Quiz 5 - Spring 2023 - Extra credit

Karan Vora (username: kv2154)

Attempt 1

Written: May 11, 2023 11:07 PM - May 11, 2023 11:18 PM

Submission View

Released: Apr 26, 2023 11:57 PM

Question 1 **1 / 1 point**

In self-attention mechanism (select all that apply)

- ✓ ◯ Given a word, its neighboring words are used to compute its context by taking a weighted sum up the word values to map the Attention related to that given word.
- ◯ Given a word, its neighboring words are used to compute its context by taking a simple average of the word values to map the Attention related to that given word.
- ◯ Given a word, its neighboring words are used to compute its context by selecting the highest of the word values to map the Attention related to that given word.
- ◯ Given a word, its neighboring words are used to compute its context by selecting the lowest of those word values to map the Attention related to that given word.

Question 2 **0 / 1 point**

Select all that is true about multi-head mechanisms in Transformer models.

- ✓ ☐ Self-attention is used in only encoder
- ⇨ ✗ ☐ encoder-decoder attention is used only in decoder
- ⇨ ✓ ☐ Both self-attention and encoder-decoder attention are used in decoder
- ✗ ☐ Both self-attention and encoder-decoder attention are used in encoder

Question 3 **1 / 2 points**

Consider the four communication schemes for gradient aggregation in distributed training:

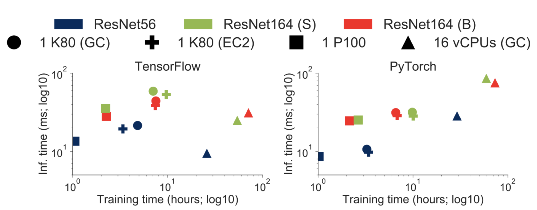
Ring AllReduce, OnetoAll, Butterfly AllReduce, Tree AllReduce
Fill the right scheme below in a and b.

a. 1. ______ is twice faster than 2. ______
b. Number of communication rounds in OnetoAll is same as in 3.______
c. Total communication time in OnetoAll is lesser than in Ring AllReduce. True or False. ______

Answer for blank # 1: Butterfly AllReduce ✓(25 %)
Answer for blank # 2: Ring AllReduce ✗ **(Tree AllReduce)**
Answer for blank # 3: Tree AllReduce ✗ **(Ring AllReduce)**
Answer for blank # 4: False ✓(25 %)

Question 4 **0.167 / 1 point**

The charts below capture TTA variability due to hardware and framework. It shows the inference and training time to 93% accuracy for different hardware, frameworks, and model architectures in DAWNBench seed entries.

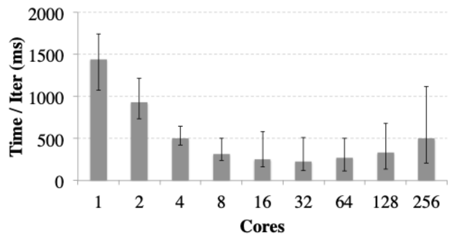


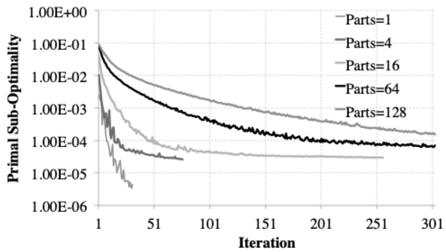
Based on this chart identify the best hardware and framework to train different architectures. For framework, your choice should be:
tensorflow or pytorch
and for hardware your choice should be one of the following:
1 K80(GC), 1 K80(EC2), 1 P100, 16 vCPUs(GC)
a. ResNet56 training:
framework: 1. __________ hardware: 2. __________
b. ResNet56 inference:
framework: 3. __________ hardware: 4. __________
c. ResNet164(S) training:
framework: 5. __________ hardware: 6. __________

Answer for blank # 1: tensorfow ✗ **(tensorflow, pytorch)**
Answer for blank # 2: P100 ✗ **(1 P100)**
Answer for blank # 3: pytorch ✓(16.67 %)
Answer for blank # 4: GC ✗ **(1 P100)**
Answer for blank # 5: pytorch ✗ **(tensorflow)**
Answer for blank # 6: EC2 ✗ **(1 P100)**

Question 5 **0 / 1 point**

Below are two charts showing the scalability of distributed optimization algorithms for machine learning. The first shows the time per iteration as the degree of parallelism (in terms of number of cores) is varied. The second chart shows the convergence of CoCoA (a distributed optimization algorithm) as we vary the degree of parallelism (in terms of number of parts; parts is same as cores).



