POLYTECHNIQUE MONTRÉAL

UNIVERSITÉ D’INGÉNIERIE

Assignment 2 Report

Hyperparameter evaluation of RNN, GRU and Transformer

Sangar Simon 1938126

Aymen Alaeddine Zeghaida 1926415

INF8225 - Intelligence artificielle : techniques probabilistes et d’apprentissage

Computer and Software engineering department

École Polytechnique de Montréal

9 april 2023

1) Hyperparameter search

a) Experimental setup

We employed the Weights and Biases “***sweeps***” functionality for conducting our parameter search.

Utilizing a randomized search technique, we explored a wide spectrum of parameters across 10 iterations for each model, with a total of 30 iterations. We started with several test iterations in order to narrow down the ranges for our parameters, and settled on a fixed number of epochs of 6 epochs for RNN/GRU and twice that amount for the Transformer.

We have 3 WandB sweeps, each responsible for random 10 iterations, with the goal of minimizing our validation loss.

b) Results

The Transformer models consistently outperformed the RNN and GRU models in terms of accuracy (**TOP1-Validation**) , which can be attributed to the fact that Transformers are more adept at processing sequential data. Unlike RNN/GRU models, Transformers can capture correlations between distant words due to their attention mechanism.

The GRU cell is a superior alternative to the RNN architecture due to its advanced design that effectively deals with the vanishing gradient problem and preserves information across multiple time steps. The number of layers is also an essential factor to consider, and a larger embedding dimension negatively impacts accuracy slightly.

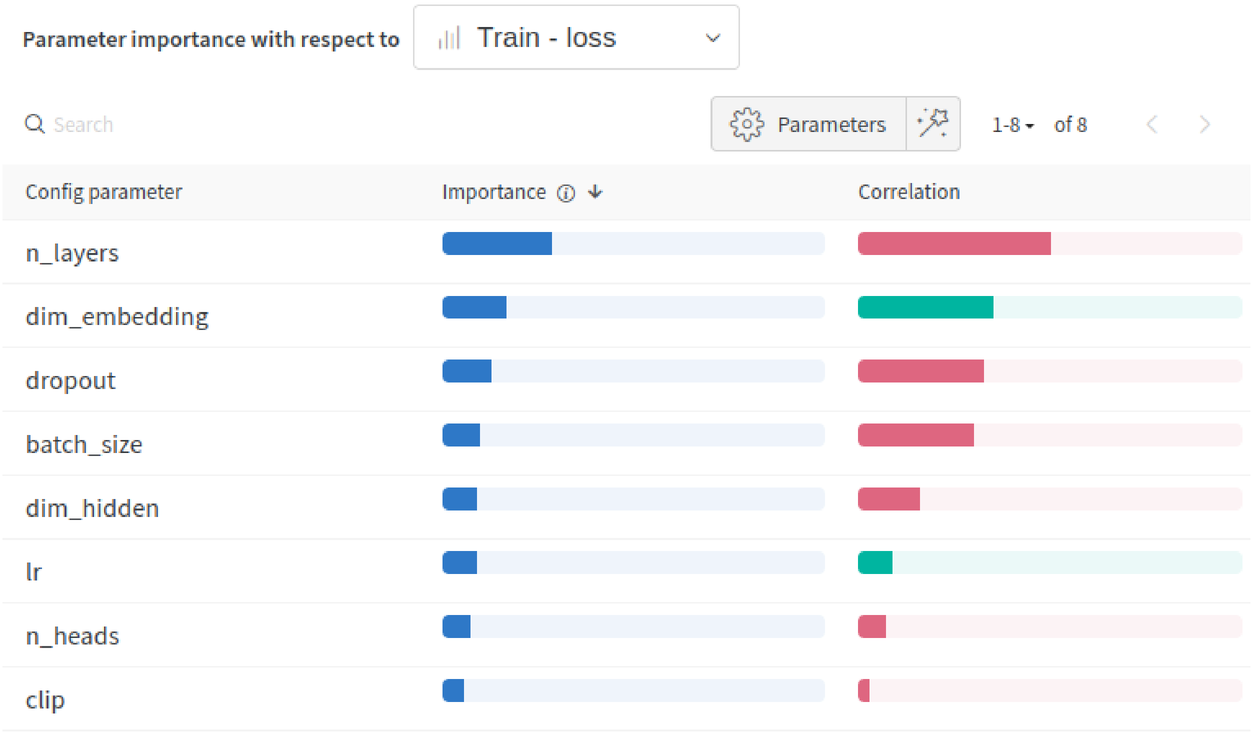
In the following table, we report the parameters as well as accuracy and loss for our best performing models, for each model type.

