EE599 Project Final Report

Section 2: Background

Answers:

1. The four main aspects are: pre-training (how to pretrain a capable LLM), adaptation (how to effectively adapt pre-trained LLMs for better use), utilization (how to use LLMs for solving various downstream tasks) and capability evaluation (how to evaluate the abilities of LLMs and existing empirical findings).

2. When the parameter scale exceeds a certain level, these enlarged language models not only achieve a significant performance improvement but also show some special abilities that are not present in small-scale language models. To discriminate the difference in parameter scale, the term large language models (LLM) is for the pre-trained language model (PLM) of significant size.

3.

- Encoder-decoder Architecture: it consists of two stacks of Transformer blocks as the encoder and decoder, respectively. The encoder adopts stacked multi-head self-attention layers to encode the input sequence for generating its latent representations, while the decoder performs cross-attention. Examples include the vanilla transformer.
- Causal Decoder Architecture: This architecture incorporates the unidirectional attention mask, to guarantee that each input token can only attend to the past tokens and itself. Examples include GPT-series models.
- Prefix Decoder Architecture: This architecture revises the masking mechanism of causal decoders, to enable performing bidirectional attention over the prefix tokens and unidirectional attention only on generated tokens. Examples include GLM-130B and U-PaLM.

4. Language modeling is a probabilistic models to predict the next word or token in a sequence, given the preceding words or tokens. Causal language modeling predicts the next word based on previous words. Mainly used for text generation. Masked language modeling predicts a masked word in a sentence using its context. Mainly uses this technique for pre-training.

5. Text classification, is a task in Natural Language Processing, where a given piece of text is assigned to one or more predefined categories or labels. The goal is to automatically classify the text into one or several categories based on its content. Various models could be used depending on the specific tasks, including traditional ones(SVM, etc.) based on TF-IDF, RNN, LSTM, and transformers.

   When using models like BERT or other transformer-based models for a new downstream text classification task, it's common to retrain or "fine-tune" the classification head. For a specific classification task, classification head (a few dense layers) would typically be added to the pre-trained model, and get fine-tuned to adjust the weights.

6. Summarization task involves condensing a longer piece of text into a shorter version, retaining only the most crucial information. The goal is to capture the essence of the original content in fewer words. T5, BERT and GPT can be fine-tuned for summarization tasks.

7. Adam and AdamW are optimizers based on adaptive estimates of lower-order moments for first-order gradient-based optimization. They are widely used to adjust the learning rates of each parameter based on the historical information of its gradients. This adaptive learning rate often leads to faster convergence and less sensitivity to the initial learning rate setting, which could save time, memory and energy, especially in large-scale traing processes. Due to its need to apply first-order and second-order momentums to further adjust parameters dynamically, it would introduce overhead. They need to maintain moving averages of past gradients and past squared gradients. This means for each parameter, Adam and AdamW store two additional values, leading to a total memory requirement of 3N.

8. Learning rate scheduler is used to adjust the learning rate during training based on the number of epochs or the number of batch iterations. Sometimes starting with a higher learning rate and reducing it later can help the model converge faster and achieve a better final performance. The torch.optim.lr_scheduler.CosineAnnealingLR scheduler in PyTorch modifies the learning rate using a cosine annealing schedule. It starts with an initial learning rate, decreases it following a cosine curve until it gets near a minimum value (eta_min), and repeats this for T_max epochs. This approach can help navigate the loss landscape and potentially avoid local minima.

9. Tokenization is a preprocessing method widely used in NLP. It aims to segment raw text into sequences of individual tokens, which are subsequently used as the inputs of LLMs. In traditional NLP research, word-based tokenization is the predominant approach, which is more aligned with human's language cognition. However, wordbased tokenization can yield different segmentation results for the same input in some languages (e.g., Chinese word segmentation), generate a huge word vocabulary containing many low-frequency words, and also suffer from the "outof-vocabulary" issue. Thus, several neural network models employ character as the minimum unit to derive the word representation (e.g., a CNN word encoder in ELMo).

