GROUP 9 PROJECT REPORT on

Translating Cultures: A Neural Approach to English-Hindi Translation

Submitted By,

Vamshi Krishna Madhavan, Akshay Manoj, Nikita Anil Yadav

University of Houston

Course: EDS 6397 – Natural Language Processing

Date: 12/02/2024

INTRODUCTION

Language bridges cultural and social gaps, while translation systems are very significant in encouraging communication among diverse linguistic groups. Of course, the attention towards English-to-Hindi and Hindi-to-English translations are increasing noticeably, because Hindi stands out as one of the most spoken languages in some regions. Nevertheless, building accurate machine translation systems for English-to-Hindi and vice versa poses special challenges in syntax, cultural subtlety, and the relatively lower availability of high-quality parallel corpora.

The recent advancements in neural machine translation in NLP have opened floodgates for more advanced and reliable solutions to such problems. The models used in this project were mBART and mT5, which uses the transformer architecture have made high impact in the tasks related to machine translation providing support to several languages. The mBART model is trained on the objective based upon "denoising" fashioned autoencoder which targets the language-agnostic representations. On the other hand, mT5 is more robust and has higher capabilities which is based on text-to-text framework, that provides more flexibility to handle various complex tasks of NLP including translation.

This project develops a neural machine translation system for translating text from English-to-Hindi and vice versa using mBART and mT5 models. Further, we emphasize the translation quality using BERT Score, showing semantic similarity and contextual alignment in translation. The study compares the efficiency of these models to feature their strengths and weaknesses, hence giving insight into their applicability for low-resource language pairs.

DATASET

For this project, we used the IIT Bombay English-Hindi Parallel Corpus[1], a freely available dataset, which contains material prepared and gathered for the task of machine translation from and into Hindi. The corpus is and excerpted resource consisting of about 1.49 million parallel segments contributed by several domains like open-source software documentation, religious texts, TED talks, legal judgments, and government websites. Notably, a big portion of these segments 694,000 approximately had not been in the public domain earlier, there by justifying the importance of the corpus in advancing the research in English-Hindi translation[2]. The dataset is highly reputed for its high-quality parallel sentences, making it effective in training and evaluating machine translation models.

Corpus ID	Source	Number of Segments
1	GNOME (OPUS) (Tiedemann, 2012)	145706
2	KDE4 (OPUS)	97227
3	Tanzil (OPUS)	187080
4	Tatoeba (OPUS)	4698
5	OpenSubs2013 (OPUS)	4922
6	HindEnCorp (Bojar et al., 2014b)	273885
7	Hindi-English Linked Wordnets (Bhattacharyya, 2010)	175175

8	Mahashabdakosh: Administrative Domain Dictionary* (Kunchukuttan et al., 2013)	66474
9	Mahashabdakosh: Administrative Domain Examples*	46825
10	Mahashabdakosh: Administrative Domain Definitions*	46523
11	TED talks (Abdelali et al., 2014)	42583
12	Indic Multi-parallel corpus (Alexandra Birch and Post, 2011)	10349
13	Judicial domain corpus - I* (Kunchukuttan et al., 2013)	5007
14	Judicial domain corpus - II* (Kunchukuttan et al., 2012)	3727
15	Indian Government corpora*	123360
16	Wiki Headlines (Provided by CMU)	32863
17	Gyaan-Nidhi Corpus*	227123
	TOTAL	1493527

Table 1: Details of the IITB English-Hindi Parallel Corpus (training set). Indicates new corpora not in the public domain previously

Performed extensive pre-processing on the text to make sure the high consistency in the data and its quality. As part of pre-processing, removed the special characters, unwanted symbols; the entire text was converted to lowercase to maintain uniform format; lastly removed the sentences that had tokens less than three to make sure there are no outliers or if the sentences were incomplete. The mBART model used the mBART50Tokenizer, capable of handling multilingual embeddings, while the mT5 model used the MT5Tokenizer for multilingual text-to-text tasks. These tokenizers helped generate source and target token embeddings that were needed during model training. Also, it added the start (<s>) and end (</s>) tokens, which marked sentence boundaries for better alignment in training.

Language	Train	Test
#Sentences	1,492,827	2,507
#Tokens (eng)	20,667,259	57,803
#Tokens (hin)	22,171,543	63,853
#Types (eng)	250,782	8,957
#Types (hin)	343,601	8,489

Table 2: Statistics of data sets

The data we split into 80% as training set and 20% as testing set, to make sure while developing and testing the model we have enough data for our models. The dataset consists of 101,786 parallel sentences in the training set containing translated sentences pairs and in the similar manner test set used to check model's generalizations on unseen data examples. There were a few inconsistencies observed in translation quality due to colloquial variations and sentence formation differences between the two languages English and

Hindi. It was also observed that there are limited number of examples for complicated sentence structures which impacted on the model's understanding capability of contextual meaning resulting direct impact on the performance.

