English/German MT

Andrew Smith

Dataset

- Transformers
- Para_crawl/ende_plain_text
- tfds.load('para_crawl/ende_plain_text', with_info=True, as_supervised=True)

- LSTM, GRU, RNN
- https://raw.githubusercontent.com/jbrownlee/Datasets/master/deu.txt

BLEU Scores

- Corpus_bleu
- From nltk.translate.bleu_score
- Usage: corpus_bleu(actual, predicted,weights)
- weights for unigrams, bigrams, trigrams and so on
- BLEU-1: weights=(1.0, 0, 0, 0))
- BLEU-2: weights=(0.5, 0.5, 0, 0))
- BLEU-3: weights=(0.3, 0.3, 0.3, 0))
- BLEU-4: weights=(0.25, 0.25, 0.25, 0.25))
- BLEU: no weights passed in

RNNs

- https://github.com/asmith1138/MachineTranslation/blob/main/ML_tensorflow.ipynb
- SimpleRNN, GRU, LSTM
- https://machinelearningmastery.com/develop-neural-machine-translation-system-keras/
- Trained with 9000 sentences
- Test was 1000 sentences

RNN

Layer (type)	Output Shape	Param #
embedding_6 (Embedding)	(None, 10, 256)	987136
simple_rnn (SimpleRNN)	(None, 256)	131328
repeat_vector_4 (RepeatVecto	(None, 5, 256)	0
simple_rnn_1 (SimpleRNN)	(None, 5, 256)	131328
time_distributed_3 (TimeDist	(None, 5, 2404)	617828

Total params: 1,867,620

Trainable params: 1,867,620

Non-trainable params: 0

src=[sei dir nicht so sicher], target=[dont be so sure], predicted=[dont be so sure]

src=[hat tom viel zu tun], target=[is tom busy], predicted=[did tom in]

src=[tom hat viel gelesen], target=[tom read a lot], predicted=[tom needs a car]

src=[jeder hat es gesehen], target=[everybody saw it], predicted=[who saw it]

src=[das ist so stumpfsinnig], target=[this is so dumb], predicted=[thats is so]

src=[alles ist gut], target=[everything is ok], predicted=[its all right]

src=[ich werde es mir ansehen], target=[ill watch it], predicted=[i know watch it]

src=[ich sagte doch bleib zuruck], target=[i said stay back], predicted=[i i for tom]

src=[ist tom allein], target=[is tom alone], predicted=[is tom alone]

src=[nehmen sie einen bus], target=[take a bus], predicted=[take a bus]

BLEU-1: 0.525032

BLEU-2: 0.398027

BLEU-3: 0.331349

BLEU-4: 0.185056

GRU

Layer (type)	Output Shape	Param #
embedding_7 (Embedding)	(None, 10, 256)	987136
gru_3 (GRU)	(None, 256)	394752
repeat_vector_5 (RepeatVecto	(None, 5, 256)	0
gru_4 (GRU)	(None, 5, 256)	394752
time_distributed_4 (TimeDist	(None, 5, 2404)	617828

Total params: 2,394,468

Trainable params: 2,394,468

Non-trainable params: 0

src=[sei dir nicht so sicher], target=[dont be so sure], predicted=[dont be so sure]

src=[hat tom viel zu tun], target=[is tom busy], predicted=[is tom a yet]

src=[tom hat viel gelesen], target=[tom read a lot], predicted=[tom has a lot]

src=[jeder hat es gesehen], target=[everybody saw it], predicted=[nobody liked that]

src=[das ist so stumpfsinnig], target=[this is so dumb], predicted=[its so so]

src=[alles ist gut], target=[everything is ok], predicted=[its is good]

src=[ich werde es mir ansehen], target=[ill watch it], predicted=[i will watch it]

src=[ich sagte doch bleib zuruck], target=[i said stay back], predicted=[i just up back]

src=[ist tom allein], target=[is tom alone], predicted=[did tom alone]

src=[nehmen sie einen bus], target=[take a bus], predicted=[take a bus]

