ELMO vs BERT for NER

1) **Memory ::**

BERT

- saved memory : 418M (including 30k word embeddings)
- memory on GPU
    - model size = 439083520 Bytes
    - model size with input batch size 128 = 439607808 Bytes
    - model size with input batch size 512 = 441180672 Bytes
- I was unable to use. batch size of 1024
- input size per 1 sentence = 4096 Bytes
    - input_ids,input_mask,segment_mask,label_ids each of length 128 and type long which is 8 bytes.
    - if we go by this ideally can fit a large batch but GPU has a lot of cached memory when pytorch loads.
- 700 is the maximum batch size that i was able to use

ELMO (with a BILSTM at top)

- saved memory : 368M (only character embeddings are used)
- memory on GPU
    - model size = 384921600 Bytes
    - model size with input batch size 128 = 401174528 Bytes
- I was unable to use a batch size of 512.
- input size per 1 sentence = 126,976
    - input_ids =128*50*8, input_mask = 128*8 , label_ids = 128*8
    - (there is some inconsistency here as I'm getting half the size allocated)

- 256 is maximum batch size that I was able to use

Comments ::

Numbers reported above does not include cached memory , Elmo's inputs are large and here I haven't considered the word embeddings as mentioned in the paper so if there are multiple GPUs with large RAM BERT will be pretty fast.

2) **Time taken while predicting**

BERT

- model_loading => 4.5 sec
- forward pass is taking very small time (0.011 seconds)

ELMO

- model loading ==> 16.48

- forward pass is taking significant time (0.96 seconds) which is expected as lstm are involved .

Comments ::

Moving logits from CPU to GPU is taking different times for the two models i.e; taking little time in ELMO based model while taking a significant time in BERT based model with the same code, I'm not sure why this is happening (raised an issue in github awaiting answers).

Issue :: https://discuss.pytorch.org/t/understanding-time-taken-from-moving-data-from-gpu-to-cpu/46680

3) **Performance**

BERT

	prec	recall	f1	support
I	0.92	0.94	0.93	280
B	0.95	0.99	0.97	303

ELMO

	prec	recall	f1	support
I	0.81	0.83	0.83	280
B	0.90	0.86	0.93	303

Comments::

This is with AGE dataset with only age related quantities tagged as BIO . Has performed extensive hyper parameter tuning but the results for BERT are far superior. I haven't performed any post processing on the tags .