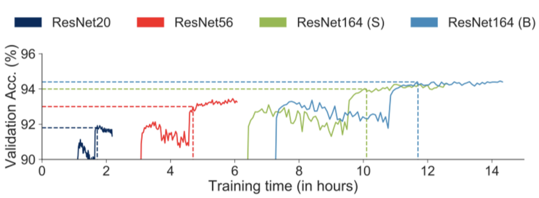


From these charts and other observations in the Hemingway paper select all that is correct.

- The time per iteration decreases initially as we increase the degree of parallelism due to reduced work per core.
- Different distributed algorithms show different degree of scaling with increased parallelism and so the right choice of algorithm for any task can be determined only by knowing the number of cores available in the cluster.
- The observations in Hemingway paper are only valid for core level parallelism and does not apply to distributed deep learning training using GPUs as in the Facebook paper (we studied in distributed training) it was observed that by increasing the degree of parallelism the time per iteration remained almost constant.
- Convergence of CoCoA is poor as we increase the degree of parallelism and hence training using less number of cores will take less time to converge compared to training using large number of cores.

Question 6 **0.25 / 1 point**

You are given 4 networks ResNet20, ResNet56, ResNet156 (S), ResNet156 (B). When training with CIFAR10 the following figure shows the convergence of these 4 networks. When optimizing for TTA which network will be preferred for following accuracy thresholds:



90% 1. ______________

92.5% 2. ______________

93.2% 3. ______________

94.1% 4. ______________

Answer for blank # 1: ResNet156 (S) ✗ **(ResNet20)**

Answer for blank # 2: ResNet20 ✗ **(ResNet56)**

Answer for blank # 3: ResNet56 ✓(25 %)

Answer for blank # 4: ResNert156 (B) ✗ **(ResNet156 (S))**

Question 7 **0 / 2 points**

Consider stochastic depth with linear survival probability of layer ℓ defined as:

$$p_\ell = 1 - \frac{\ell}{L}(1 - p_L)$$

What is the effective depth of the network during training time when:

1. p_L = 0.75 and L=23 ______________

2. p_L = 0.4 and L=29 ________________

Answer for blank # 1: 17.25 ✖ **(20)**

Answer for blank # 2: 11.6 ✖ **(20)**

Question 8 **0.6 / 1 point**

Select all that is true about Pytorch DataParallel?

- ✔ ☐ Using DataParallel you can also perform multi-GPU training over multiple machines however it will be much slower than DistributedDataParallel
- ⇨ ✔ ☐ If you want to do minimal code changes to run your training over multiple GPUs on a single machine then DataParallel is preferable
- ⇨ ✔ ☐ At the start of every forward path an updated model is replicated from one GPU to all the other GPUs
- ✖ ☐ DataParallel benefits from overlapping gradient computation with gradient reduction thus speeding up the training by reducing the communication overhead
- ⇨ ✖ ☐ One of the GPUs ends up doing majority of the processing while other GPUs are underutilized

Question 9 **0.5 / 1 point**

Which of the following are true about Transformers?

- ✔ ☐ Transformers consist of a single encoder and decoder network.
- ⇨ ✔ ☐ Unlike RNNs, transformers process the entire input at once.
- ⇨ ✔ ☐ Residuals are applied to attention layers to combat the vanishing gradients problem.
- ✖ ☐ Weight sharing occurs between different encoders in a transformer network.