   Recently, subword tokenizers have been widely used in Transformer based language models, typically including BytePair Encoding tokenization, WordPiece tokenization and Unigram tokenization.

10. LLaMA was trained on a diverse dataset including English CommonCrawl(67%), C4(15%), Github(4.5%), Wikipedia(4.5%), Books(4.5%), ArXiv(2.5%), and StackExchange(2.0%). Training set for LLaMA 65B and LLaMA 33B is 1.4 trillion tokens. For smallest model, LLaMA 7B, is trained on one trillion tokens.

11. Perplexity is defined as the exponentiated average negative log-likelihood of a sequence.

    Given a sequence $X = (x_0, x_1, \ldots, x_t)$,

    $$PPL(X) = \exp -\frac{1}{t}\sum_{i}^{t} \log p_\theta(x_i|x_{<i})$$

    In the formula, each of the item $\log p_\theta(x_i|x_{<i})$ is the log-likelihood of the i-th token conditioned on the preceding tokens according to our model. Intuitively, it can be thought of as an evaluation of the model's ability to predict uniformly among the set of specified tokens in a corpus. Importantly, this means that the tokenization procedure has a direct impact on a model's perplexity which should always be taken into consideration when comparing different models.

12. Generating text from LLM can be done using various decoding methods.

- Greedy Search: Selects the highest probability word at each timestep, but may miss high probability words hidden behind a low probability word.
- Beam Search: Keeps the most likely `num_beams` hypotheses at each timestep, selecting the hypothesis with the highest overall probability, offering a more robust alternative to greedy search.
- Sampling: Picks the next word randomly based on its conditional probability distribution, potentially leading to incoherent text. Adjusting the `temperature` parameter can control the randomness.
- Top-K Sampling: Filters the K most likely next words and redistributes the probability mass among them. However, a fixed K may lead to sub-optimal choices of next words.
- Top-p Sampling: Dynamically chooses a set of words whose cumulative probability exceeds a threshold p, adapting to the next word's probability distribution.

13. Prompt learning refers to a method of training machine learning models, especially large language models like those in the GPT series. Instead of fine-tuning a model on a specific task with labeled data, prompt learning involves providing the model with a series of prompts or questions and having the model generate responses based on its pre-trained knowledge.

14. In-context learning refers to a process where a system learns from the specific environment or situation it's in, rather than from explicit instructions or pre-defined datasets. It allows for adaptive behavior based on immediate surroundings or experiences. In AI, it's exemplified by models that adjust their responses based on immediate user input or context.

15. These terms refer to the number of examples provided to the model to guide its response in the process of prompt learning.

- Few-shot: The model is given several examples to infer the desired task.
- One-shot: Only one example is provided.
- Zero-shot: No examples are given; the model is expected to understand the task solely from the prompt.

16. Instruction tuning is a method employed to refine the performance of LLM to better align with human preferences or specific tasks. LLaMA is fine-tuning on instructions data rapidly leads to improvements on MMLU(Massive Multitask Language Understanding). They observe that a very small amount of fine-tuning improves the performance on MMLU, and further improves the ability of the model to follow instructions. Instruction tuning is a form of supervised training because it involves providing the model with specific instructions and then adjusting the model based on the feedback or the desired outcomes.

17. As is mentioned in the descriptions, Alpaca is a dataset of 52,000 instructions and demonstrations generated by OpenAI's text-davinci-003 engine. This instruction data can be used to conduct instruction-tuning for language models and make the language model follow instruction better.

    Examples:

    ```
    Instruction: Explain why the following fraction is equivalent to 1/4.
    ```

    An instruction describes the task the model should perform. Each of the 52K instructions is unique.

```
Input: 4/16
```

An input is an optional context or input for the task. For example, when the instruction is "Summarize the following article", the input is the article. Around 40% of the examples have an input.

```
Output: The fraction 4/16 is equivalent to 1/4 because both numerators and
denominators are divisible by 4. Dividing both the top and bottom numbers by 4
yields the fraction 1/4.
```

This is simply the answer to the instruction as generated by text-davinci-003.

