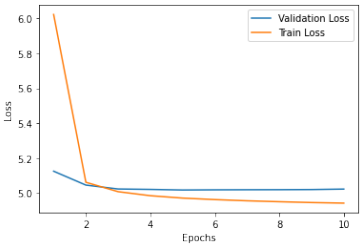
HW2 Report

OUDAOUD El Mehdi

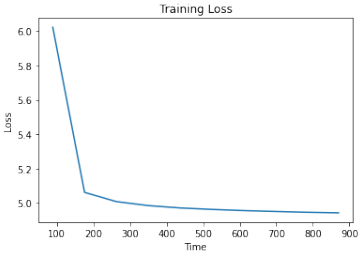
November 2022

1 Question 3.1

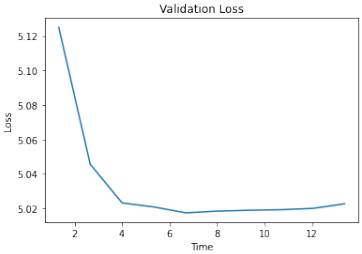
1.1 Seq 2 Seq



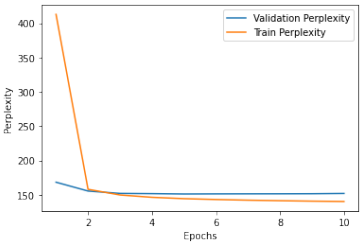
(a) Loss over epochs



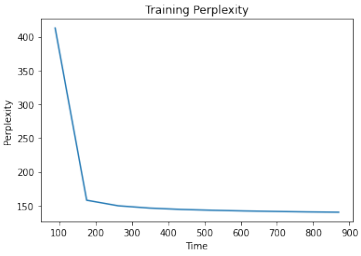
(b) Trainig Loss over Time



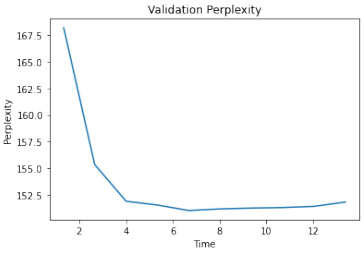
(c) Validation Loss over Time



(d) Perplexity over epochs

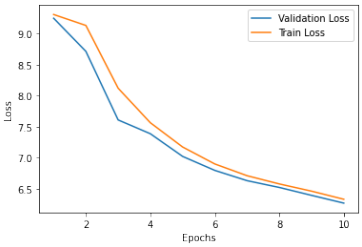


(e) Trainig Perplexity over Time

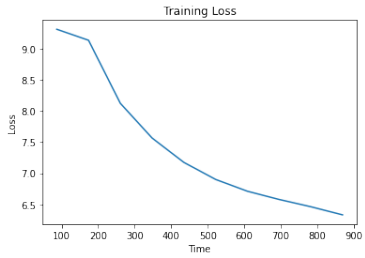


(f) Validation Perplexity over Time

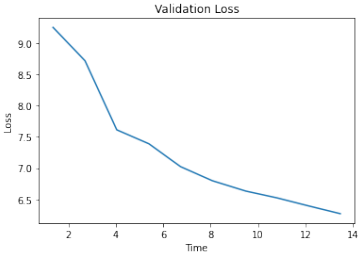
Figure 1: Loss and Perplexity over epochs and Time of seq2seq model with adam; no dopout,and greedy decoding



