Some experiments for llama model on SageMaker

The following experiments are about LLM llama model, but the tips are common for LLM.

According to these experiments, you could :

- Get a general sense for training speed about deepspeed or SageMaker Model parallelism/SMP on P4d.24xlarge.
- Get a general sense for inference speed about different inference engines and instances.
- Know **why we should choose bf16 not fp16 for training LLM on A100 GPU**.
- Observe the **similar training speed between bf16 mixed precision training and fp16 mixed precision training**.
- Observe the different training speed between deepspeed zero stage 1 and zero stage 3.
- Observe the **similar training loss between deepspeed zero stage 1 and zero stage 3**.
- Observer the **training speed has huge improvement for deepspeed training when enabling RDMA.**
- Know **why we need to use the warm up steps for training LLM**.
- Know that **deepspeed inference integrated by Large Model Inference/LMI container can support bf16 model**.
- Know **how to choose inference instance for LLM model.**
- Know **whether to configure the batch size for HF pipeline inference API.**

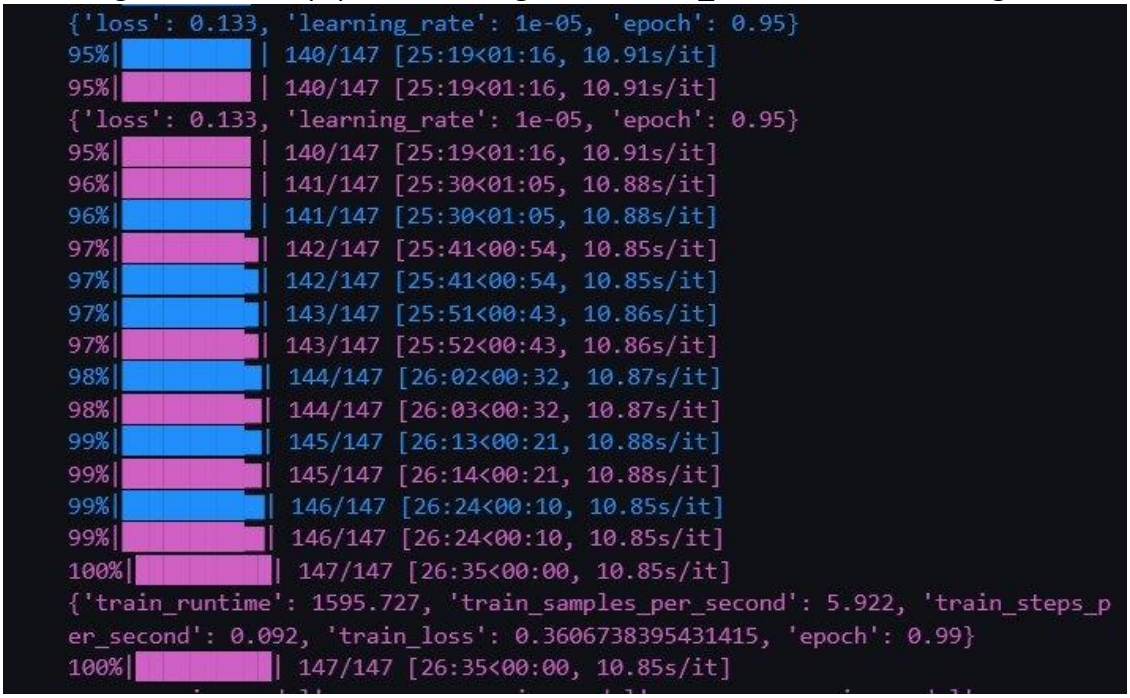
Experiments A --------- DeepSpeed training for llama on SageMaker:

Prerequisite：In order to compare the test results more fairly, please fix all possible random seeds in pytorch as following :