An extra section is text, which is the combination of the instruction, input and output formatted with the prompt template used by the authors for fine-tuning their models.

18. Human alignment in LLM is the process of adjusting the models to align with human values, minimizing biases, and ensuring that they are useful and not harmful to users. It's important as it improves usability, mitigates misunderstandings and biases, and allows for better control over the models in accordance with human expectations.

Section 3: Preliminary

3.1 Gradient Accumulation

1. For each mini batch, the loss function is

$$Loss_k = \frac{1}{8 \cdot 4} \sum_{i=8k+1}^{8k+8} (y_i - \hat{y}_i)^2, k = 0, 1, 2, 3$$

If we add the losses of all the batches together, it would be:

$$TotalLoss = \frac{1}{32} \sum_{i=1}^{32} (y_i - \hat{y}_i)^2$$

And that would be the same with the loss of the whole batch.

2. In essence, it is natural that the results may differ, since the mean and standard deviation might be different in each mini batch.

Consider $\hat{y}_i = Wx_i + b_i$.

If $x_i' = \frac{x_i - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}}$, then $\hat{y_i}' = W\frac{x_i - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} + b_i$.

In the loss function mentioned above,

$$Loss_k = \frac{1}{32} \sum_{i=8k+1}^{8k+8} (y_i - \hat{y_i}')^2, k = 0, 1, 2, 3$$

We cannot simply sum them up together to get the total loss.

3.2 Gradient Checkpointing

1. During model inference, we do not have to store the gradient used for calculations in the back propagation process, which might also involve gradient accumulation mentioned above.

2. The poor strategy: $O(1)$ memory, $O(n^2)$ computation steps.

   The general strategy: $O(\sqrt{n})$ memory, $O(n)$ steps.

3.3.1 SVD and truncated SVD

1. Matrix rank is the maximum number of linearly independent rows or columns in a matrix.

2. The Singular Value Decomposition (SVD) decomposes a matrix into three other matrices, A is represented as $A = U\Sigma V^T$:

- U: An orthogonal matrix containing the left singular vectors.
- Σ: A diagonal matrix containing the singular values.
- V^T: An orthogonal matrix containing the right singular vectors.

3. When an orthogonal matrix is multiplied by its transpose, the result is the identity matrix. This is a fundamental property of orthogonal matrices.

4. Rank is n

5. Keeping only the top-k singular values, it is essentially performing a form of lossy compression or dimensionality reduction on the image data. As k increases, more singular values are retained, leading to a less compressed representation of the image. The reconstructed image will be a closer approximation to the original image, retaining more of the original detail.

6. If the rank is r which r < n, The singular value matrix will have non-zero singular values only for the rank of r, and the rest will be zeros.

7. $A = U_k S_k, B = V_k^T$ where U_k is the first k columns of U, S_k is the first k rows and columns of S, and V_k^T is the first k rows of V^T. The reconstruction $W \approx AB$ is a good approximation especially when singular values beyond the top-k are near zero, as they contribute less to the structure of W.

8. Broad Spectrum of Singular Values: Slow decay of singular values can lead to significant information loss upon truncation. Near-Singular Matrices: Near-singular matrices have singular values that are very close to zero. When applying truncated SVD to such matrices, the method may fail to distinguish between relevant and irrelevant singular values, potentially leading to poor approximations. Noise Sensitivity: Truncated SVD can be sensitive to noise in the data. If a matrix has noisy data, the singular values corresponding to the noise may be mistaken for significant singular values.

3.3.2 LoRA

1. $r < n$

2. In the back-propagation, since $r << n$, the gradient computation for $n * r$ matrix A and $r * n$ matrix B has a much lower computational complexity compared to on the full $n * n$ matrix W.