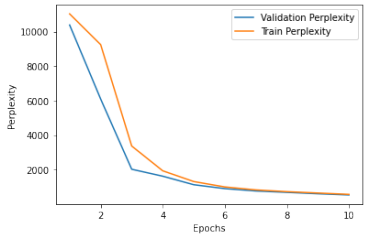
(a) Loss over epochs



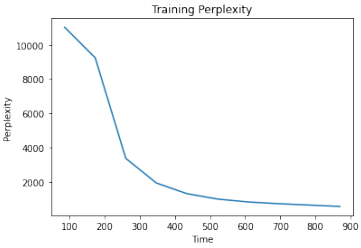
(b) Trainig Loss over Time



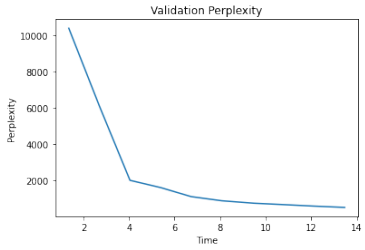
(c) Validation Loss over Time



(d) Perplexity over epochs

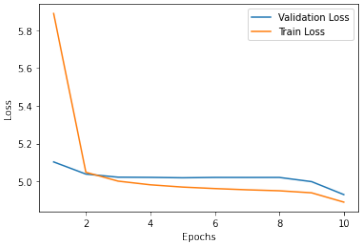


(e) Training Perplexity over Time

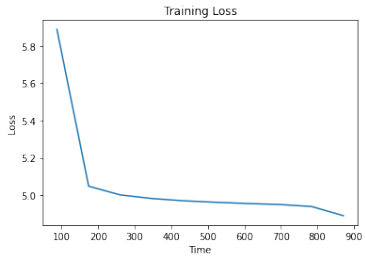


(f) Validation Perplexity over Time

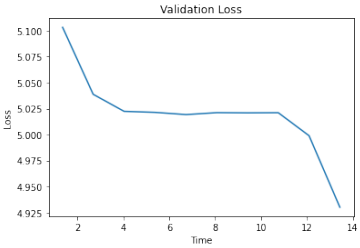
Figure 2: Loss and Perplexity over epochs and Time of seq2seq model with SGD optimizer, with dopout,and greedy decoding



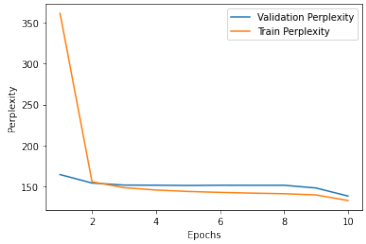
(a) Loss over epochs



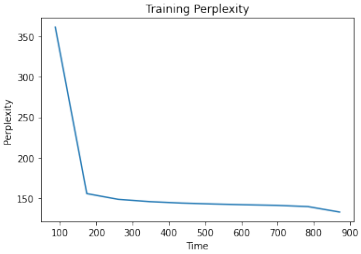
(b) Trainig Loss over Time



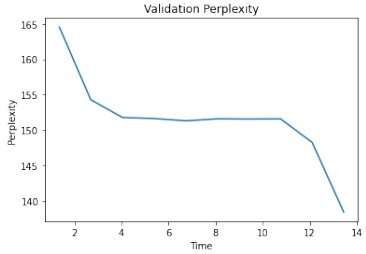
(c) Validation Loss over Time



(d) Perplexity over epochs

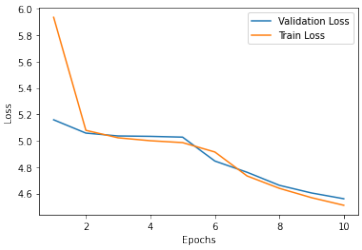


(e) Trainig Perplexity over Time

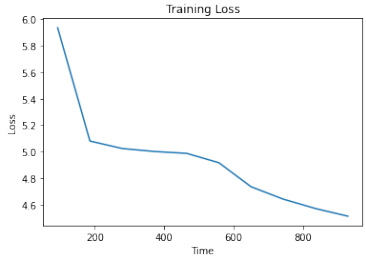


(f) Validation Perplexity over Time

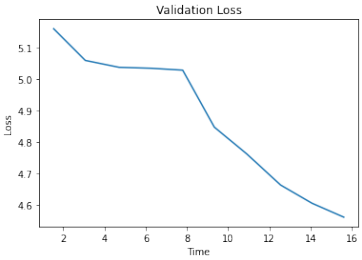
Figure 3: Loss and Perplexity over epochs and Time of seq2seq model with Adam optimizer, with dropout,and greedy decoding