METHODOLOGY

To implement the project, we followed the methodology involving the implementation of the two transformer-based models mBART and mT5 for English-to-Hindi and Hindi-to-English translations. The mBART[3] and mT5[4] pretrained models were referenced from Hugging Face which is a machine learning and Data Science Community that has open source pretrained machine learning models. These models we furthermore fine-tuned by tweaking their hyper-parameters as both models leverage multilingual embeddings to perform translation tasks of natural language processing.

The mBART model has been trained under denoising objective which later we fine-tuned on the English-to-Hindi data from the dataset. As the pre-trained model, it allowed mBART to learn the language-agnostic representation leveraging state-of-the-art transformer architecture.

In this task, the mBART50Tokenizer has been used to tokenize the input sentences into embeddings that the model can work with. Similarly, the mT5 model also works on a text-to-text framework and converts all tasks into a unified input-output text format. It uses the MT5Tokenizer for encoding, making it compatible with its multilingual architecture.

In training the models, a variety of hyperparameters were selected with care. For the mBART models, the training configurations used a learning rate of 1e-5, batch size of 32, and a linear learning rate scheduler. The gradient clipping was done to keep the training stable using a threshold of 1.0, and an early stopper with five-epoch patience was done to avoid overfitting.

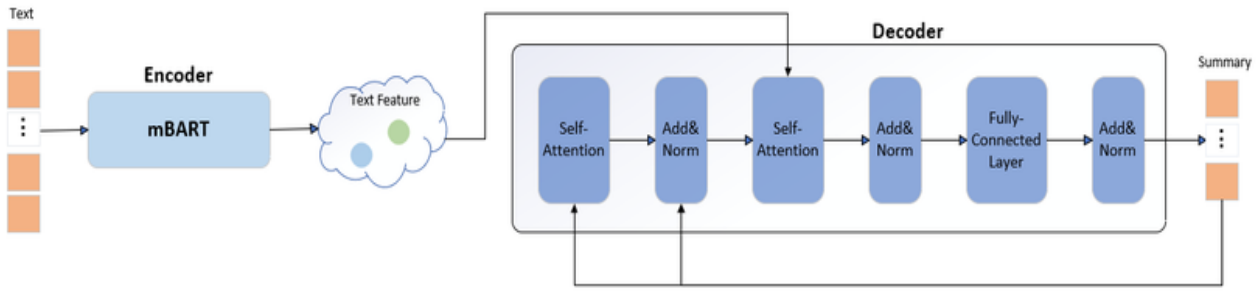

Figure 1: mBART Model Architecture

The models in this study, namely mBART and mT5, were trained on a carefully chosen number of epochs to balance performance optimization against computational efficiency. The fine-tuning of the mBART models was conducted for 3 epochs; this proved sufficient in achieving high translation quality. This relatively small number of epochs was selected because the pre-training on multilingual data for mBART had already been

done, so only a little fine-tuning was required to get it working on the English-Hindi translation task. Additionally, early stopping with a patience of five epochs was implemented, which helped model to prevent from overfitting and saving computational resources as it stops training process if there are no improvements are observed in consecutive evaluation steps.

By contrast, the mT5 model was fine-tuned for 5 epochs since it had to learn more task-specifically and align its pre-trained general-purpose representations better with the translation requirements. The extra number of epochs allowed the model to refine the outputs of the decoder while profiting from the frozen encoder layers that retained the pre-trained multilingual embeddings. This approach helped the mT5 model adapt better to the complexities of English-Hindi translation. In both models, the chosen number of epochs ensured the efficiency of training, retaining strong performance metrics.

The fine-tuning for the mT5 model involved the learning rate of 3e-6, a batch size of 16, and cosine learning rate schedule with warmup ratio of 20%. In this experiment, both models are trained using AdamW optimizer with mixed precision to speed up computations in FP16 format.

While fine-tuning the mT5 model, some of its layers were kept frozen to enhance computational efficiency and focus only on task-specific learning. Being pre-trained on a large multilingual corpus, the mT5 model already captures robust language representations in its encoder layers. By freezing these layers, we ensured that the model retained its pre-trained linguistic knowledge and directed its learning capacity toward the task-specific adjustments needed in the later layers, particularly the decoder. This approach allowed the model to adapt its output generation to align with the nuances of English-Hindi translation, without the need to relearn basic language representations. Freezing layers reduced the risk of overfitting of the model and increased its generalization to new, unseen data.