3. Before inference, we can explicitly compute and replace W0 with $W = W_0 + AB$, after that it compute the same as the original model.

3.4 Mixed Precision Training

1. Directly using FP16 in every situation would cause a loss of precision, especially when the value of the gradients are very small. By maintaining an FP32 master copy of the weights, the updates are applied to this copy with higher precision. After the update, the FP32 weights can be again turned to FP16 for the following forward and backward passes.

   Although FP16 weight storages are additional, we could save up more space if we store activations in FP16, which always take up a significant portion of the memory. Changing from FP32 to FP16 for activations would directly cut down the memory requirements by half, and this would normally lead to a net decrease in memory usage.

Section 4: LLaMA Inference

4.3 Model Architecture

The KV cache structure is used to accelerate inference. It is removed according to the requirement.

```
# 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
```

Also modified the code for mask to adapt the change

```
# mask = torch.triu(mask, diagonal=start_pos + 1).type_as(h)
mask = torch.triu(mask, diagonal=1).type_as(h)
```

Apart from deleting code related to this part in `model.py`, we also modify part of the code in `generation.py` as below:

```
for cur_pos in range(min_prompt_len, total_len):

    logits = self.model.forward(tokens[:, 0:cur_pos], prev_pos)
```

in order to fetch all token values at one time.

For `fairscale.nn.model parallel.layers`, we replace them with corresponding `nn.Linear` layers.

```
# Attention
self.wq = nn.Linear(args.dim, args.n_heads * self.head_dim, bias=False)
self.wk = nn.Linear(args.dim, args.n_heads * self.head_dim, bias=False)
self.wv = nn.Linear(args.dim, args.n_heads * self.head_dim, bias=False)
self.wo = nn.Linear(args.n_heads * self.head_dim, args.dim, bias=False)
# FeedForward
self.w1 = nn.Linear(dim, hidden_dim)
self.w2 = nn.Linear(hidden_dim, dim)
self.w3 = nn.Linear(dim, hidden_dim)
# Transformer
self.tok_embeddings = nn.Embedding(params.vocab_size, params.dim)

self.output = nn.Linear(params.dim, params.vocab_size, bias=False)
```

4.5 Run Inference

To generate text by directly running `python -m ...`, we also disable distributed features. Specifically, we hardcode the environment variables in `generation.py`:

```
# Set environment variables for distributed training
os.environ['RANK'] = '0'
os.environ['WORLD_SIZE'] = '1'
os.environ['MASTER_ADDR'] = 'localhost'
os.environ['MASTER_PORT'] = '12345'
```

This is the command we use to run inference:

```
python example_text_completion.py --ckpt_dir /project/saifhash_1190/llama2-7b
        --tokenizer_path /project/saifhash_1190/llama2-7b/tokenizer.model
```

We test with some new texts.

Zero shot prompts

```
What is machine learning?
> How does it work?
A lot of people have heard about machine learning, but few know what it actually
is.
Machine learning is a type of artificial intelligence (AI) that allows computers
to
learn from data without being explicitly programmed.
Machine learning algorithms are used to make predictions based on patterns in
data. For
```

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

```
To maintain a healthy work-life balance, it's essential to
> have a healthy lifestyle. It is not always easy to maintain a healthy lifestyle,
especially when you have a busy schedule. However, it is important to make time
for yourself and your health. This article will discuss how to maintain a healthy
work-life balance with a healthy lifest
```

==========

Few shot prompts

```
Find synonyms for the following words:
        quick => fast
        happy => joyful
        difficult => challenging
        intelligent =>
> smart
        silly => dumb
        simple => easy
        fast => quick
        loud => noisy
        dumb => silly
        scary => spooky
        noisy => loud
        quiet => silent
        smart => intelligent
        challenging => difficult
        easy =>
```

