An Error Report on attempts to train a pure transformer model in sockeye

First attempt with those parameters (according to “Attention is all you need” base parameters)**:**

```
--encoder transformer
--decoder transformer
--num-layers NUM_LAYERS
                        Number of layers for encoder & decoder. Use "x:x"
to
                        specify separate values for encoder & decoder.
                        Default: (6, 6).#again, default is alright
  --transformer-model-size TRANSFORMER_MODEL_SIZE
                        Number of hidden units in transformer layers. Use
                        "x:x" to specify separate values for encoder &
                        decoder. Default: (512, 512).#I think default is
alright, see 3.1 d_model
  --transformer-attention-heads TRANSFORMER_ATTENTION_HEADS
                        Number of heads for all self-attention when using
                        transformer layers. Use "x:x" to specify separate
                        values for encoder & decoder. Default: (8, 8).
             #In the paper see section 3.2.2
             #In this work we employ h = 8 parallel attention layers, or
heads.
  --transformer-feed-forward-num-hidden
TRANSFORMER_FEED_FORWARD_NUM_HIDDEN
                        Number of hidden units in transformers feed
forward
                        layers. Use "x:x" to specify separate values for
                        encoder & decoder. Default: (2048, 2048). #is the
same as d_ff
  --transformer-activation-type {gelu,relu,swish1}
```

```
Type activation to use for each feed forward layer.
                        Default: relu.
  --transformer-positional-embedding-type learned #was said to acquire
equal results in paper
  --transformer-preprocess rn #please check section 3.1 in paper
  --transformer-postprocess rn #please check section 3.1 in paper
--transformer-dropout-attention TRANSFORMER_DROPOUT_ATTENTION
                        Dropout probability for multi-head attention.
Default:
```

```
                        0.1. #Id say leave it default, same as P_drop
 --transformer-dropout-act TRANSFORMER_DROPOUT_ACT
```

Dropout probability before activation in feed-forward
block. Default: 0.1. #I d say leave it default
--transformer-dropout-prepost TRANSFORMER_DROPOUT_PREPOST
Dropout probability for pre/postprocessing blocks.
Default: 0.1.#I d say leave it default
--optimizer adam

```
es/sockeye/utils.py", line 118, in check_condition
    raise SockeyeError(error_message)
sockeye.utils.SockeyeError: No GPUs found, consider running on the CPU with --us
e-cpu
```

```
[08:02:17] src/imperative/./imperative_utils.h:76: GPU support is disabled. Comp
ile MXNet with USE_CUDA=1 to enable GPU support.
```

Comment: The training always stopped after 10 min giving this error. We tried to reinstall the environments, but nothing worked during whole saturday. Installing MXnet 9.0 instead of MXnet 8.0 made the model train, but it only delivered BLEU Score 1.0011473925146788e-07'.

Second attempt with the following sbatch train file:

```
#!/bin/bash

#The name of the job is test_job
#SBATCH -J ATT1_PRJ

#The job requires 1 compute node
#SBATCH -N 1

#The job requires 1 task per node
#SBATCH --ntasks-per-node=1

#SBATCH --exclude=falcon3

#The maximum walltime of the job is a 8 days
#SBATCH --time=6-23:59:59

#SBATCH --mem=30G

#Leave this here if you need a GPU for your job
#SBATCH --partition=gpu

#SBATCH --gres=gpu:tesla:1
```

```
# OUR COMMANDS GO HERE

module load python/3.6.3/CUDA-8.0

source activate mtenv-cuda8-1

python -m sockeye.train --disable-device-locking \
                        --device-ids 0 \
                        -s data/bpe.cleaned.tc.tok.train.et \
                        -t data/bpe.cleaned.tc.tok.train.en \
                        -vs data/bpe.cleaned.tc.tok.dev.et \
                        -vt data/bpe.cleaned.tc.tok.dev.en \
                        -o experiments/model \
                        --encoder transformer \
                        --decoder transformer \
                        --num-layers 6:6 \
              --transformer-model-size 512:512 \
                        --transformer-attention-heads 8:8 \
              --transformer-feed-forward-num-hidden 2048:2048 \
              --transformer-activation-type relu \
              --transformer-positional-embedding-type learned \
              --transformer-preprocess dn \
              --transformer-postprocess dn \
              --transformer-dropout-attention 0.1 \
              --transformer-dropout-act 0.1 \
              --transformer-dropout-prepost 0.1 \
                        --max-seq-len 100 \
                        --batch-type sentence \
                        --batch-size 64 \
                        --checkpoint-frequency 3377 \
                        --optimizer adam \
                        --initial-learning-rate 0.001 \
                        --optimized-metric perplexity \
                        --min-num-epochs 3 \
                        --max-num-checkpoint-not-improved 4
```

This is the error that we get when we use `--transformer-preprocess rn` and `--transformer-postprocess rn`

```
  File "/gpfs/hpchome/jayasing/.conda/envs/mtenv-cuda8-1/lib/python3.6/site-pack
ages/sockeye/transformer.py", line 230, in __call__
    assert 'r' not in self.sequence, "Residual connection not allowed if no prev
ious value given."
AssertionError: Residual connection not allowed if no previous value given.
```

Afterwards we tried without those two parameters, but the resulting model was also rubbish.

We tried some more parameters, but nothing did work out so far. We will try to reduce parameters now and train a simple default transformer model in sockeye.