Table 1 Best performing model per (TOP-1 accuracy)

Model type	**Batch**	**Clip**	**embed**	**hidden**	**Dropout**	**learning rate**	**heads**	**layers**	**Validation TOP1**
Transformer	512	8	762	605	0.1626	0.0008301	11	9	**0.7307**
GRU	256	9	402	144	0.2394	0.0009384	11	8	**0.6146**
RNN	512	4	471	284	0.2919	0.0009016	2	4	**0.5313**

Out of a total of 30 runs, the Transformer model with the parameters below achieved the highest validation accuracy of 73% in 12 epochs, while the GRU achieved 61% and RNN 53%.

The following figure (Figure 1) represents the parameter importance with respect to the train accuracy.



Parameter importance with respect to Train - loss

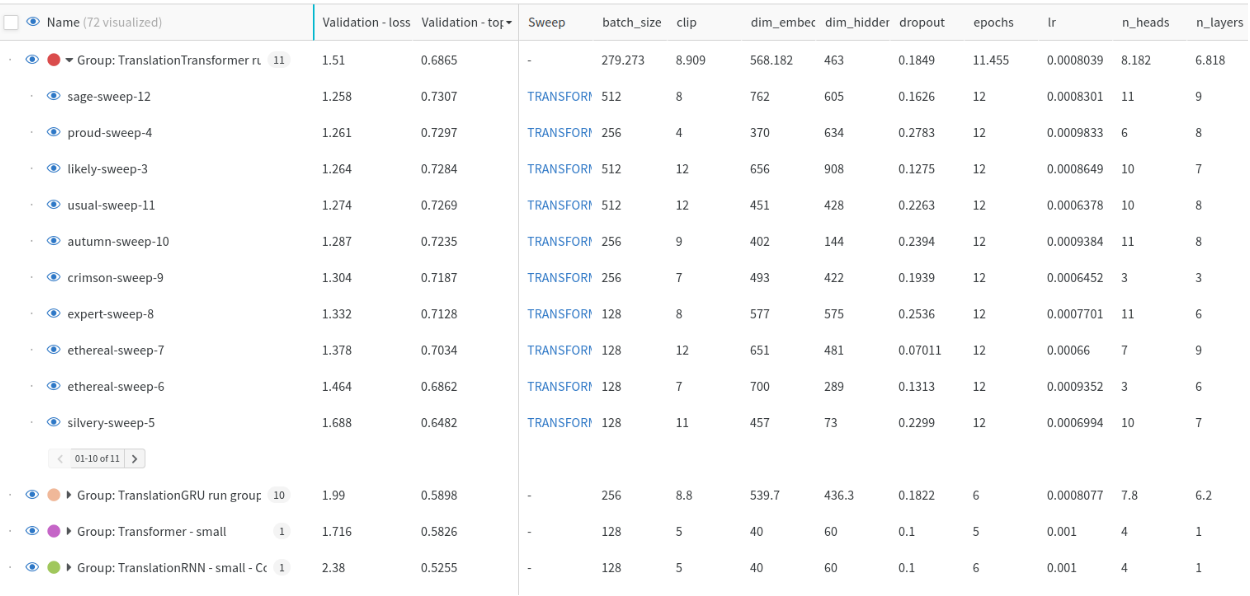
Search

Config parameter	Importance ⓘ ↓	Correlation
n_layers	🟦	🟥
dim_embedding	🟦	🟩
dropout	🟦	🟥
batch_size	🟦	🟥
dim_hidden	🟦	🟥
lr	🟦	🟩
n_heads	🟦	🟥
clip	🟦	🟥