==========

```
Match the event to the year:

        Moon Landing => 1969
        Fall of the Berlin Wall => 1989
        Invention of the Internet => 1983
        Start of World War I =>
> 1914
        Start of World War II => 1939
        Start of the Vietnam War => 1964
        Start of the Korean War => 1950
        Start of the Cold War => 1947
        Start of the Cuban
```

==========

Section 5: LLaMA Training

5.8 Analysis

LoRA settings

lora_rank = 16 lora_alpha = 32, lora_dropout = 0.05

Trainable parameter count

Total parameters: 6747639808 Trainable parameters: 8388608 Percentage of trainable parameters: 0.12%

GPU memory usage

40402MiB / 40960MiB

Gradient checkpoints

A checkpoint is placed during the forwarding process of the TransformerBlock as follows:

```
...
def forward(
        self,
        x: torch.Tensor,
        start_pos: int,
        freqs_cis: torch.Tensor,
        mask: Optional[torch.Tensor],
    ):
        h = x + self.attention.forward(
            self.attention_norm(x), start_pos, freqs_cis, mask
        )
        # Checkpoint
        def check_forward(x):
            return self.feed_forward.forward(self.ffn_norm(x))
        out = h + checkpoint(check_forward, h, use_reentrant=True)
        return out
...
```

Training loss for every epoch:

Epoch 0: Average Loss = 1.325083038618483 Epoch 1: Average Loss = 1.0072221773462813 Epoch 2: Average Loss = 0.7484890972584303 Epoch 3: Average Loss = 0.5021964082850953 Epoch 4: Average Loss = 0.3443939249542737

Comprehensive analysis

For each of the techniques applied (low precision, gradient accumulation, gradient checkpointing, LoRA), analyze whether or not it can contribute to memory reduction (in terms of model weight, model activation, model gradient, and optimizer state), computational cost, and potential impact on model accuracy.

Technique	Memory Reduction	Computational Cost Reduction	Impact on Accuracy
Low Precision	Reduce gradient and optimizer	Reduce loss calculations	Decrease accuracy
Gradient Accumulation	Reduce gradient storage only	Reduce gradient calculations	No impact on accuracy
Gradient Checkpointing	Reduce gradient storage	Increase optimizer and loss calculations	No impact on accuracy
LoRA	Reduce model gradient and optimizer state	Only calculate gradient for LoRA parameters	Decrease accuracy compared to training on all parameters

Example prompts and results (fine-tuned model)

Explain why a given topic is important. Environmental protection

> Environmental protection is essential for maintaining the health of our planet for future generations. Human activities such as emissions, farming, and deforestation are contributing to climate change, air and water pollution, reduction in biodiversity, and other environmental threats. By placing a priority on environmental protection, we can reduce the speed of environmental damage and conserve natural resources for our own and future generations' use.

==========

Rewrite the following sentence without changing the meaning. The sales report was not accurate.

> The sales report was inaccurate.

==========

Explain why the given definition is wrong. A mole is an animal that lives underground.

> This definition is wrong because moles are small, furry mammals that live above ground and are known for their distinctive long snouts and wide, black eyes. They are most commonly found in grassy areas and wooded areas.

==========

Generate a password of 8 characters.

> 1q2w3e4r s5t6u7v8

==========

Below are 4 random questions beyond the range of [0:200].

Name two different ways to reduce poverty in a community.

> 1. Improve education opportunities and access to healthcare, 2. Increase access to financing and investment opportunities, 3. Provide subsidies and incentives, 4. Foster entrepreneurship and

> creativity, 5. Create jobs and

==========

Convert the following number from base 10 to hexadecimal: 278

> 11 07 def 11 07 def def 11 07 def 11 07 def 11 07 def 11 07 def 11 07 def 11 07 def 11

==========

Suggest a suitable card game for a birthday party with 10 year old kids.

