Decoding Strategies

Stochastic Decoding & CS/CD

Learning goals

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

SAMPLING MOTIVATION

- *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.

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"

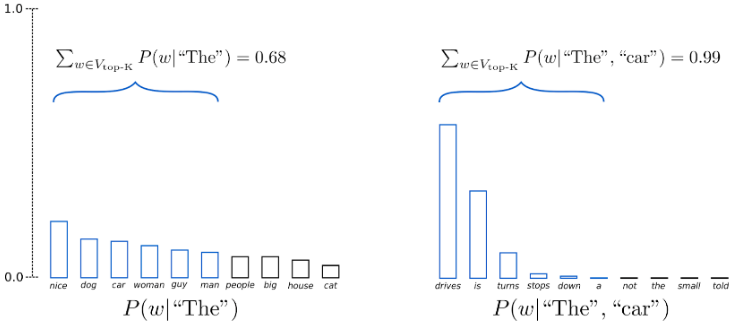
- Sampling with low temperature: *", during the Second World War, during the final months for his three most talented young players, the coach, Harry Gregg said this"*
- Sampling with high temperature: *"— well. Nowhere you call back my call, not on time; never the two on account my four. Do not come." 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 ▸ Fan et al., 2018

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:



▸ huggingface, Patrick von Platen

TOP-K SAMPLING

Prompt: "Once upon a time"

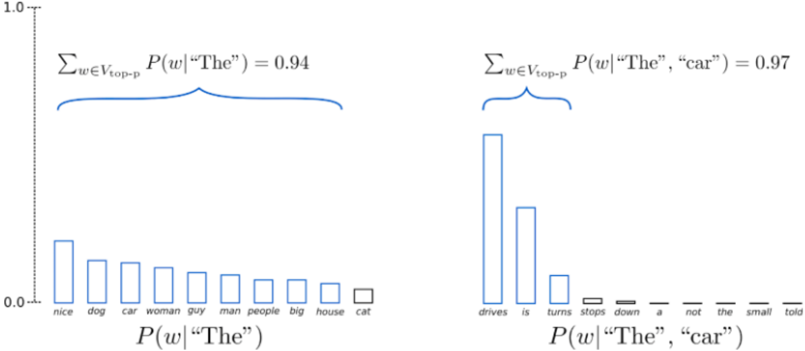
- Top-k , $k = 100$: *"when I was young the internet was a mysterious landscape full of new and exciting ideas. I read ebooks, watched videos, read short stories"*

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

$\Rightarrow$ Make k dynamic

TOP-*P* (NUCLEUS) SAMPLING ▸ Holtzman et al., 2019

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 $p = 0.92$:



▸ huggingface, 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 ▸ Su et al., 2022

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

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

CONTRASTIVE SEARCH FORMULA

▸ huggingface, Tian Lan

Let's have a closer look at the formula for Contrastive Search:

$$x_t = \underset{v \in V^{(k)}}{argmax} \left\{ (1-\alpha) \times p_\theta(v|\mathbf{x}_{<t}) - \alpha \times (max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\}) \right\}$$

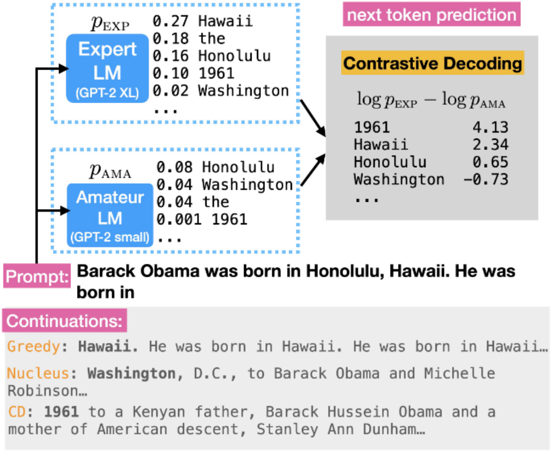
- x_t is the output token and $\mathbf{x}_{<t}$ the context
- $V^{(k)}$ is the set of top-k predictions from the models probability distribution (this is the same k as in the top-k sampling from earlier)
- $p_\theta(v|\mathbf{x}_{<t})$, the *model confidence*, is the probability of a candidate token v given the context
- $max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\}$, the *degeneration penalty*, measures how similar v is to the context, $s()$ is the cosine similarity between the token representations

CONTRASTIVE SEARCH FORMULA

- The degeneration penalty is defined as the maximum cosine similarity between the token representation of v, i.e h_v, and of all tokens in the context $\mathbf{x}_{<t}$
- h_v is computed by the language model given the concatination of v and $\mathbf{x}_{<t}$
- In order to maximize the formula we want v to have a high probability and a low degeneration penalty
- Intuitively, a larger degeneration penalty of v means it is more similar (in the representation space) to the context, therefore more likely leading to the problem of model degeneration
- α determines how much weight to give to each component
- For $\alpha = 0$ we only consider the probability and contrastive search becomes greedy search

CONTRASTIVE DECODING

Li et al., 2023



p_{EXP}

Expert LM (GPT-2 XL)

```
0.27 Hawaii
0.18 the
0.16 Honolulu
0.10 1961
0.02 Washington
...
```

p_{AMA}

Amateur LM (GPT-2 small)

```
0.08 Honolulu
0.04 Washington
0.04 the
0.001 1961
...
```

next token prediction

Contrastive Decoding

$\log p_{\text{EXP}} - \log p_{\text{AMA}}$

```
1961          4.13
Hawaii        2.34
Honolulu      0.65
Washington   -0.73
...
```

Prompt: **Barack Obama was born in Honolulu, Hawaii. He was born in**

Continuations:

Greedy: **Hawaii.** He was born in Hawaii. He was born in Hawaii…

Nucleus: **Washington**, D.C., to Barack Obama and Michelle Robinson…

CD: **1961** to a Kenyan father, Barack Hussein Obama and a mother of American descent, Stanley Ann Dunham…

- Contrastive decoding exploits the contrasts between expert and amateur LM of different sizes by choosing tokens that maximize their log-likelihood difference (read the paper if you are interested, not going into more detail here!)