Nucleus Sampling for Open-ended and Directed Text Generation : "The Curious Case of Neural Text DeGeneration" by Holtzman et al. [1]

Thomas Boudras, Mathilde Kretz, and Alexandre Ngau

Université Paris Dauphine

January 24, 2024

Introduction

- LLM topic concerned → decoding.
- Several methods are already in use:
  - Greedy search
  - Beam search
  - Top-k sampling
- New method → Nucleus sampling/Top-p sampling.

Greedy search and Top-k sampling

- Top-k sampling:
  - Select the k most likely sequences
  - Create a new distribution from these k tokens
  - Select the next token with this new distribution
- Greedy search:
  - Select the next token: most likely to follow the previous sequence
  - Special case of top-k sampling with $k = 1$

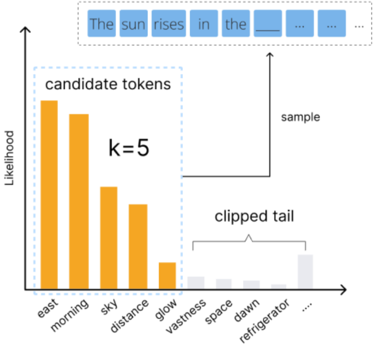

Figure 1: Top-k sampling diagram [1]

[1] Image source : https://www.megaputer.com/mastering-language-models-a-deep-dive-into-input-parameters/

Beam search

- The starting point is the repetition of w times the prompt
- For each of these prompts, create w possible sequences with the w most likely tokens
- We keep the w most likely suites/beams cumulatively
- We repeat from the 2nd point
- In the end, we keep the most likely cumulative result

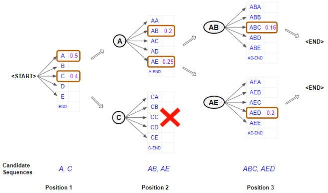

Figure 2: Beam search diagram

2

[2] Image source: https://www.megaputer.com/mastering-language-models-a-deep-dive-into-input-parameters/

Nucleus Sampling

- The most probable tokens are selected one by one until the sum of their associated probabilities is p
- Create a new distribution from these tokens
- Select the next token with this new distribution

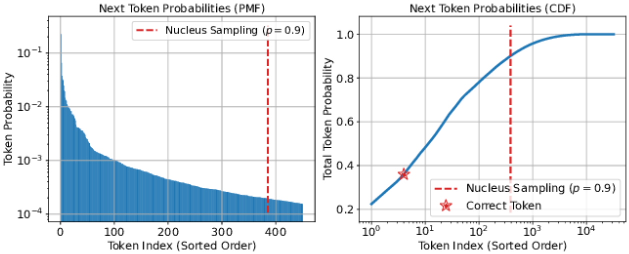

Figure 3: Beam search diagram

3

[3] Image source: https://www.researchgate.net/figure/Nucleus-sampling-can-have-unpredictable-behavior-and-lead-to-many-low-probability-tokens_fig3_370841754

Perplexity

- Goal : test the quality and consistency of text generation
- Based on the tokens by tokens probability of occurence : negative loglikelihood probability

$$Perplexity = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(n_i)\right)$$

- The lower the perplexity, the more likely (not necessarily diversified)

Results obtained for text generation 1/2

Parameters	Results	Perplexity
`Context`	Out on one of the busiest Colorado highways north of Denver in 2009, there were 57 bicyclists found dead, and eight of those were cyclists, according to victim recovery agencies.	
`origin text`	Surveillance video that showed injuries to 11 people was released today. Three women who died of forgotten injuries showed up drooling, bruising and constipation. [...]	
`greedy`	The Colorado Department of Transportation said it was investigating the deaths. The Colorado Department of Transportation said it was investigating the deaths. The Colorado Department of Transportation said it was investigating the deaths. The Colorado Department of Transportation said it was	2.73
$k = 20$	The Colorado Department of Transportation said the fatalities involved the following incidents: - A bicyclist died after sustaining a head injury that struck him or her head on the road near Lakeland on Tuesday. A pedestrian was also killed following a traffic accident,	9.602
$k = 640$	Another cyclist, who never spent a day in jail, ended up living in a van behind Morrison Hill as a consolation prize in 2008, through last year. The judge said it went too far to call that irony. Justice Bjergsen's injunction against	29.857

Table 1: Results of generating 50 tokens for various settings. Repetitions, Incoherences.

Results obtained for text generation 2/2

Parameters	Results	Perplexity
$k = 40$, $t = 0.7$	A bicyclist was found dead in the backseat of a car on the same highway, near the Denver International Airport. Authorities say the body was not identified. The crash happened in the southbound lanes of Interstate 84 near the intersection of U	4.00
`beam search`, $w = 16$	According to the Colorado Department of Public Safety, more than 1,000 bicyclists were killed in Colorado in 2009, and more than 1,000 bicyclists were killed in Colorado in 2010, according to the Colorado Department of Public Safety	2.16
$p = 0.7$	The two deaths of bicyclists occurred at Route 24 and 52, near the University of Colorado Boulder campus.Those incidents occurred on Feb. 2, 2009, and May 6, 2009, respectively. Those incidents occurred in both years.	11.13
$p = 0.95$	The scenario is "nice to think of when you're growing up," said Santa Fe Police Chief Jeff Garibaldi. Garcaldi said he and the department have heard that there have been a handful of crashes and a string of accidents involving	34.173

Table 2: Results of generating 50 tokens for various settings. Repetitions, Incoherences.

Novelty approach : Nucleus Sampling for Translation

- A new approach: using nucleus sampling for translation
- Results :

Text Type	Text	Perplexity
Input English Sentence	And yet halting the payments did not stabilize German politics.	
Target French Translation	Et pourtant, l'arrêt des paiements n'a pas stabilisé la politique allemande.	1677.87
Beam Search ($b = 4$) Translation	Et pourtant, arrêter les paiements n'a pas stabilisé la politique allemande.	2063.07 ($ADP = 385.20$)
Nucleus Sampling ($p = 0.7$) Translation	Et pourtant l'arrêt des paiements n'a pas stabilisé la politique allemande.	3763.16 ($ADP = 2085.29$)

Table 3: Example of a worse absolute difference perplexity and a better translation

Improving our work and conclusion

- Improvements :
  - Calculate the perplexity of the original text (if possible)
  - Introduce new human quality measure
  - Continue to develop translation
- Conclusion :
  - Open Generation works well
  - We're still working on the translation

References

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” 2020.