CS726 Programming Assignment – 3 Report

Saksham Rathi (22B1003)
Sharvanee Sonawane (22B0943)
Deeksha Dhiwakar (22B0988)
Department of Computer Science,
Indian Institute of Technology Bombay

1 Task 0: Introduction to LLM Decoding Techniques

Here is how, we are getting the logits from the model. To speed up the process, we are also using cache, which stores the previous model run, and passes it for the next run. This speeds up the process by roughly 5 times.

```
for i in range(self.max_output_len):
  outputs = self.model(
      input_ids=current_ids,
      past_key_values=past_key_values,
      use_cache=True
  )
  logits = outputs.logits
  past_key_values = outputs.past_key_values
  logit_last_token = logits[:, -1, :]
```

1.1 Greedy Decoding

Here is how greedy decoding is implemented:

```
next_token = torch.argmax(logit_last_token, dim=-1)
```

That is, we are taking the token corresponding to the maximum probability (dim = -1 stores the probabilities across the vocabulary).

Here are some sample runs:

```
Example: 1/50

Reference: an appearance is a bunch of attributes related to the service person like their
shoes clothes tie jewellery hairstyle makeup watch cosmetics perfume etc

Ground Truth: service is the combination of many qualities for people such as their
clothes shoes ties accessories makeup hairstyle cosmetics etc


Example: 10/50

Reference: send rama with the sage and send lakshmana too

Ground Truth: ram with the sage and lakshman also send
```

Here are the final score values:

```
BLEU: 0.31440443213296393
ROUGE-1: 0.3571874622955566
ROUGE-2: 0.13222295518311183
ROUGE-LCS: 0.27441622904852214
```

1.2 Random Sampling with Temperature Scaling

Here is how random sampling is implemented with temperature τ:

```
  next_token_logits = next_token_logits / self.tau
  probs = torch.softmax(next_token_logits, dim=-1)
  next_token = torch.multinomial(probs, num_samples=1)
```

1.2.1 $\tau = 0.5$

Sample run:

```
Example: 11/50

Reference: but mangal pandeys brave deed was done through devotion to a high and
noble principle

Ground Truth: parantu mangal pande ne yah saahsik kaarnama aeek unche aur shreshtha
siddhant ke pratip sampan ke liye kiya
```

Scores:

```
BLEU: 0.2838258164852255
ROUGE-1: 0.28984430881105616
ROUGE-2: 0.10543762417930613
ROUGE-LCS: 0.22789346051514345
```

1.2.2 $\tau = 0.9$

Sample run:

```
Example: 23/50

Reference: indus valley civilisation is known for its technological knowledge
in a variety of fields

Ground Truth: trickle of technology reach down to the people in the different
regions of sindh ghat
```

Scores:

```
BLEU: 0.15806715806715804
ROUGE-1: 0.1595327517775904
ROUGE-2: 0.035088629933956283
ROUGE-LCS: 0.1170539764521488
```

1.3 Top-k Sampling

Here is how Top-k sampling is implemented by selecting the greatest k logits:

```
topk_logits, topk_indices = torch.topk(logit_last_token, self.k, dim=-1)
topk_probs = nn.functional.softmax(topk_logits, dim=-1)
next_token_idx = torch.multinomial(topk_probs, num_samples=1)
next_token = topk_indices.gather(-1, next_token_idx).squeeze(-1)
```

This method first selects the top-k logits (highest probabilities) from the vocabulary, normalizes them using softmax to form a probability distribution, and then samples the next token from this distribution.

1.3.1 $k = 5$

Sample run:

```
Example: 9/50

Reference: each city in punjab has varied preferences like people in amritsar are
particularly fond of amritsari kulche stuffed paranthas and milk products

Ground Truth: "every city of punjab is loved by everyone just like amritsar people
are special amritsar kulche barvaan parathas and milk products"
```

Scores:

```
BLEU: 0.24757281553398058
ROUGE-1: 0.24637532114470517
ROUGE-2: 0.06982793622619132
ROUGE-LCS: 0.18258561511936278
```

1.3.2 $k = 10$

Sample run:

```
Example: 4/50

Reference: ashoka started making extensive use of stone for sculptures and great
monuments whereas the previous tradition consisted of working with wood and clay

Ground Truth: ashok started to create statues sculptures and spectacular monuments
by using stones whereas the old traditional way is the use of wood and earth
```

Scores:

```
BLEU: 0.24899274778404515
ROUGE-1: 0.2583936169635166
ROUGE-2: 0.07972418861869043
ROUGE-LCS: 0.1984233891953222
```