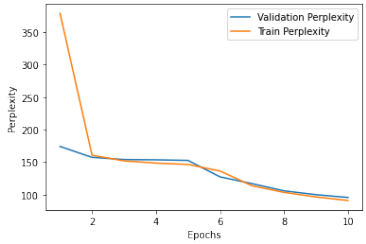
(a) Loss over epochs



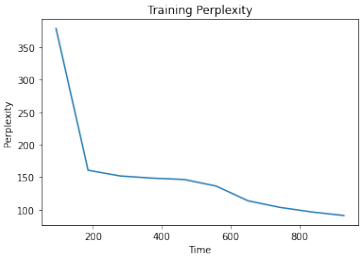
(b) Trainig Loss over Time



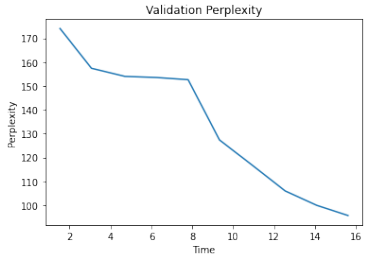
(c) Validation Loss over Time



(d) Loss over epochs



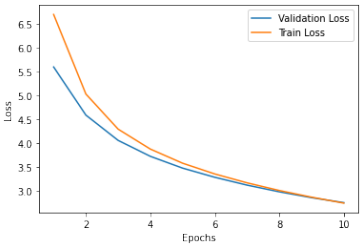
(e) Trainig Loss over Time



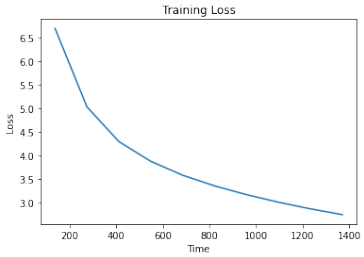
(f) Validation Loss over Time

Figure 4: Loss over epochs and Time of seq2seq model with Adam optimizer, with dopout,and random Temperature decoding

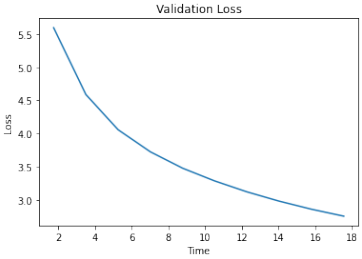
1.2 Transformer



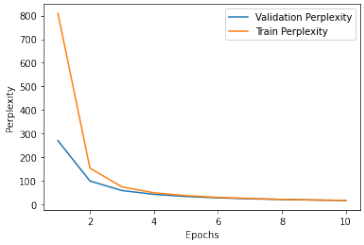
(a) Loss over epochs



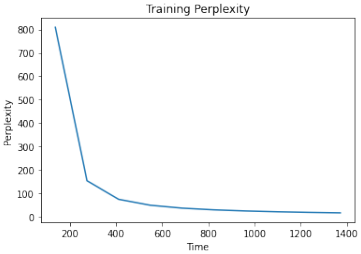
(b) Trainig Loss over Time



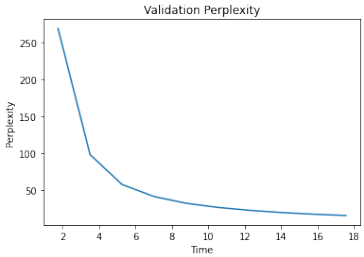
(c) Validation Loss over Time



(d) Perplexity over epochs

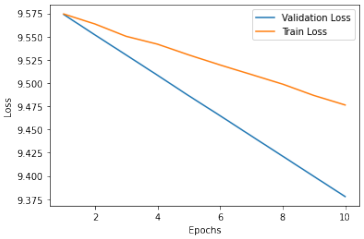


(e) Trainig Perplexity over Time

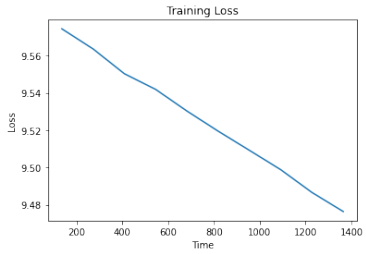


(f) Validation Perplexity over Time

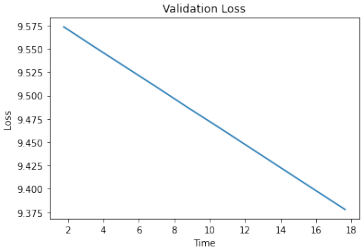
Figure 5: Loss and Perplexity over epochs and Time of transformer with adam; no dopout,and greedy decoding