Figure 1. Parameter importance wrt train accuracy

The analysis indicates that the number of layers and the learning rate are of the utmost importance. The significance of the number of layers can be attributed to the fact that a deeper neural network can potentially enhance or reduce performance. Interestingly, there appears to be a strong negative correlation between the number of layers and the top-1 accuracy on the training data. It was observed that increasing the number of layers results in decreased accuracy, which can be expected given our limited training time. A deeper network requires more time to train to achieve optimal results. This is supported by the positive correlation observed between the number of epochs and the accuracy.

The figure 2 showcases the random parameters that we tested with the transformer model.



Name (72 visualized)	Validation - loss	Validation - top	Sweep	batch_size	clip	dim_embec	dim_hidder	dropout	epochs	lr	n_heads	n_layers
Group: TranslationTransformer ru 11	1.51	0.6865	-	279.273	8.909	568.182	463	0.1849	11.455	0.0008039	8.182	6.818
sage-sweep-12	1.258	0.7307	TRANSFORI	512	8	762	605	0.1626	12	0.0008301	11	9
proud-sweep-4	1.261	0.7297	TRANSFORI	256	4	370	634	0.2783	12	0.0009833	6	8
likely-sweep-3	1.264	0.7284	TRANSFORI	512	12	656	908	0.1275	12	0.0008649	10	7
usual-sweep-11	1.274	0.7269	TRANSFORI	512	12	451	428	0.2263	12	0.0006378	10	8
autumn-sweep-10	1.287	0.7235	TRANSFORI	256	9	402	144	0.2394	12	0.0009384	11	8
crimson-sweep-9	1.304	0.7187	TRANSFORI	256	7	493	422	0.1939	12	0.0006452	3	3
expert-sweep-8	1.332	0.7128	TRANSFORI	128	8	577	575	0.2536	12	0.0007701	11	6
ethereal-sweep-7	1.378	0.7034	TRANSFORI	128	12	651	481	0.07011	12	0.00066	7	9
ethereal-sweep-6	1.464	0.6862	TRANSFORI	128	7	700	289	0.1313	12	0.0009352	3	6
silvery-sweep-5	1.688	0.6482	TRANSFORI	128	11	457	73	0.2299	12	0.0006994	10	7
01-10 of 11												
Group: TranslationGRU run group 10	1.99	0.5898	-	256	8.8	539.7	436.3	0.1822	6	0.0008077	7.8	6.2
Group: Transformer - small 1	1.716	0.5826	-	128	5	40	60	0.1	5	0.001	4	1
Group: TranslationRNN - small - Cc 1	2.38	0.5255	-	128	5	40	60	0.1	6	0.001	4	1

Figure 2 Parameters searched Transformer

Upon initial observation, it is apparent that the Transformer model outperforms the other two models significantly, with the lowest accuracy recorded being 0.648. Unlike the GRU/RNN models, it appears that the number of layers does not have a significant impact on the accuracy.

This is evident from the fact that the best performing model in the first line has 9 layers . It is also notable that batch size appears to have a substantial effect on the accuracy of the models.

We provide in the following link access to all visualizations and sweep configurations. Visit : https://wandb.ai/8225_team_/INF8225%20-%20TP3

2) Beyond

Incorporating the activation functions and optimizers into the parameter search would provide valuable insights into their impact on model performance. To reduce search time, a random search method was utilized, but a grid search approach can be employed in the next step to explore the combination of parameters that yielded the highest accuracy. This exhaustive search method will test all possible combinations of the specified parameters. Such a search would further optimize the model and potentially provide deeper insights into the interplay between the different parameters.