Freezing the layers had one more very useful side effect it drastically reduced computational cost. The mT5 model has an enormous number of parameters, meaning that its fine-tuning required a substantial amount of hardware resources. By freezing the encoder layers, memory consumption dropped down considerably, which correspondingly decreased training time and gave significant boosts in efficiency on weaker hardware. This choice was a balanced trade-off it leveraged the pre-trained strengths of mT5 but put its capacity to focus on decoding task-specific outputs. In this way, fine-tuning remained computationally feasible.

To optimize performance and resource utilization, the training process was performed with the BF16 precision format as part of the hyperparameter configuration. BF16 is a numerical format widely used in modern machine learning frameworks to improve the efficiency of deep learning training while maintaining computational accuracy. It reduces memory usage and accelerates computations compared to full 32-bit floating-point (FP32) precision.

The hyperparameter configuration uses BF16 precision to balance efficiency and resource utilization during model training. By using less memory, BF16 enables the effective training of large models such as mBART and mT5, which use a lot of memory for their multilingual embeddings and large parameter sizes. It accelerates computations on hardware that supports BF16 operations, such as TPUs and NVIDIA Ampere GPUs, thereby reducing overall training time without sacrificing model performance. Unlike FP16, which can be prone to certain problems like gradient underflow or overflow due to limited range, BF16 maintains the same exponent range as FP32 and hence is numerically stable during training. Besides, the capability

of handling extensive matrix operations by BF16 most makes it more compatible with transformer-based models and large datasets like that of the IIT Bombay English-Hindi Parallel Corpus. Among all the precisions, choosing this format was critical to have a balance of performance, stability, and computational efficiency.

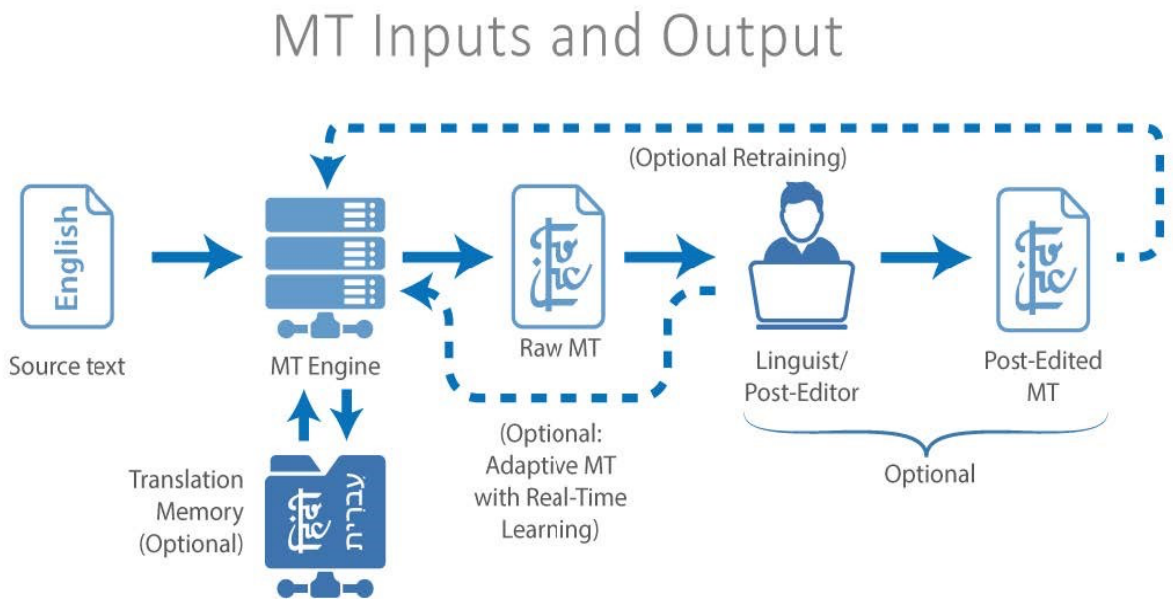

Figure 2: mT5 Model Architecture

Effect on Model Performance of BF16 was instrumental in efficiently training the models while maintaining a performance comparable to FP32 precision. By employing BF16, the training was optimized both in terms of speed and memory for fine-tuning mBART and mT5 on resource-constrained hardware without sacrificing quality in the resulting translations. This choice was critical in handling the complexity of translation tasks and the large-scale parallel corpus effectively.