BLEU-1: 0.531213

BLEU-2: 0.401636

BLEU-3: 0.335122

BLEU-4: 0.183399

LSTM

Layer (type)	Output Shape	Param #
embedding_8 (Embedding)	(None, 10, 256)	987136
lstm_6 (LSTM)	(None, 256)	525312
repeat_vector_6 (RepeatVecto	(None, 5, 256)	0
lstm_7 (LSTM)	(None, 5, 256)	525312
time_distributed_5 (TimeDist	(None, 5, 2404)	617828

Total params: 2,655,588

Trainable params: 2,655,588

Non-trainable params: 0

src=[sei dir nicht so sicher], target=[dont be so sure], predicted=[dont be so sure]

src=[hat tom viel zu tun], target=[is tom busy], predicted=[is tom a now]

src=[tom hat viel gelesen], target=[tom read a lot], predicted=[tom is a dog]

src=[jeder hat es gesehen], target=[everybody saw it], predicted=[who found it]

src=[das ist so stumpfsinnig], target=[this is so dumb], predicted=[its so bad]

src=[alles ist gut], target=[everything is ok], predicted=[its is right]

src=[ich werde es mir ansehen], target=[ill watch it], predicted=[i will it it]

src=[ich sagte doch bleib zuruck], target=[i said stay back], predicted=[i lost tom home]

src=[ist tom allein], target=[is tom alone], predicted=[is tom alone]

src=[nehmen sie einen bus], target=[take a bus], predicted=[take a bus]

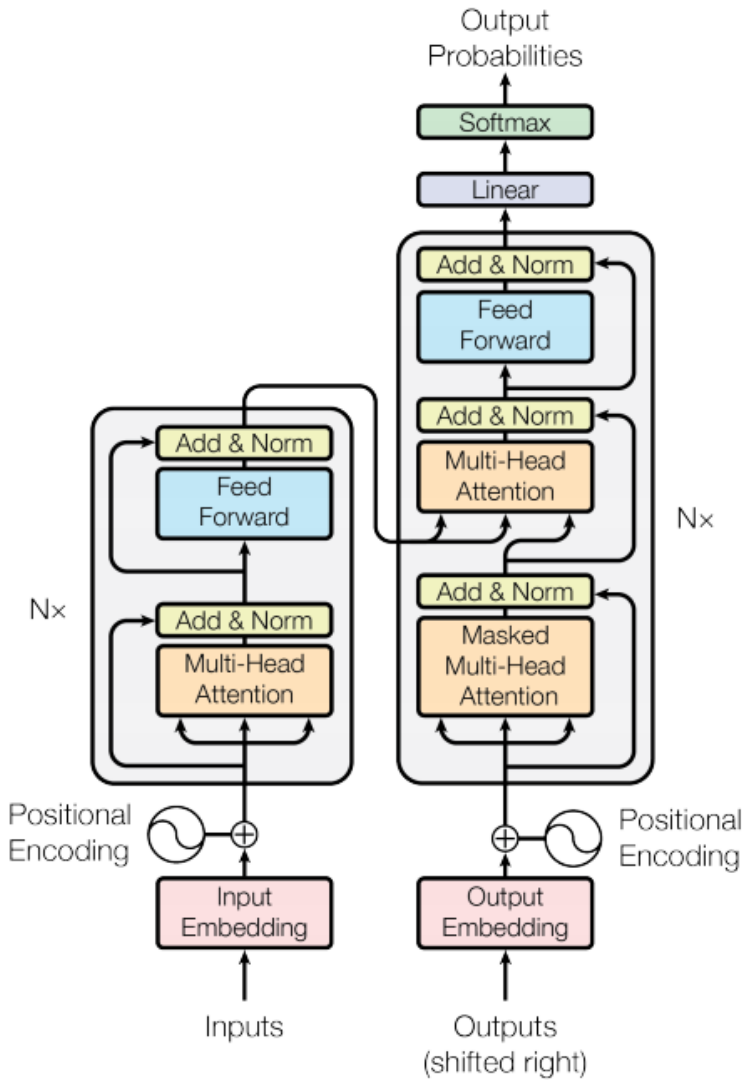
BLEU-1: 0.489177

BLEU-2: 0.355798

BLEU-3: 0.283467

BLEU-4: 0.145477

Transformers



Tokenizer

- Trained tokenizers as BertTokenizer
- Para_crawl dataset
- https://github.com/asmith1138/MachineTranslation/blob/main/MT_subwords_tokenizer.ipynb
- https://www.tensorflow.org/tutorials/tensorflow_text/subwords_tokenizer

