**EE 599: Systems for Machine Learning**
**Final Year Project - Phase 2**

# Efficient Fine-tuning of LLM on a Single GPU

*Sheik Dhawood Ashfaq Asick Ali*
*Sam Devavaram Jebaraj*

Time taken with KV caching: 48.36289 seconds
Time taken without KV caching: 93.8209 seconds

The presence of KV caching likely speeds up inference by reducing the time required for key-value pair retrieval, enhancing overall efficiency. The variation in outputs could be attributed to differences in GPU computation kernels or Python package versions. These variances can lead to differences in floating-point calculations, potentially affecting the output.

Given Prompts =

```
 [
           # For these prompts, the expected answer is the natural
continuation of the prompt
        "I believe the meaning of life is",
        "Simply put, the theory of relativity states that ",
        """A brief message congratulating the team on the launch:

        Hi everyone,

        I just """,
          # Few shot prompt (providing a few examples before asking
model to complete more);
        """Translate English to French:

        sea otter => loutre de mer
        peppermint => menthe poivrée
        plush girafe => girafe peluche
        cheese =>""",
    ]
```

Changes in code:

**In Model Py**: Commenting off the KV Caching in the Attention Module and directly compute keys and values from the input tensors.

```
        super().__init__()
        self.n_kv_heads = args.n_heads if args.n_kv_heads is None else args.n_kv_heads
        self.n_local_heads = args.n_heads
        self.n_local_kv_heads = self.n_kv_heads
        self.n_rep = self.n_local_heads // self.n_local_kv_heads
        self.head_dim = args.dim // args.n_heads

        self.wq = nn.Linear(args.dim, args.n_heads * self.head_dim, bias=False)
        self.wk = nn.Linear(args.dim, self.n_kv_heads * self.head_dim, bias=False)
        self.wv = nn.Linear(args.dim, self.n_kv_heads * self.head_dim, bias=False)
        self.wo = nn.Linear(args.n_heads * self.head_dim, args.dim, bias=False)

        '''self.cache_k = torch.zeros(
            (
                args.max_batch_size,
                args.max_seq_len,
                self.n_local_kv_heads,
                self.head_dim,
            )
        ).cuda()
        self.cache_v = torch.zeros(
            (
                args.max_batch_size,
                args.max_seq_len,
                self.n_local_kv_heads,
                self.head_dim,
            )
        ).cuda()'''
```

```
        '''self.cache_k = self.cache_k.to(xq)
        self.cache_v = self.cache_v.to(xq)

        self.cache_k[:bsz, start_pos : start_pos + seqlen] = xk
        self.cache_v[:bsz, start_pos : start_pos + seqlen] = xv

        keys = self.cache_k[:bsz, : start_pos + seqlen]
        values = self.cache_v[:bsz, : start_pos + seqlen]'''
        keys = xk
        values = xv
```

**In Generation Py**: Making the previous position to be 0, thus it doesn't update and no slicing occurs.

```
        for cur_pos in range(min_prompt_len, total_len):
            with torch.no_grad():
                logits = self(tokens[:, prev_pos:cur_pos], prev_pos)
            if temperature > 0:
                probs = torch.softmax(logits[:, -1] / temperature, dim=-1)
                next_token = sample_top_p(probs, top_p)
            else:
                next_token = torch.argmax(logits[:, -1], dim=-1)

            next_token = next_token.reshape(-1)
            # only replace token if prompt has already been generated
            next_token = torch.where(
                input_text_mask[:, cur_pos], tokens[:, cur_pos], next_token
            )
            tokens[:, cur_pos] = next_token

            eos_reached |= (~input_text_mask[:, cur_pos]) & (
                next_token == tokenizer.eos_id
            )

            #prev_pos = cur_pos
```

| Without KV Caching | With KV Caching |
|---|---|
| I believe the meaning of life is<br>> to learn to love.<br>Love is not a feeling. It is a decision.<br>I believe the meaning of life is to learn to love.<br>Love is not a feeling. It is a decision. It is a commitment. It is a conscious choice of the will and the intellect.<br>There are many<br><br>==========<br><br>Simply put, the theory of relativity states that<br>> 1) the speed of light is constant for all observers and 2) the laws of physics are the same for all observers.<br>The theory of relativity is a very important concept in physics, but it is also one of the most misunderstood.<br>There are a lot of misconceptions about<br><br>==========<br>A brief message congratulating the team on the launch:<br><br>Hi everyone,<br><br>I just<br>><br><br><a href="https://www.google.com">Google</a><br><br>your website.<br><br>I hope you enjoy the new look and feel.<br><br>I'll be in touch soon to discuss your next project.<br><br>Best<br>==========<br><br>Translate English to French: | I believe the meaning of life is<br>> to learn to love.<br>Love is not a feeling. It is a decision.<br>I believe the meaning of life is to learn to love.<br>Love is not a feeling. It is a decision. It is a commitment. It is a conscious choice of the will and the intellect.<br>There are many<br><br>==========<br><br>Simply put, the theory of relativity states that<br>> 1) the speed of light is constant for all observers and 2) the laws of physics are the same for all observers.<br>The theory of relativity is a very important concept in physics, but it is also one of the most misunderstood.<br>There are a lot of misconceptions about<br><br>==========<br>A brief message congratulating the team on the launch:<br><br>Hi everyone,<br><br>I just<br>><br><br><a href="https://www.google.com">Google</a><br><br>your website.<br><br>I hope you enjoy the new look and feel.<br><br>I'll be in touch soon to discuss your next project.<br><br>Best<br>==========<br><br>Translate English to French: |

```
        sea otter => loutre de mer
        peppermint => menthe poivrée
        plush girafe => girafe peluche
        cheese =>
> fromage
        penguin => pinguin
        handbag => sac à main
        mug => tasse
        chocolate => chocolat
        chocolate => chocolat
        chocolate => chocolat
        chocolate => choc

==================================
```

```
        sea otter => loutre de mer
        peppermint => menthe poivrée
        plush girafe => girafe peluche
        cheese =>
> fromage
        penguin => pinguin
        handbag => sac à main
        mug => tasse
        toothpaste => dentifrice
        t-shirt => tee-shirt
        pencil => crayon
        parrot => perroquet

==================================
```

**Custom prompt**: “Best Sunset hike in Los Angeles”

**Time taken with KV caching:** 32.03430461883545 secs

**Time taken without KV caching:** 35.742812633514404 secs

**Both Outputs:**

Best Hiking place in Los Angeles for sunset
>
L.A. is known for its beautiful sunsets, and there are many hiking trails in the city that offer stunning views of the setting sun. Here are some of the best hiking places in Los Angeles for sunset:
1. Griffith Park: Griffith Park is a