The overall workflow was initiated with data preprocessing, which cleaned the raw sentences, tokenized them, and added special tokens. Then, the preprocessed data was divided into training and testing sets, which were used to feed the respective models. After training, the models were evaluated by BERT Score, in which the contextual alignment between translations and references was measured. This evaluation showed the strengths and limitations of each model in preserving semantic and contextual accuracy.

RESULTS

The evaluation of our neural machine translation models, mBART and mT5, yielded some interesting results about their effectiveness and shortcomings while performing the tasks of English-to-Hindi and Hindi-to-English translation. Performance metrics were measured with the inclusion of BERT Score for semantic similarity and contextual alignment, and qualitative analysis pointed out strengths and areas of improvement.

Model	Precision (%)	Recall (%)	F1 Score (%)
mBART (English-to-Hindi)	95.5	92	93.7
mBART (Hindi-to-English)	98.4	98.9	98.7
mT5 (English-to-Hindi)	77.3	77.5	77.2

Table 3: Models Evaluation based on BERT Score

The evaluation results reflect that mBART worked amazingly in both English-to-Hindi and Hindi-to-English machine translations, which gave excellent precision, recall, and F1 score values. It effectively generated grammatically correct, semantically rich translations with preservation of contextual nuances, especially in Hindi-to-English, whose near-perfect scores speak to the solidity of multilingual embeddings and the fine-tuned decoder. On the other hand, the mT5 model, though universal for general text-to-text tasks, yielded lower scores due to its broader pre-training objectives. While mT5 was performing well for simple sentences, it produced poor results on complex syntactic and semantic structures, which implies that more task-oriented fine-tuning is needed.

Qualitative Analysis

Besides numerical measures, sample translations were subjected to some qualitative evaluations; mBART has shown high linguistic precision as it handles idiomatic expressions well in maintaining the semantic integrity of the source.

mT5 was versatile enough for text-to-text, but sometimes did not yield the intended meaning of idiomatic and culturally specific phrases. For example, sentences with nested clauses or rare vocabulary often remained incomplete.

After performing comparison across models, the better performance of mBART across both translation directions underlines the advantage of its pre-training being explicitly for translation tasks. In contrast, mT5, with its flexibility, revealed limitations given its general-purpose pre-training. This comparative analysis underlines the importance of task-specific pre-training and fine-tuning for achieving optimal results in machine translation.

Sample outputs

To demonstrate the model's performance and working; there are sample outputs attached further for each developed model for English-to-Hindi and Hindi-to-English translations. This helps to clearly see how the models understands and how translation capabilities.

```
Vamshi's English-to-Hindi Model Evaluation:
Input Sentence 1: How are you?
Translation: आप कैसे हैं?
Reference: आप कैसे हैं?
Input Sentence 2: This is a test of the translation model.
Translation: यह अनुवाद मॉडल का परीक्षण है
Reference: यह अनुवाद मॉडल का परीक्षण है।
Input Sentence 3: The weather today is quite pleasant.
Translation: आज मौसम काफी सुखद है
Reference: आज का मौसम काफी सुहावना है।
Input Sentence 4: Could you please help me?
Translation: क्या आप मेरी मदद कर सकते हैं?
Reference: क्या आप कृपया मेरी मदद कर सकते हैं?
Input Sentence 5: Machine learning is transforming many industries.
Translation: मशीन शिक्षण कई उद्योगों को बदल रहा है
Reference: मशीन लर्निंग कई उद्योगों को बदल रही है।
```

Figure 3: Output of mBART English-to-Hindi Translation Model

```
Akshay's Hindi-to-English Model Evaluation:
Input Sentence 1: आपका नाम क्या है?
Translation: what is your name?
Reference: What is your name?
Input Sentence 2: आज का मौसम कैसा है?
Translation: what is the weather like today?
Reference: How is the weather today?
Input Sentence 3: यह तकनीक बहुत प्रभावी है।
Translation: this technique is very effective.
Reference: This technology is very effective.
Input Sentence 4: मुझे आपकी सहायता चाहिए।
Translation: i need your help.
Reference: I need your assistance.
Input Sentence 5: कृपया मुझे रास्ता दिखाइए।
Translation: please show me the way.
Reference: Please show me the way.
```

