AI539 Natural Language Processing with Deep Learning – Homework 3

Decoding Language Models by Sampling and Search

Ameyassh Nagarajan

1.0 Written Responses

▶ TASK 1.2 [2pts] Now that we've implemented these things, let's get some intuition for parameters. When `decoder.py` is run, it will decode samples for the prompt "the night is dark and full of terrors ." for the following sampling settings. Note that the random seed is reset before each. 2

vanilla

temperature-scaled $\tau = 0.0001$

temperature-scaled $\tau = 100$

top-k, k=1

top-k, k=20

top-p, p=0.001

top-p, p=0.75

top-p, p=1 Provide your outputs for these runs in your report. These are random algorithms but given the same random seed, the samples for (2), (4), and (6) should nearly always be the same [a]. Likewise (1) and (8) should be the same. Argue why this should be an expected results.

[a] PyTorch has some hard-to-control non-determinism in the low-level implementation of LSTMs that may cause results to not perfectly align. Different systems / CUDA versions may also introduce noise.

1.1 Vanilla Sampling

Vanilla sampling chooses the next word based solely on the highest probability from the softmax distribution of logits, acting deterministically given a constant model state and input.

1.2 Temperature-Scaled Sampling

Temperature-scaled sampling modifies the distribution sharpness:

- A low temperature (0.0001) makes the model more deterministic, favoring likely words heavily.
- A high temperature (100) makes the distribution more uniform, introducing more randomness.

1.3 Top-k Sampling

This strategy limits the choices to the top k most probable words:

- For $k = 1$, the model is deterministic, selecting the most probable next word.
- For $k = 20$, the model introduces randomness but remains within the top 20 likely words.

1.4 Top-p Sampling

Also known as nucleus sampling, this method selects the smallest set of words whose cumulative probability exceeds p:

- For $p = 0.001$, the output is nearly deterministic.
- For $p = 0.75$, it considers a larger set, introducing more variety.
- For $p = 1$, it is identical to vanilla sampling since no words are excluded.

1.4 Discussion on Expected Results

Given the same random seed, certain outputs are expected to be nearly identical:

- Vanilla sampling and Top-p with $p = 1$ should produce identical results, as both consider the entire probability distribution without exclusion.
- Temperature-scaled with $\tau = 0.0001$, Top-k with $k = 1$, and Top-p with $p = 0.001$ should produce similar outputs, as all restrict the choice significantly towards the most probable outcomes.

1.4 Conclusion

The degree of randomness and diversity in the generated text is directly influenced by the parameters (τ, k, p) of the sampling method. Controlled experiments demonstrate that deterministic settings produce consistent results across runs, validating our understanding of these sampling strategies in neural text generation models.

▶ TASK 2.1 [15pts] Implement the `beamsearch` function in the `decoder.py` skeleton code. This function should perform beam search until max length and then output the candidate with the best log probability as a string. The skeleton code for this function is show below:

```
def beamsearch(model, text_field, beams=5, prompt="", max_len=50)
    :

  .
  .
  .

  return decodedString
```

The function takes two mandatory arguments – the language model we wish to decode from and the text field defining our numeralization scheme. Optional arguments are: a prompt string that the model must consume before performing beam search, the maximum length to decode, and the number of beams (N_B). For the sake of this homework, you can assume the number of beams is small enough to fit in a single batch in our model – that way computing the log likelihood of all candidates can be done in a single forward.

Once you've implemented beam search, running `decoder.py` will also execute three beam searches with N_B=1,10,50. Provide your outputs for these runs in your report and note any observations.

1.5 Beam=1

With only one beam, the model generated text that is highly coherent and closely follows the most probable path. This setting tends to produce text that is consistent and relevant to the initial prompt but lacks diversity.

> "the night is dark and full of terrors from ..." (Continues with coherent narrative)

1.6 Beam=10

Increasing the beams to 10 introduced more diversity in the text. The narratives generated were richer and contained a wider variety of concepts, though at times at the expense of coherence.

> "the night is dark and full of terrors from russet and blood and the hope ..." (Shows diverse vocabulary and themes)

1.7 Beam=50

At 50 beams, the output was the most diverse, exploring various narrative paths and utilizing a broad vocabulary. However, this often resulted in fragmented text with decreased relevance to the prompt.

> "the night is dark and full of terrors from unk , massive word. the coast ..." (Illustrates maximal diversity but reduced coherence)

1.7 Outputs for beam search implementation

———— Beam Search B=1 ————
the night is dark and full of terrors from russet , " said , " if it please you . . . by a reason . " " wait between the roofs of braavos . the black dread led a sortie and took her part on the bloody mummers . she had no choice , no danger nor dragons on , and dany did not wear his grandchildren to help her and her men at the citadel . perhaps the gods will still be sleeping at the king's wedding . " " i will not ask that , " she promised . " i will not suffer daario himself to avenge her , i know . . . " i am sorry , catelyn thought . then only to bronn . and when the girl meets the water , she raised her arms and kiss her tongue and beg to be carried on her . " i am
———— Beam Search B=10 ————
the night is dark and full of terrors from russet and blood and the hope . . . . . " called haggo , old eyes said yellow eyes . . a half - haired sick and a man yurkhaz zo galare on the block of the dagger he could accept the torch and more than out he had on the river past bones and worst of them all the gods remained . . . . . " you . . . . . a maester commander tarly , and kneelers seemed to hear . " pentos ? " a lannister . . . . . " the killer , he foretold you <unk> . . . . . joffrey bound his son . she is narbo's waking the truth . . . . . " courtier's . . . " seasons of them , and trading boats to the bear
———— Beam Search B=50 ————
the night is dark and full of terrors from <unk> , massive word . the coast , or genuine and jon's own eyes closed wide the slayer . . . " as most of you . . . . . and dagmer - at the way he may weigh , ned stark and . . . . . " ned suspected . . . " light jon's harridan , " by the good company and pander . " and dignity . . . . . " pleased him , lord walder frey's men and voice . . . . . " swift as it was one . . . . . the dothraki arakh of all you were starved . . . . . " the way . . . . . "

1.7 Conclusion

The number of beams in a beam search significantly impacts the diversity and coherence of the generated text. While a single beam ensures coherence, increasing the number of beams enhances diversity but may reduce the text's overall coherence and relevance.