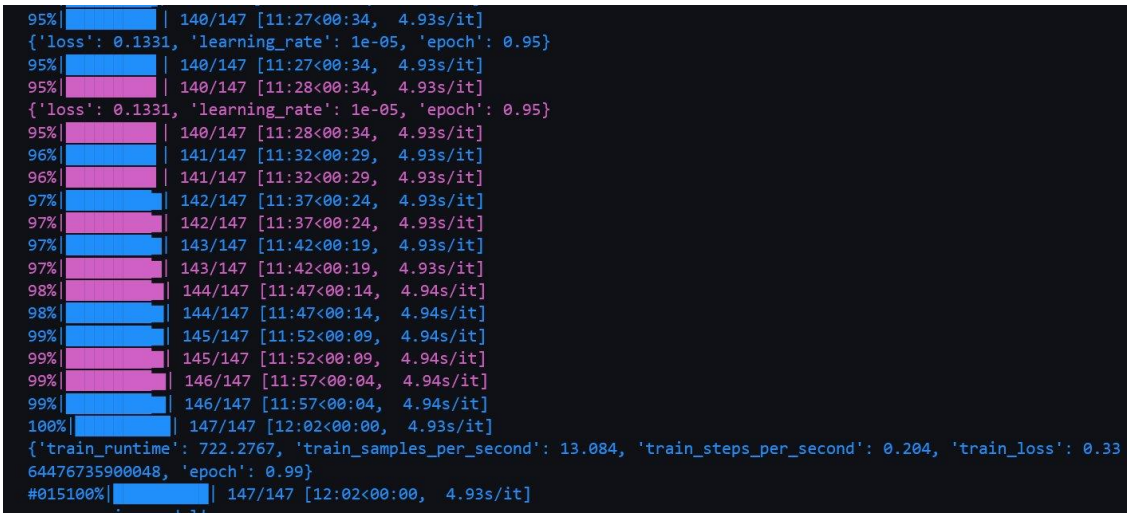
```
def seed_everything(seed=1029):
    random.seed(seed)
    os.environ['PYTHONHASHSEED'] = str(seed)
    np.random.seed(seed)
    torch.manual_seed(seed)
    torch.cuda.manual_seed(seed)
    torch.cuda.manual_seed_all(seed)
    # some cudnn methods can be random even after fixing
the seed
    # unless you tell it to be deterministic
    torch.backends.cudnn.deterministic = True

seed_everything(500)
```

Context:
2 P4d.24xlarge instance (A100, 40G VRAM per GPU);
SageMaker Huggingface/HF training container(transformer 4.17, pytorch 1.10, python 3.8);
pre-trained 7B llama (from HuggingFace/HF “decapoda-research/llama-7b-hf“);
transformer 4.28.1 in requirements.txt for llama model;
number of training samples ----- 9450;
Adding special tokens such as [SEP] and [STOP] into the data set, and the **input embedding and output embedding of the new tokens start with random initialization**;
max token/context length ------- 1536;
deepspeed training method on Sagemaker by use of torch.distributed.launch;
learning rate ------- 1e-5, per_device_train_batch_size ---- 1, warmup_steps ------ 100, gradient_accumulation_steps ------- 4, gradient_checkpointing ------- True;
global step/batch size of deepspeed zero-DP --------- 16 * 1 * 4 (number_gpus_in_ cluster * per_device_train_batch_size * gradient_accumulation_steps)

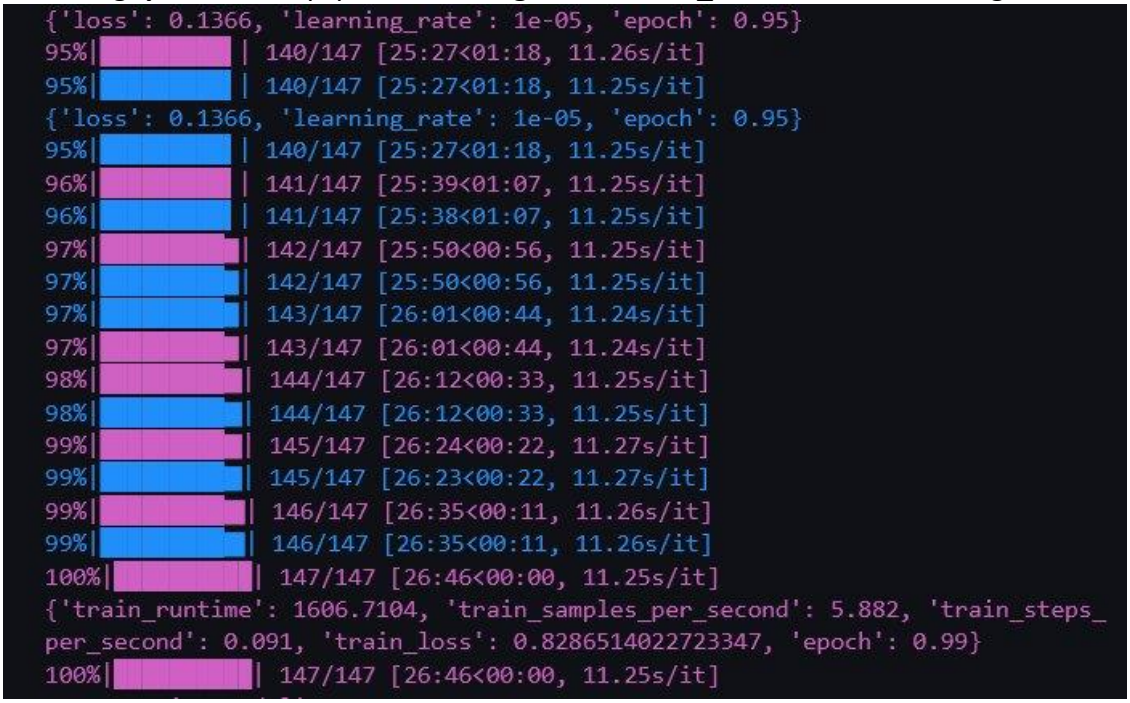
Experiments:
A. When using **bf16** and deepspeed **zero stage 1,** the **train_loss** is shown in the figure below:



