributed. Visualization for $k = 6$ in two sampling steps:



▸ Hugging Face, Patrick von Platen

TOP-*K* SAMPLING

Prompt: "Once upon a time"

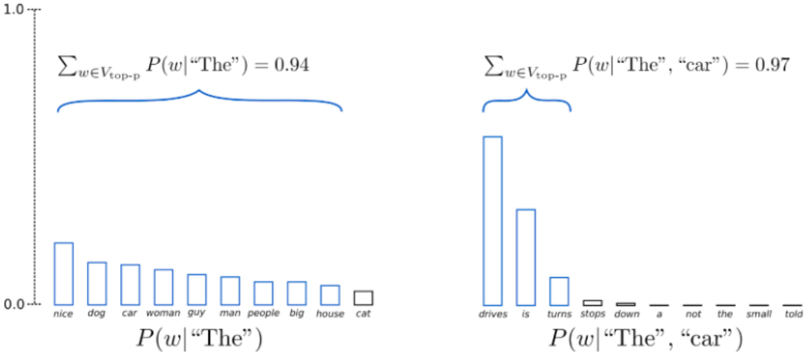
- Top-*k* , $k = 100$: *"when I was young the internet was a mysterious landscape full of new and exciting ideas. I read ebooks, watched videos, read short stories"*

The quality has improved, but the fixed *k* might be counterproductive

$\Rightarrow$ Make *k* dynamic

TOP-*P* (NUCLEUS) SAMPLING

Top-*p* sampling chooses from the smallest possible set of tokens whose cumulative probability exceeds the probability threshold *p*. The probability mass is then redistributed accordingly. Visualization with a threshold $p = 0.92$:



▸ Hugging Face, Patrick von Platen

TOP-*P* (NUCLEUS) SAMPLING

Prompt: "Once upon a time"

- Top-*p* , $p = 0.92$: *"there were four major political parties in the United States. Since then, however, they have become even more of a novelty. For the past few decades, there have been only two."*

SOTA for many years, default decoding strategy in various GPT versions, but sometimes erratic depending on *p* and the sampled tokens.

Question: Can there be a balance of coherence and diversity?
$\Rightarrow$ Contrastive search

CONTRASTIVE SEARCH

$$x_t = \underset{v \in V^{(k)}}{\arg\max} \left\{ (1-\alpha) \times \underbrace{p_\theta(v|\boldsymbol{x}_{<t})}_{\text{model confidence}} - \alpha \times \underbrace{(\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\})}_{\text{degeneration penalty}} \right\}$$

When generating output, contrastive search jointly considers:

- The probability predicted by the language model to maintain the semantic coherence between the generated text and the prompt.
- The similarity with respect to the previous context to avoid degeneration (as in Greedy or Beam search)

$\Rightarrow$ An "ideal" token should have a high probability and bring diversity to the story.

Empirical studies suggest $k \in \{5, 8, 10, 15\}$ and $\alpha \in \{0.4, 0.5, 0.6\}$

▸ Su & Collier, 2023 ▸ Su & Xu, 2022 ▸ Su et al., 2022