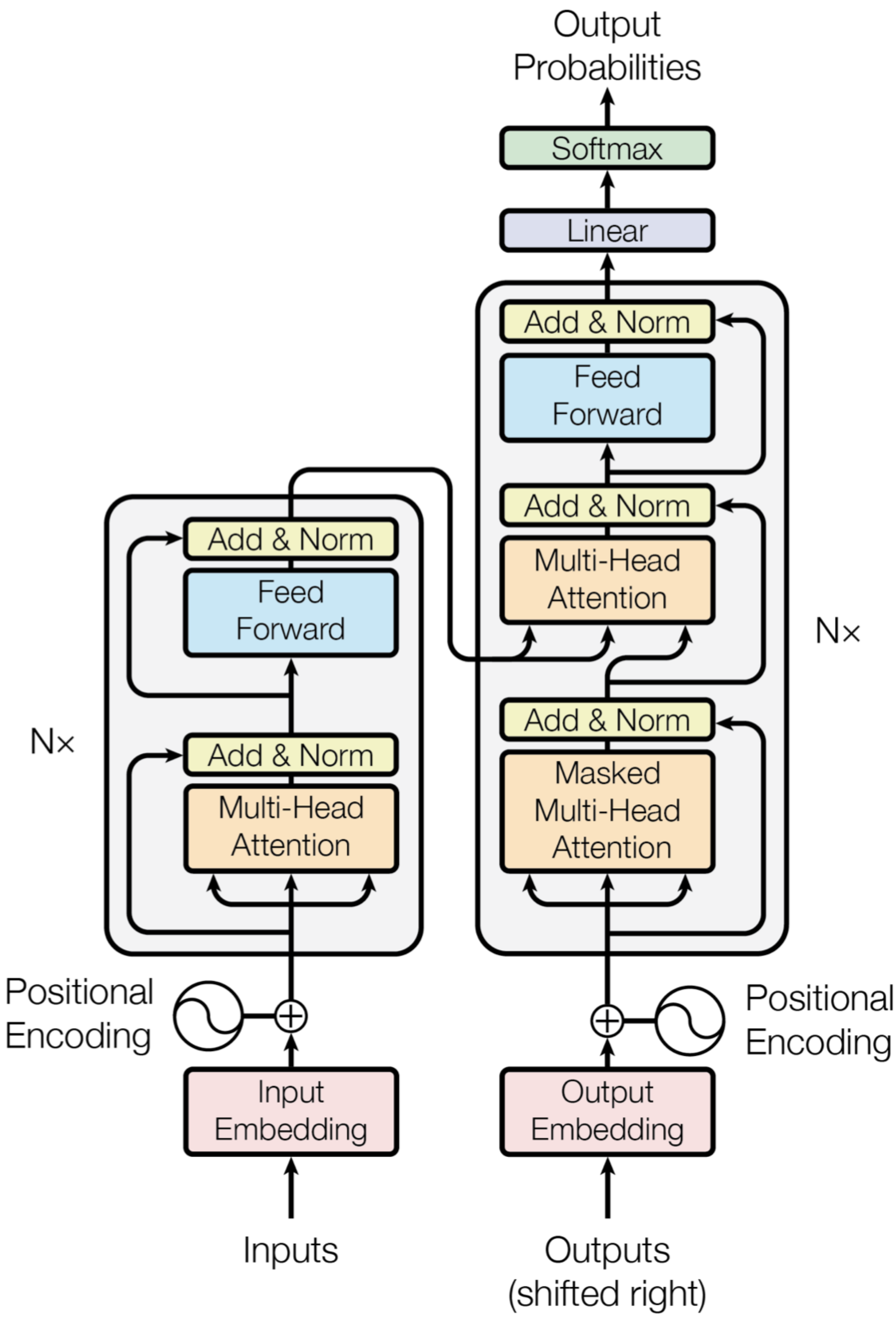
Question 10 **2 / 2 points**

Stochastic depth is a regularization technique proposed for Residual networks which "enables the seemingly contradictory setup to train short networks and use deep networks at test time". During training in stochastic depth, for each mini-batch, a subset of layers is dropped and bypassed with the identity function. Select all that is true for stochastic depth regularization technique. In the following constant depth refers to regular training without stochastic depth and survival probability is the probability of not dropping a layer.

- ✔ ☐ Deeper networks show significantly large improvement in performance with stochastic depth compared to shallow networks.
- ✔ ☐ If the survival probability is not chosen correctly stochastic depth can result in higher test error compared to constant depth.
- ✔ ☐ Stochastic depth can save training time substantially without compromising accuracy.
- ✔ ☐ With stochastic depth one can train very deep networks effectively as the magnitude of gradients in stochastic depth is mostly larger than their values in constant depth network.
- ✔ ☐ Stochastic depth always improves performance compared to constant depth for network of any number of hidden layers.

Question 11 **0 / 1 point**

Following is the architecture of a transformer network

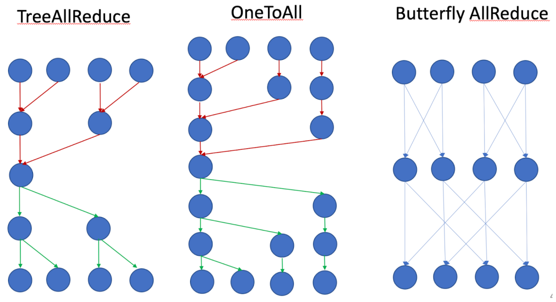


What information does the decoder take from the encoder for its second block of multi-head mechanism?

- ☐ Value
- ☐ Key
- ☐ input sequence
- ☐ Query

Question 12 **1 / 1 point**

Below are the three communication schemes:



Which of these has the same communication cost (in terms of the number of communication rounds) as Ring All-Reduce with scatter-gather?

- ○ TreeAllReduce
- ✓ ○ OneToAll
- ○ Butterfly AllReduce

Attempt Score: 6.517 / 15 - F

Overall Grade (highest attempt): 6.517 / 15 - F

Done