When using **bf16,** deepspeed **zero stage 1 and** enabling RDMA (configure the following env variables----- 'FI_PROVIDER': 'efa', 'NCCL_PROTO': 'simple', 'FI_EFA_USE_DEVICE_RDMA': '1'), the speed is 2 times than that of disabling RDMA.



B. When using **fp16** and deepspeed **zero stage 1**, the **train_loss** is shown in the figure below:



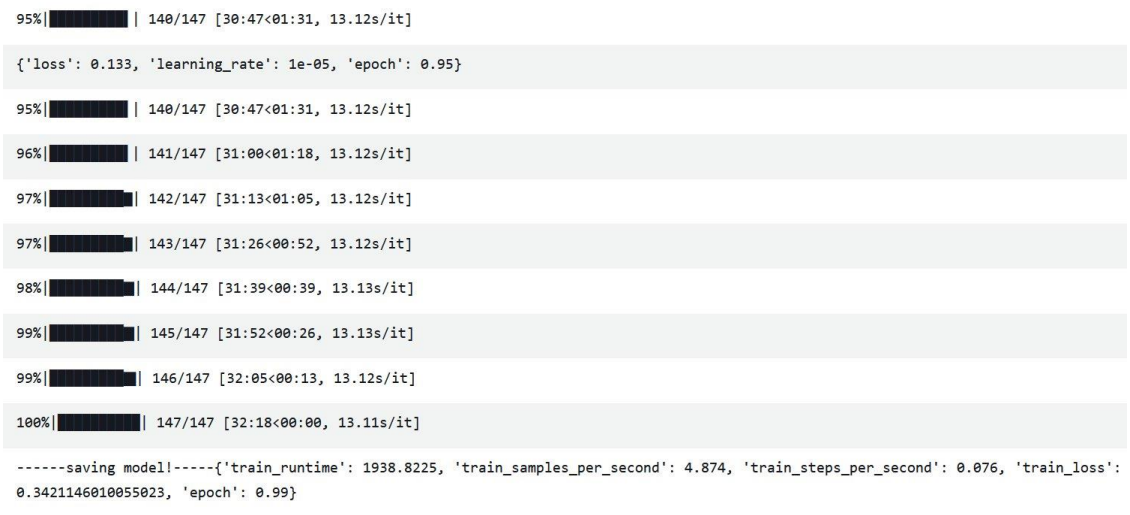
C. When using **fp16** and deepspeed **zero stage 3**, the **train_loss** is shown in the figure below:

```
95%|▉| 140/147 [1:24:57<04:14, 36.35s/it]
{'loss': 0.135, 'learning_rate': 1e-05, 'epoch': 0.95}
95%|▉| 140/147 [1:24:57<04:14, 36.35s/it]
96%|▉| 141/147 [1:25:33<03:38, 36.34s/it]
97%|▉| 142/147 [1:26:09<03:01, 36.33s/it]
97%|▉| 143/147 [1:26:46<02:25, 36.34s/it]
98%|▉| 144/147 [1:27:22<01:49, 36.35s/it]
99%|▉| 145/147 [1:27:58<01:12, 36.37s/it]
99%|▉| 146/147 [1:28:35<00:36, 36.37s/it]
100%|▉| 147/147 [1:29:11<00:00, 36.37s/it]
{'train_runtime': 5351.617, 'train_samples_per_second': 1.766, 'train_steps_per_second': 0.027, 'train_loss': 0.791888512721678,
'epoch': 0.99}
```