3 Transformer Models

https://www.tensorflow.org/tutorials/text/transformer

TensorFlow Example

- num_layers = 4
- d_model = 128
- dff = 512
- num_heads = 8
- dropout_rate = 0.1

Transformer Paper

- num_layers = 6
- d_model = 512
- dff = 2048
- num_heads = 8
- dropout_rate = 0.1

Custom

- num_layers = 16
- d_model = 512
- dff = 2048
- num_heads = 16
- dropout_rate = 0.1

Tensorflow Example Transformer

Epoch 1 Loss 6.6696 Accuracy 0.0883
Epoch 2 Loss 5.1106 Accuracy 0.1675
Epoch 3 Loss 4.2967 Accuracy 0.2524
Epoch 4 Loss 3.6377 Accuracy 0.3448
Epoch 5 Loss 3.1756 Accuracy 0.4142
Epoch 6 Loss 2.9090 Accuracy 0.4541
Epoch 7 Loss 2.7322 Accuracy 0.4808
Epoch 8 Loss 2.6057 Accuracy 0.5000
Epoch 9 Loss 2.5090 Accuracy 0.5149
Epoch 10 Loss 2.4342 Accuracy 0.5265
Epoch 11 Loss 2.3711 Accuracy 0.5362
Epoch 12 Loss 2.3199 Accuracy 0.5439
Epoch 13 Loss 2.2746 Accuracy 0.5514
Epoch 14 Loss 2.2377 Accuracy 0.5573
Epoch 15 Loss 2.2036 Accuracy 0.5625
Epoch 16 Loss 2.1749 Accuracy 0.5669
Epoch 17 Loss 2.1472 Accuracy 0.5715
Epoch 18 Loss 2.1247 Accuracy 0.5752
Epoch 19 Loss 2.1031 Accuracy 0.5783
Epoch 20 Loss 2.0830 Accuracy 0.5815
Epoch 21 Loss 2.0645 Accuracy 0.5845
Epoch 22 Loss 2.0491 Accuracy 0.5869
Epoch 23 Loss 2.0332 Accuracy 0.5895
Epoch 24 Loss 2.0182 Accuracy 0.5917
Epoch 25 Loss 2.0055 Accuracy 0.5935
Epoch 26 Loss 1.9924 Accuracy 0.5958
Epoch 27 Loss 1.9807 Accuracy 0.5976
Epoch 28 Loss 1.9695 Accuracy 0.5993
Epoch 29 Loss 1.9597 Accuracy 0.6008
Epoch 30 Loss 1.9501 Accuracy 0.6024
Epoch 31 Loss 1.9399 Accuracy 0.6041
Epoch 32 Loss 1.9313 Accuracy 0.6054
Epoch 33 Loss 1.9228 Accuracy 0.6067
Epoch 34 Loss 1.9150 Accuracy 0.6078
Epoch 35 Loss 1.9084 Accuracy 0.6091
Epoch 36 Loss 1.9002 Accuracy 0.6104
Epoch 37 Loss 1.8939 Accuracy 0.6113
Epoch 38 Loss 1.8868 Accuracy 0.6125
Epoch 39 Loss 1.8802 Accuracy 0.6135
Epoch 40 Loss 1.8737 Accuracy 0.6146
Epoch 41 Loss 1.8671 Accuracy 0.6156
Epoch 42 Loss 1.8622 Accuracy 0.6163
Epoch 43 Loss 1.8568 Accuracy 0.6174
Epoch 44 Loss 1.8501 Accuracy 0.6182
Epoch 45 Loss 1.8465 Accuracy 0.6188
Epoch 46 Loss 1.8413 Accuracy 0.6200
Epoch 47 Loss 1.8375 Accuracy 0.6201
Epoch 48 Loss 1.8325 Accuracy 0.6207
Epoch 49 Loss 1.8275 Accuracy 0.6218
Epoch 50 Loss 1.8233 Accuracy 0.6225

