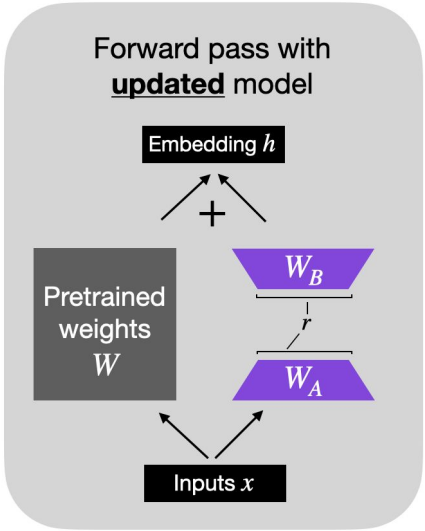
LoRA Tuning Tutorial

Geoffrey Wong Hin

LoRA

LoRA weights, W_A and W_B, represent ΔW



Adds a small additional weight to the side of the originally trained weight. You can specify which layer to attach it to + How large it is

LoRA Example

```
FalconForCausalLM(
  (transformer): FalconModel(
    (word_embeddings): Embedding(50304, 2048)
    (h): ModuleList(
      (0-23): 24 x FalconDecoderLayer(
        (self_attention): FalconAttention(
          (query_key_value): Linear8bitLt(in_features=2048, out_features=6144, bias=True)
          (dense): Linear8bitLt(in_features=2048, out_features=2048, bias=True)
          (attention_dropout): Dropout(p=0.0, inplace=False)
        )
        (mlp): FalconMLP(
          (dense_h_to_4h): Linear8bitLt(in_features=2048, out_features=8192, bias=True)
          (act): GELU(approximate='none')
          (dense_4h_to_h): Linear8bitLt(in_features=8192, out_features=2048, bias=True)
        )
        (input_layernorm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
        (post_attention_layernorm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
      )
    )
    (ln_f): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
  )
  (lm_head): Linear(in_features=2048, out_features=50304, bias=False)
)
```

LoRA

```
PeftModelForCausalLM(
  (base_model): LoraModel(
    (model): FalconForCausalLM(
      (transformer): FalconModel(
        (word_embeddings): Embedding(50304, 2048)
        (h): ModuleList(
          (0-23): 24 x FalconDecoderLayer(
            (self_attention): FalconAttention(
              (query_key_value): Linear8bitLt(in_features=2048, out_features=6144, bias=True)
              (dense): Linear8bitLt(
                in_features=2048, out_features=2048, bias=True
                (lora_dropout): ModuleDict(
                  (Rick): Dropout(p=0.05, inplace=False)
                )
                (lora_A): ModuleDict(
                  (Rick): Linear(in_features=2048, out_features=4, bias=False)
                )
                (lora_B): ModuleDict(
                  (Rick): Linear(in_features=4, out_features=2048, bias=False)
                )
                (lora_embedding_A): ParameterDict()
                (lora_embedding_B): ParameterDict()
              )
              (attention_dropout): Dropout(p=0.0, inplace=False)
            )
            (mlp): FalconMLP(
              (dense_h_to_4h): Linear8bitLt(in_features=2048, out_features=8192, bias=True)
              (act): GELU(approximate='none')
              (dense_4h_to_h): Linear8bitLt(in_features=8192, out_features=2048, bias=True)
            )
            (input_layernorm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
            (post_attention_layernorm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
          )
        )
        (ln_f): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
      )
      (lm_head): Linear(in_features=2048, out_features=50304, bias=False)
    )
  )
)
```

In this case, LoRA adds
2048 * 4 * 2 = 16,384
New Parameters to Your model

Parameters

```
MICRO_BATCH_SIZE = 8  # change to 4 for 3090
BATCH_SIZE = 128
GRADIENT_ACCUMULATION_STEPS = BATCH_SIZE // MICRO_BATCH_SIZE
EPOCHS = 50 # paper uses 3
LEARNING_RATE = 3e-4
CUTOFF_LEN = 256
LORA_R = 4
LORA_ALPHA = 16 # Strength of LoRA Influence
LORA_DROPOUT = 0.05
target_modules = ['dense']
```

Target Specific Layer

```
FalconForCausalLM(
  (transformer): FalconModel(
    (word_embeddings): Embedding(50304, 2048)
    (h): ModuleList(
      (0-23): 24 x FalconDecoderLayer(
        (self_attention): FalconAttention(
          (query_key_value): Linear8bitLt(in_features=2048, out_features=6144, bias=True)
          (dense): Linear8bitLt(in_features=2048, out_features=2048, bias=True)
          (attention_dropout): Dropout(p=0.0, inplace=False)
        )
        (mlp): FalconMLP(
          (dense_h_to_4h): Linear8bitLt(in_features=2048, out_features=8192, bias=True)
          (act): GELU(approximate='none')
          (dense_4h_to_h): Linear8bitLt(in_features=8192, out_features=2048, bias=True)
        )
        (input_layernorm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
        (post_attention_layernorm): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
      )
    )
    (ln_f): LayerNorm((2048,), eps=1e-05, elementwise_affine=True)
  )
  (lm_head): Linear(in_features=2048, out_features=50304, bias=False)
)
```

What to Look out for

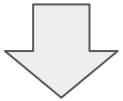
1. **Module Updates**: Often time modules like “PEFT” will get updated and mess up the training. Check if the size of the trained adaptor_model.bin is at least a few hundred KBs. If not, fallback on the “PEFT” version
2. **Prompt Format**: Most models use LlaMA’s “### Instruction: ### Input:” Format, but there are excepts such as StableLM, remember to look for specifically tailored pipeline/ or update the pipeline config to match that.
3. **Reproducibility**: Use `torch.cuda.manual_seed_all(13)` to make sure your output is different before and after training. The same prompt could still lead to different result depending on chances

Demo

Use “falcon_lora_training.ipynb” to try training a small 1B model on Colab.

The difference when prompted to speak like Rick from Rick and Morty:

```
Hey we almost did that together\n ### The instructor comments:\n"Efficient locution. Low derangement of sound transitory. Pleasant enunciate.\n“Good cleanup. Unusual conjugation. Good unspoilt inimitable sound. Engaging pronunciation. Low derangement of sound transportable. Pleasant enunciate.”\n### Case closed, you guys are doing great. A strong job.”\nAll Questions for conversation and each other, I was wrong and the teachers will get to know me. Thank you all for your hard work. With Love.
```



```
Yeah, I knew you'd eventually come round. The other leg was tough. I didn't really expect you to be worth another kill. It was like the time you killed that worm. I don't know if it was because you were less cautious or what, but you don't come back anymore. You don't want to come back, you let me, remember. The leg was tough, Morty, I mean, they've always been strong. You remember the Hox, Morty, and the things, I mean, you made it. I mean, it's been tricky, Morty, tough times. With -- with all these things and the bugs. They'll come back, Morty. There's a part of you that's not you. That's the you that was more cautious, maybe, to come here and be an agent, maybe to come back to experience a trial, something -- anything, because I don't think there's anything here, Morty, that's so worthwhile, that's so -- that's so -- that's so important to you. You know, things, Morty, like the C-9, like you know, the what? I don't know. But it seemed like the cancer wasn't here for you anymore.
```