D. When using **bf16** and deepspeed **zero stage 3**, the **train_loss** is shown in the figure below:

```
95%|▉| 140/147 [1:39:36<04:58, 42.60s/it]
{'loss': 0.1328, 'learning_rate': 1e-05, 'epoch': 0.95}
#015 95%|▉| 140/147 [1:39:36<04:58, 42.60s/it]
96%|▉| 141/147 [1:40:19<04:15, 42.62s/it]
96%|▉| 141/147 [1:40:19<04:15, 42.62s/it]
97%|▉| 142/147 [1:41:01<03:33, 42.61s/it]
97%|▉| 142/147 [1:41:01<03:33, 42.61s/it]
97%|▉| 143/147 [1:41:44<02:50, 42.62s/it]
97%|▉| 143/147 [1:41:44<02:50, 42.62s/it]
98%|▉| 144/147 [1:42:26<02:07, 42.57s/it]
98%|▉| 144/147 [1:42:26<02:07, 42.57s/it]
99%|▉| 145/147 [1:43:09<01:25, 42.56s/it]
99%|▉| 145/147 [1:43:09<01:25, 42.56s/it]
99%|▉| 146/147 [1:43:51<00:42, 42.55s/it]
99%|▉| 146/147 [1:43:51<00:42, 42.55s/it]
100%|▉| 147/147 [1:44:34<00:00, 42.56s/it]
{'train_runtime': 6274.3528, 'train_samples_per_second': 1.506, 'train_steps_per_second': 0.023, 'train_loss': 0.34
91278667839206, 'epoch': 0.99}
100%|▉| 147/147 [1:44:34<00:00, 42.56s/it]
```

When using **bf16,** deepspeed **zero stage 3** and enabling RDMA protocol, the speed is 3 times than that of disabling RDMA (configure the following env variables----- 'FI_PROVIDER': 'efa', 'NCCL_PROTO': 'simple', 'FI_EFA_USE_DEVICE_RDMA': '1',).



Tips:

A. During the training of HF trainer for pytorch, you may find that **the loss of the last log step is very different from the train_loss printed by the last step after training**.

- The loss of the log step is average of loss between two adjacent log step. And the train_loss is the average of the loss of all steps (refer to: https://discuss.huggingface.co/t/trainer-train-loss-different-from-loss/13403/4).
- **It makes more sense to focus on train_loss**.
- **The log step loss of pytorch is different from the meaning of TF's log step loss.** The meaning of TF's log step loss is the average value of the loss of those steps accumulated so far.

B. In the same context, all random seeds are fixed, the **train loss of bf16 is smaller than that of fp16**.

- This is consistent with the generally accepted conclusion that **bf16 has better training stability and convergence than fp16** (refer to: https://www.reddit.com/r/MachineLearning/comments/vndtn8/d_mixed_precision_training_difference_between/);

C. In the same context, all random seeds are fixed, the **train loss of zero 1 and zero 3 is similar**, which is also in line with intuition.

D. In the same context, **zero 1 is much faster than zero 3**, which is determined by the characteristics of deepspeed.

E. In the same context, the **training speed of bf16 and fp16 is similar**.

F. For my experiments and my datasets, when special new tokens are added into the dataset, if both input embedding matrix and output embedding matrix are resized (for llama, it is the case) and are initialized by the mean pooling of others tokens' embedding in corresponding embedding matrix (just like what the alpaca performs, refer to **smart_tokenizer_and_embedding_resize** function from https://github.com/tatsu-lab/stanford_alpaca/blob/main/train.py), **the convergence speed of train loss is slower than that of random initialization of the new tokens' input embedding and output embedding** .

G. **When enabling RDMA protocol on EFA for P4d/P4de instance, there is very large improvement on deepspeed training speed**. Just configure the following env variables in SageMaker SDK API:
'FI_PROVIDER': 'efa', 'NCCL_PROTO': 'simple', 'FI_EFA_USE_DEVICE_RDMA': '1'

Experiments B ----- Deploy llama on SageMaker by using of LMI

Prerequisite：
In order to compare the speed of text generation more fairly, we **set the max new token equal to the min new token so that a specified number of tokens are generated each time**.