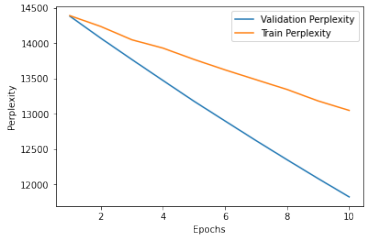
(a) Loss over epochs



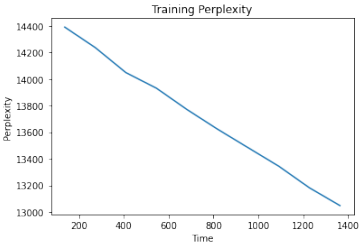
(b) Trainig Loss over Time



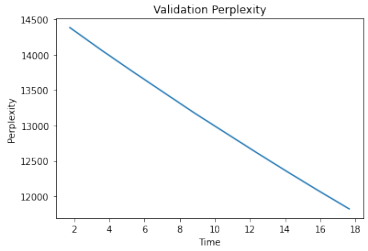
(c) Validation Loss over Time



(d) Perplexity over epochs

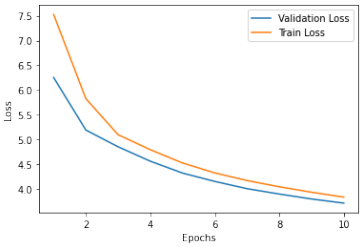


(e) Trainig Perplexity over Time

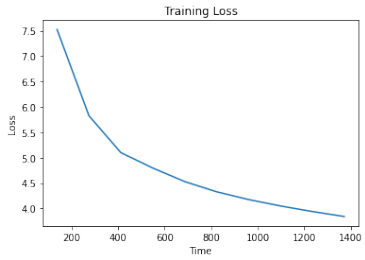


(f) Validation Perplexity over Time

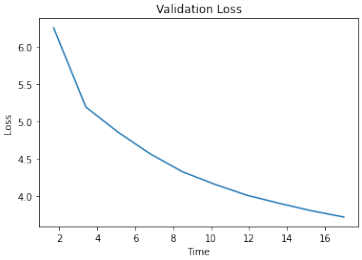
Figure 6: Loss and Perplexity over epochs and Time of transformer with SGD Optimizer with dopout,and greedy decoding



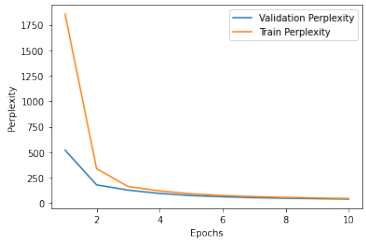
(a) Loss over epochs



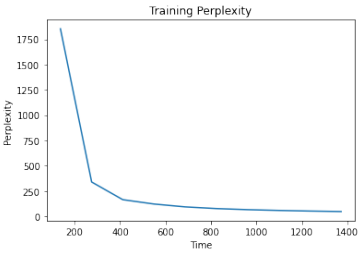
(b) Trainig Loss over Time



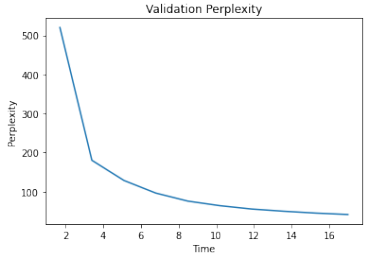
(c) Validation Loss over Time



(d) Perplexity over epochs

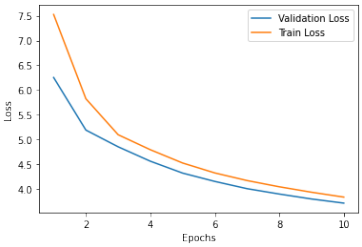


(e) Trainig Perplexity over Time

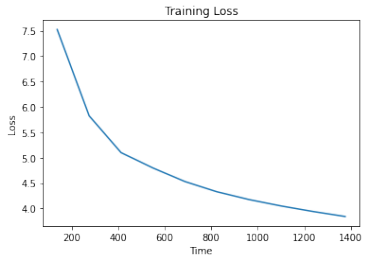


(f) Validation Perplexity over Time

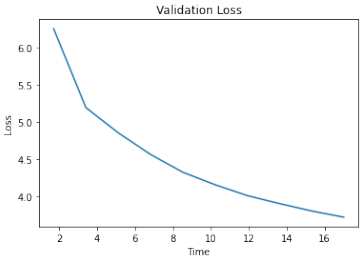
Figure 7: Loss and Perplexity over epochs and Time of transformer with Adam Optimizer with dopout,and greedy decoding