Final Scores

BLEU-1: 0.196988

BLEU-2: 0.443833

BLEU-3: 0.614231

BLEU-4: 0.666208

BLEU: 0.666208

GitHub

Transformer from Paper

Epoch 1 Loss 6.3979 Accuracy 0.1026
Epoch 2 Loss 5.0926 Accuracy 0.1621
Epoch 3 Loss 4.6122 Accuracy 0.1990
Epoch 4 Loss 4.2422 Accuracy 0.2376
Epoch 5 Loss 3.9796 Accuracy 0.2671
Epoch 6 Loss 3.7978 Accuracy 0.2887
Epoch 7 Loss 3.6574 Accuracy 0.3057
Epoch 8 Loss 3.5427 Accuracy 0.3195
Epoch 9 Loss 3.4450 Accuracy 0.3319
Epoch 10 Loss 3.3601 Accuracy 0.3424
Epoch 11 Loss 3.2835 Accuracy 0.3522
Epoch 12 Loss 3.2148 Accuracy 0.3607
Epoch 13 Loss 3.1500 Accuracy 0.3693
Epoch 14 Loss 3.0927 Accuracy 0.3771
Epoch 15 Loss 3.0379 Accuracy 0.3841
Epoch 16 Loss 2.9859 Accuracy 0.3906
Epoch 17 Loss 2.9381 Accuracy 0.3971
Epoch 18 Loss 2.8904 Accuracy 0.4036
Epoch 19 Loss 2.8472 Accuracy 0.4095
Epoch 20 Loss 2.8067 Accuracy 0.4151
Epoch 21 Loss 2.7662 Accuracy 0.4204
Epoch 22 Loss 2.7279 Accuracy 0.4260
Epoch 23 Loss 2.6918 Accuracy 0.4308
Epoch 24 Loss 2.6572 Accuracy 0.4358
Epoch 25 Loss 2.6234 Accuracy 0.4405
Epoch 26 Loss 2.5914 Accuracy 0.4448
Epoch 27 Loss 2.5600 Accuracy 0.4495
Epoch 28 Loss 2.5295 Accuracy 0.4536
Epoch 29 Loss 2.5009 Accuracy 0.4582
Epoch 30 Loss 2.4730 Accuracy 0.4621
Epoch 31 Loss 2.4450 Accuracy 0.4658
Epoch 32 Loss 2.4189 Accuracy 0.4702
Epoch 33 Loss 2.3933 Accuracy 0.4740
Epoch 34 Loss 2.3679 Accuracy 0.4777
Epoch 35 Loss 2.3435 Accuracy 0.4811
Epoch 36 Loss 2.3200 Accuracy 0.4849
Epoch 37 Loss 2.2970 Accuracy 0.4879
Epoch 38 Loss 2.2738 Accuracy 0.4914
Epoch 39 Loss 2.2525 Accuracy 0.4950
Epoch 40 Loss 2.2308 Accuracy 0.4982
Epoch 41 Loss 2.2100 Accuracy 0.5014
Epoch 42 Loss 2.1900 Accuracy 0.5045
Epoch 43 Loss 2.1688 Accuracy 0.5079
Epoch 44 Loss 2.1518 Accuracy 0.5104
Epoch 45 Loss 2.1313 Accuracy 0.5137
Epoch 46 Loss 2.1121 Accuracy 0.5167
Epoch 47 Loss 2.0945 Accuracy 0.5194
Epoch 48 Loss 2.0761 Accuracy 0.5225
Epoch 49 Loss 2.0590 Accuracy 0.5251
Epoch 50 Loss 2.0416 Accuracy 0.5279

BLEU-1: 0.187514

BLEU-2: 0.433029

BLEU-3: 0.605215

BLEU-4: 0.658049

BLEU: 0.658049

GitHub

Custom Transformer

- Never finished running
- https://github.com/asmith1138/MachineTranslation/blob/main/MT_eng_deu_transformer_custom.ipynb

References

- How to Develop a Neural Machine Translation System from Scratch
- datasets/datasets at master · huggingface/datasets
- List of Transformer tutorials for Deep Learning – MachineCurve
- Easy Machine Translation with Machine Learning and HuggingFace Transformers – MachineCurve
- Transformer model for language understanding | TensorFlow Core
- Subword tokenizers | TensorFlow Core
- https://raw.githubusercontent.com/jbrownlee/Datasets/master/deu.txt
- https://www.kite.com/python/docs/nltk.bleu_score.corpus_bleu
- https://www.tensorflow.org/datasets/catalog/para_crawl