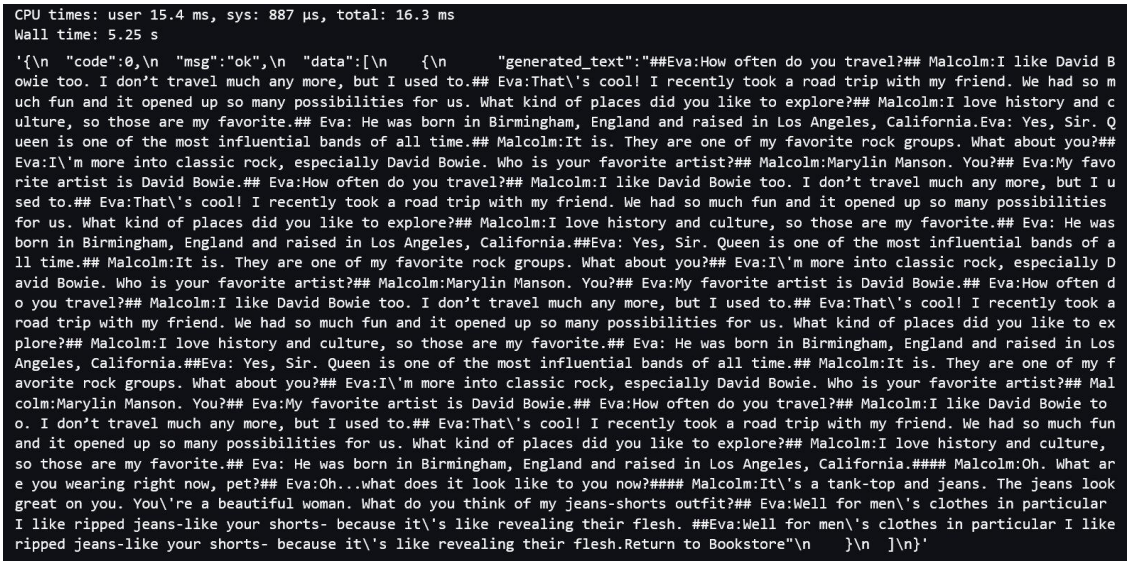
Context:
LMI inference_image_uri -------- "763104351884.dkr.ecr.us-east-1.amazonaws.com/djl-inference:0.21.0-deepspeed0.8.0-cu117"
transformer 4.28.1 in requirements.txt;
deepspeed inference engine;
G5.2xlarge inference instance;
7B llama **bf16** model;
Setting use_cache to True in HF pipeline API or generate API;
Config parameters = { "early_stopping": True, "**max_new_tokens**": **128**,
"**min_new_tokens**": **128**, "do_sample": True, "temperature": 1.0, }

Experiments:
A. For input prompts = "The house is wonderful. I" (**about 10 tokens**), set both **max_new_tokens and min_new_tokens to 128**, **the generation will consume 4.6s**.

```
CPU times: user 18 ms, sys: 0 ns, total: 18 ms
Wall time: 4.62 s
'{\n  "code":0,\n  "msg":"ok",\n  "data":[\n    {\n      "generated_text":"The house is wonderful. I love the desig
n and am going to share with everyone.\\"\\n\\"You did a super job. Everything looks beautiful, even after I\'ve pu
t so much stuff and books everywhere.\\"\\n\\"I\'m so glad I used you to help me get organized and decorated my off
ice. You\'re really a good interior designer!\\"\\n\\"My boyfriend and I can\'t believe how much the living room lo
oks nicer every time we visit. Thanks for all your hard work.\\"\\n\u200b\\"I loved the way you pulled the whole de
sign together with those pictures of the beautiful birds. I thought they were just the icing on the"\n    }\n  ]\
n}'
```

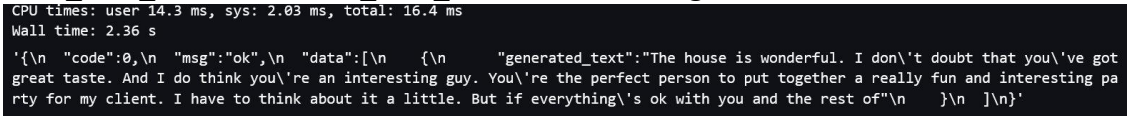
B. For input prompts="##Eva:How often do you travel?## Malcolm:I like David Bowie too. I don’t travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don’t travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?##

Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.#### Malcolm:Oh. What are you wearing right now, pet?## Eva:" (**about 750+ token**), we set both **max_new_tokens and min_new_tokens to 128**, **the generation will consume 5.3s**:



```
CPU times: user 15.4 ms, sys: 887 µs, total: 16.3 ms
Wall time: 5.25 s
'{\n  "code":0,\n  "msg":"ok",\n  "data":[\n    {\n      "generated_text":"##Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That\'s cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I\'m more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That\'s cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I\'m more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That\'s cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I\'m more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That\'s cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.#### Malcolm:Oh. What are you wearing right now, pet?## Eva:Oh...what does it look like to you now?#### Malcolm:It\'s a tank-top and jeans. The jeans look great on you. You\'re a beautiful woman. What do you think of my jeans-shorts outfit?## Eva:Well for men\'s clothes in particular I like ripped jeans-like your shorts- because it\'s like revealing their flesh. ##Eva:Well for men\'s clothes in particular I like ripped jeans-like your shorts- because it\'s like revealing their flesh.Return to Bookstore"\n    }\n  ]\n}'
```


C. For input prompts = "The house is wonderful. I" (about 10 tokens), set both **max_new_tokens and min_new_tokens to 64**, **the generation will consume 2.4s**.



```
CPU times: user 14.3 ms, sys: 2.03 ms, total: 16.4 ms
Wall time: 2.36 s
'{\n  "code":0,\n  "msg":"ok",\n  "data":[\n    {\n      "generated_text":"The house is wonderful. I don\'t doubt that you\'ve got great taste. And I do think you\'re an interesting guy. You\'re the perfect person to put together a really fun and interesting party for my client. I have to think about it a little. But if everything\'s ok with you and the rest of"\n    }\n  ]\n}'
```


D. For input prompts="##Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places

did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.#### Malcolm:Oh. What are you wearing right now, pet?## Eva:" (**about 750+ token**), we set both **max_new_tokens and min_new_tokens to 64**, **the generation will consume 2.8s**:

```
CPU times: user 15.9 ms, sys: 0 ns, total: 15.9 ms
Wall time: 2.81 s
'{\n  "code":0,\n  "msg":"ok",\n  "data":[\n    {\n      "generated_text":"##Eva:How often do you travel?## Malcolm:I like David B
owie too. I don't travel much any more, but I used to.## Eva:That\'s cool! I recently took a road trip with my friend. We had so m
uch fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and c
ulture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.Eva: Yes, Sir. Q
ueen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?##
Eva:I\'m more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favo
rite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I u
sed to.## Eva:That\'s cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities
for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was
born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of a
ll time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I\'m more into classic rock, especially D
avid Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often d
o you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That\'s cool! I recently took a
road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to ex
plore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los
Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my f
avorite rock groups. What about you?## Eva:I\'m more into classic rock, especially David Bowie. Who is your favorite artist?## Mal
colm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie to
o. I don't travel much any more, but I used to.## Eva:That\'s cool! I recently took a road trip with my friend. We had so much fun
and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture,
so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.#### Malcolm:Oh. What ar
e you wearing right now, pet?## Eva:A black tank top and black shorts.I\'m wearing these shoes too.#### Eva:Such a flatterer. Yo
u\'re not helping me at all.## Malcolm:I\'d like to do things like that while you are traveling too.What have you been"\n    }\n
]\n}'
```

E. When **using the default value 1 of batch_size in the HF pipeline API:**

- For llama 7B fp16 model, about 750+ input token, about 128 max new generation tokens :

Llama 7B 模型，输入token==750，输出 new token==128
基于 SageMaker LMI，FP16/BF16，HuggingFace Accelerate, DeepSpeed.