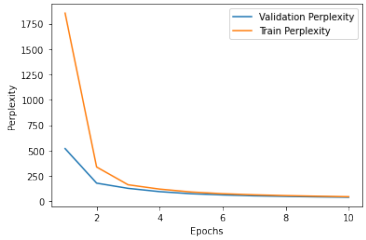
(a) Loss over epochs



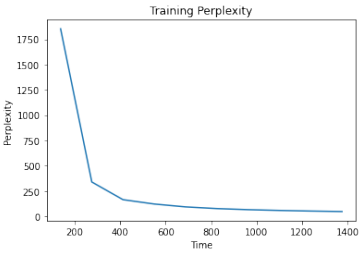
(b) Trainig Loss over Time



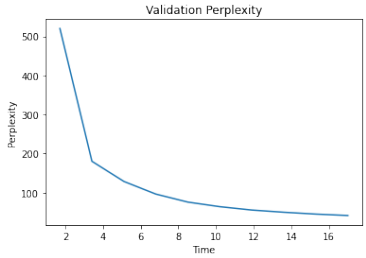
(c) Validation Loss over Time



(d) Perplexity over epochs



(e) Trainig Perplexity over Time



(f) Validation Perplexity over Time

Figure 8: Loss and Perplexity over epochs and Time of transformer with Adam Optimizer with dopout,and random temperature decoding

2 Question 3.2

Architecture	use_dropout	Optimizer	Decoding	Train Loss	Train PPL	Val Loss	Val PPL
GRU	False	Adam	Greedy	4.94	140.08	5.017	151.017
GRU	True	SGD	Greedy	6.3	562.82	6.27	528.978
GRU	True	Adam	Greedy	4.89	132.992	4.93	138.38
GRU	**True**	**Adam**	**Random**	**4.51**	**91.07**	**4.56**	**95.68**
Transformer	**False**	**Adam**	**Greedy**	**2.74**	**15.55**	**2.754**	**15.70**
Transformer	True	SGD	Greedy	9.47	13047.93	9.38	11822.38
Transformer	True	Adam	Greedy	3.84	46.58	3.72	41.26
Transformer	True	Adam	Random	3.84	46.58	3.72	41.26

Architecture	use_dropout	Optimizer	Decoding	Test BLEU-1	Test BLEU-2
GRU	False	Adam	Greedy	24.05 %	14.34 %
GRU	True	SGD	Greedy	3.54 %	0 %
GRU	True	Adam	Greedy	27.18 %	16.11 %
GRU	**True**	**Adam**	**Random**	**33.80 %**	**16.87 %**
Transformer	**False**	**Adam**	**Greedy**	**45.32 %**	**59.51 %**
Transformer	True	SGD	Greedy	0.0 %	0.026 %
Transformer	True	Adam	Greedy	23.04 %	36.57 %
Transformer	True	Adam	Random	22.00 %	37.82 %

The best configuration for GRU is : Adam optimizer, use_dropout=True, random with temperature decoding.

The best configuration for Transformer is : Adam optimizer, use_dropout=False, greedy decoding

3 Question 3.3

If concerned with Time, the best model would be the GRU with Adam optimizer, use_dropout=False, greedy decoding, since the train_time average is 87 sec, the test_time is 1.31 and the valid_time average is 1.33.

If concerned with generalization performance, the best model would be the Transformer Model with Adam optimizer, use dropout=True, ran- dom with temperature decoding.

Naturally, the dopout delays the model training and evaluation as shown when comapring the configuration 1. and 2. of each model.

However, the dropout is supposed to improve the generalization performance of the model which is the case for GRU configuration 1. vs 3.; but is not the case for TRansformers conf 1. vs conf 3.

4 Question 3.4

The Transformer model is better performing because the losses (validation and Training) are better, also it got higher blue scores and the perplexity (Training and Validation) is better as well (lower).

5 Question 3.5

I expected the Transformer models with dropout to have better generalization performance, which was not the case, this could be due to the fact that few epochs were demonstrated or the model was not deep enough for the changes to show.

Also, with the optimizer SGD, the model scored the lowest but still has a good generalization performance (Training and Validation have close values). But the reason SGD optimizer performs less is due to the fact that it is slower so it might need more epochs.

6 Question 3.6

All the GRU models have few to no comnsumption of GRU (the usage is bound between 0GiB and 2 GiB).

For the TRansformer Models.

Transformer Model configuration : SGD Transformer, with dropout, and greedy strategy decoding $->$ 1408 MiB

Transformer Model configuration : Adam Transformer, No dropout, and greedy strategy decoding $->$ 436 MiB

Transformer Model configuration : Adam Transformer, with dropout, and greedy strategy decoding $->$ 0 MiB

Transformer Model configuration : Adam Transformer, with dropout, and random strategy $->$ 4 MiB

7 Question 3.7.

The overfitting is more present in the learning curves of the first configurations of each model. (Adam Optimizer, No dropout and greedy decoding). One can solve overfitting problems by data augmentation (use of synonymes, or different formulations) or by adding a dropout like the other configurations.