> - Casino Royale
> - War
> - Crazy Eights
> - Uno
> - Candy Land
> - Chutes and Ladders
> - Tic-Tac-Toe
> - Mancala
> - Memory
> - Happy Birthday

==========

Explain the difference between a personal computer and a server

> A personal computer (PC) is a general-purpose computing device designed for individual or home use. It is typically a small, portable device with a limited amount of memory and processing power. A server, on the other hand, is a more specialized and powerful computing device designed for use in business,

==========

Example prompts and results (original model)

Explain why a given topic is important. Environmental protection

> is a major issue that is faced by the world today. Many people are not aware of the negative impacts that they have on the environment. Environmental protection is important because it will help to conserve natural resources, improve the quality of the environment and also improve human health. Environmental protection is the process of protecting the environment from pollution and degradation. It involves the use of technology to reduce pollution and improve the quality of the environment. The importance of environmental protection is that it will help to conserve natural resources. Natural resources are the resources that are found in nature. These resources are used by humans for their survival. Natural resources include water, air, land, forests, minerals and animals. Natural resources are important because they are used by humans for their survival. They are also important because they are used by animals for their survival. The importance of environmental protection is that it will help to improve the quality of the environment. The quality of the environment is the condition of the environment. The quality of the environment is important because it will help to improve the quality of life. The quality of the environment is important because it will help to improve the

==================================

Rewrite the following sentence without changing the meaning. The sales report was not accurate.

> The sales report was not accurate. The sales report was not accurate

==================================

Explain why the given definition is wrong. A mole is an animal that lives underground.

> An animal that lives underground is an animal that lives in a hole in the ground. Explain why the given definition is wrong. A mole is an animal that lives in a hole in the ground. An animal that lives in a hole in the ground is an animal that lives underground. Q: Explain why the given definition is wrong. A mole is an animal that lives underground. A mole is an animal that lives underground. Q: Explain why the given definition is wrong. A mole is an animal that lives in a hole in the ground. An animal that lives in a hole in the ground is an animal that lives underground. Q: Explain why the given definition is wrong. A mole is an animal that lives in a hole in the ground. Q: Explain why the given definition is wrong. A mole is an animal that lives underground. Q: Explain why the given definition is wrong. A mole is an animal that lives in a hole in the ground. Q: Explain why the given definition is wrong. A mole is an animal that lives in a hole

==================================

Generate a password of 8 characters.

> I have a password field in a form, and I want to generate a password of 8 characters. I'm trying this code, but it is not working:
>
> ```
> \begin{code}
> function generatePassword() {
> var length = 8;
> var chars = "ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789";
> var result = "";
> for (var i = 0; i < length; i++) {
> result += chars.charAt(Math.floor(Math.random() * chars.length));
> }
> return result;
> }
> document.getElementById("password").value = generatePassword();
> \end{code}
> ```
>
> I'm trying to generate a password of 8 characters. But I'm getting this error: Uncaught TypeError: Cannot read property 'length' of undefined Comment: Your code is missing the closing `)` on the

> return statement Comment: `var length = 8;

==========

Correspondingly, below are the 4 answers to the questions out of [:200].

Name two different ways to reduce poverty in a community.

> Name two different ways to reduce poverty in a community.
>
> 1. How does the government help reduce poverty?
> 2. What are some things that people can do to help reduce poverty?
> 3. What are some ways to help reduce poverty?
> 4. What are some ways that people

==========

Convert the following number from base 10 to hexadecimal: 278

> Convert the following number from base 10 to hexadecimal: 278. Convert the following number from base 10 to hexadecimal: 278.

Solution:

```
import math
def convert_to_hexadec
```

==========

Suggest a suitable card game for a birthday party with 10 year old kids.

> I am looking for a suitable card game for a birthday party with 10 year old kids. I have heard of Uno and Apples to Apples but they are not the best choices. I would like to know of a suitable card game. card-games party I've

==========

Explain the difference between a personal computer and a server

> . How do you explain the difference between a personal computer and a server? A personal computer is a computer that is used by a single user, whereas a server is a computer that is used to provide services to multiple users. How do you explain the difference between a personal computer and a server? A personal

==========