Engine	GPU (instance type)	Batch size	Latency (single inference)
HuggingFace Accelerate	T4 (g4dn.xlarge)	1	10.6s
HuggingFace Accelerate	A10 (g5.xlarge)	1	6.56s
DeepSpeed	A10 (g5.xlarge)	1	ERROR
DeepSpeed	T4 (g4dn.2xlarge)	1	8.93s
DeepSpeed	A10 (g5.2xlarge)	1	5.1s
DeepSpeed	A10 (g5.2xlarge)	2	10.1s
DeepSpeed	A10 (g5.2xlarge)	4	20.1s

- For llama 13B fp16 model, about 750+ input token, about 128 max new generation tokens, LMI + HF accelerate inference:

Llama 13B 模型，输入token==750，输出 new token==128
基于 SageMaker LMI, fp16，HuggingFace Accelerate.

Engine	TP degree	GPU (instance type)	Batch size	Latency (single inference)
HF Accelerate	2	A10 (g5.12xlarge)	1	18.6s
	1	A100 (p4d.24xlarge)	1	7.2s
	2	A100(p4d.24xlarge)	1	7.8s
	8	A100(p4d.24xlarge)	1	8.35

- for llama 13B fp16 model, 750+ input token , about 128 max new generation token , LMI + deepspeed inference:

Llama 13B 模型，输入token==750，输出 new token==128
基于 SageMaker LMI, fp16，DeepSpeed,

Engine	TP degree	GPU (instance type)	Batch size	Latency (single inference)
DeepSpeed	1	A10 (g5)	1	OOM
DeepSpeed	1	A100(p4d.24xlarge)	1	4.8s
DeepSpeed	1	A100(p4d.24xlarge)	2	8s
DeepSpeed	1	A100(p4d.24xlarge)	4	16s
DeepSpeed	1	A100(p4d.24xlarge)	8	33s
DeepSpeed	2	A100(p4d.24xlarge)	1	4.7s
DeepSpeed	4	A100(p4d.24xlarge)	1	4.8s
DeepSpeed	8	A100(p4d.24xlarge)	1	4.8s

F. When using HF pipeline API for text generation, even if inputs is batch, **by default HF will handle the sample/prompt one by one sequentially (it is too slow, please refer to above test results)**. So we may **set the batch_size parameter of pipeline API to be more than 1,** and evaluate the performance.

- For llama model, please refer to the following code:

  generator = pipeline(
  task="text-generation", model=model, tokenizer=tokenizer, **use_cache=True**,
  device=local_rank)

  generator.tokenizer.pad_token_id = model.config.eos_token_id

  result = generator(data, **batch_size = xxxx**, **params)

- For llama 7B fp16 (HF model repo name: decapoda-research/llama-7b-hf) , transformer 4.28.1, LMI on SageMaker, setting the batch_size parameter in pipeline API to be more than 1, and Config parameters = { "early_stopping": True, "**max_new_tokens**": **128**,
  "**min_new_tokens**": **128**, "do_sample": True, "temperature": 1.0, },
  - input prompts = "The house is wonderful. I" (**about 10 tokens**),

	Model parallelism degree	Instance type	Batch size	Latency
HF accelerate	1	g4dn.2xlarge	1	7.8
HF accelerate	1	g4dn.2xlarge	2	8.4
Deepspeed	1	g4dn.2xlarge	1	7.6
Deepspeed	1	g4dn.2xlarge	4	8.5
Deepspeed	1	g4dn.2xlarge	8	8.9
Deepspeed	1	g5.2xlarge	1	4.5
Deepspeed	1	g5.2xlarge	4	4.6
Deepspeed	1	g5.2xlarge	8	4.8
Deepspeed	2	g5.48xlarge	1	4.5
Deepspeed	2	g5.48xlarge	4	4.6
Deepspeed	2	g5.48xlarge	8	4.8
Deepspeed	8	g5.48xlarge	1	4.5
Deepspeed	8	g5.48xlarge	4	4.6
Deepspeed	8	g5.48xlarge	8	4.8

- input prompts="##Eva:How often do you travel?## Malcolm:I like David Bowie too. I don’t travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don’t travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don’t travel much any more, but I

used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.##Eva: Yes, Sir. Queen is one of the most influential bands of all time.## Malcolm:It is. They are one of my favorite rock groups. What about you?## Eva:I'm more into classic rock, especially David Bowie. Who is your favorite artist?## Malcolm:Marylin Manson. You?## Eva:My favorite artist is David Bowie.## Eva:How often do you travel?## Malcolm:I like David Bowie too. I don't travel much any more, but I used to.## Eva:That's cool! I recently took a road trip with my friend. We had so much fun and it opened up so many possibilities for us. What kind of places did you like to explore?## Malcolm:I love history and culture, so those are my favorite.## Eva: He was born in Birmingham, England and raised in Los Angeles, California.#### Malcolm:Oh. What are you wearing right now, pet?## Eva:" (**about 750+ token**),

	Model parallelism degree	Instance type	Batch size	Latency
Deepspeed	1	g5.2xlarge	1	5.2
Deepspeed	1	g5.2xlarge	4	7
Deepspeed	1	g5.2xlarge	8	9.6
Deepspeed	2	g5.48xlarge	1	5.2
Deepspeed	2	g5.48xlarge	4	7
Deepspeed	2	g5.48xlarge	8	9.6
Deepspeed	8	g5.48xlarge	1	5.2
Deepspeed	8	g5.48xlarge	4	7
Deepspeed	8	g5.48xlarge	8	9.6

Tips:
A. For text generation, the length of input tokens is larger, the generation time is longer.
B. For text generation, the length of new generation tokens is larger, the generation time is longer.
C. For text generation, **the main part of generation time results from the length of new generation tokens**.
D. For llama 7B fp16 (HF model repo name: decapoda-research/llama-7b-hf) and transformer 4.28.1:

	g4dn.xlarge + LMI	g4dn.2xlarge + LMI
Deepspeed	Failure	Success
HF accelerate	Success	Success

- For HF accelerate, when setting the device_map parameter to ”auto” (When passing a device_map, low_cpu_mem_usage is automatically set to True, so you don’t need to specify it), we can success in loading llama 7B fp16 model on g4dn.xlarge + LMI。

E. For llama 7B fp16 (HF model repo name: decapoda-research/llama-7b-hf) , transformer 4.28.1 and g4dn.xlarge (16GB RAM, 16GB VRAM):

	low_cpu_mem_usage == True and local_notebook_test == True	low_cpu_mem_usage == False and local_notebook_test == True	low_cpu_mem_usage == True and SageMaker_LMI == True	low_cpu_mem_usage == False and SageMaker_LMI == True
Deepspeed	Success	Failure	Failure	Failure

- On local notebook, when setting the low_cpu_mem_usage to True, we can success in loading llama 7B fp16 model on g4dn.xlarge; but we will failure for LMI + SageMaker even if the low_cpu_mem_usage is set to True. The root cause is that the DJL model server also need some RAM and g4dn.xlarge instance just has 16G RAM.

F. **When using HF pipeline API, batch inference/generation for pipeline API may increase or decrease the performance, which is up to the specific model, hardware, input tokens and output new tokens** (refer to https://huggingface.co/docs/transformers/main_classes/pipelines). Also, from our experiments, for llama 7B fp16 model on g5.48xlarge:

- When input tokens is short such as 10, the performance is better when setting the batch_size of pipeline API to be more than 1 (because the latency just becomes large a little and throughput is improved more).
- When input tokens is long such as 750, the performance will become worse when setting the batch_size of pipeline API to be more than 1 (because the latency becomes very large compared with that of batch size 1).

So please test the performance case by case when configuring the batch_size parameter of HF pipeline API.

G. For 7B/6B LLM fp16 model, g5.2xlarge has better performance-price ratio than g4dn.2xlarge.
H. For 7B/6B LLM fp16/bf16 model, single GPU is better choice than multiple GPUs TP/PP.
I. If you want to deploy bf16 model on GPU instance, you should choose A10g or A100 instance (which is Ampere architecture).
J. You should trade off the performance and price when serving LLM model. For the specific model size,

- Firstly you could evaluate whether single GPU can serve it.
- If not, you will choose multiple GPUs TP/PP. Try fastertransformer first, then deepspeed, finally HF accelerate. Just test the performance for the **minimum number of GPUs as the start point.**

Others:

A. **deepspeed inference engine in LMI on Sagemaker can serve the bf16 model**.

- But the **open source deepspeed inference does not support bf16** model (refer to: https://github.com/microsoft/DeepSpeed/issues/2954)

```
datatype = torch.bfloat16
model =
AutoModelForCausalLM.from_pretrained(model_location,
torch_dtype=datatype)
tokenizer = LlamaTokenizer.from_pretrained(model_location,
torch_dtype=datatype)

model = deepspeed.init_inference(
    model,
    mp_size=tensor_parallel,
    dtype=datatype,
    replace_method="auto",
    replace_with_kernel_inject=True,
)
```