Figure 4: Output of mBART Hindi-to-English Translation Model

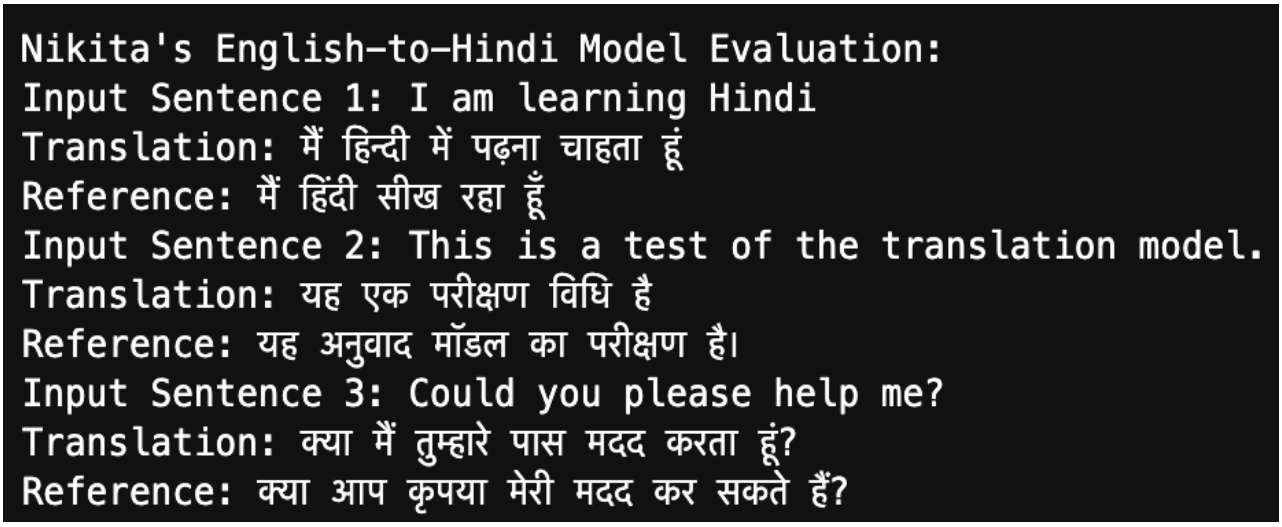

Figure 5: Output of mT5 English-to-Hindi Translation Model

DISCUSSION

The results from our experiments highlight the capabilities and challenges that lie with neural machine translation models while dealing with English-Hindi and Hindi-English translations. The scores obtained for mBART and mT5 models both look very promising, showing their strengths in multilingual embeddings for linguistic features. However, their behaviors were different based on sentence structural complexity, contextual preservations, and quality of data.

Model Performance Insights

After observing the BERT Score for both the models the observation was recorded as mBART model performed well compared to mT5 considering precision, recall and F1 score provided by BERT Score for the translations. The reason behind these results might be as an encode-decoder architecture of mBART; as it was specifically trained for machine translation tasks enabling it to capture deeper contextual alignments, contextual meanings and the grammatical sense effectively. On the contrary the mT5 is more versatile for text-to-text based complicated tasks but it struggled to achieve good results due to broader pre-training objective that has been fine-tuned for complex NLP tasks and not just specialized for translations. The poor performance of mT5 model in terms of accuracy clearly suggests that it could not fully capture the language specific nuances as effectively as mBART.

Challenges and Observations

Despite their strengths, both models faced certain challenges with respect to translation of the following aspects:

High Validation loss and Poor Generalization[5][6]: During training, both models showed relatively high validation losses, which may indicate the overfitting difficulties in generalizing to unseen data.

Handling Diverse Tokenization: the fact that Hindi and English words are tokenized differently imposed problems that prevented the generation of embeddings the meaning of the input during both training and evaluation.

Contextual Errors: Both models have at one time or another generated translation that did not catch the semantic intent of the source text, particularly those with idiomatic expressions and culturally specific terms. The problem also speaks to the challenge of achieving semantic equivalence across languages with different grammatical structures and cultural contexts.

Sentence Handling: Sentences with a complex structure, such as an embedded clause or long dependencies, were difficult for both models and resulted in incomplete or grammatically incorrect translations. These may be related to the relatively scarce representation of such sentences in the dataset.

Dataset Quality and Noise: Although the IIT Bombay Parallel Corpus is a high-quality resource, variations in translation style and the occasional inconsistency have introduced noise into the training process. For example, colloquial variation and mismatched sentence boundary have affected model performance, especially for mT5.

Evaluation Metrics: Although BERT Score was chosen based on the reasons for the assessment of semantic similarity and contextual alignment, it doesn't really capture grammatical errors or stylistic deviations in translations. Not having BLEU among the evaluation metrics-though, admittedly, for a good reason-creates some limits of comparability with existing studies in the field.

Results Significance

The high performance of mBART signifies that it is more appropriate for a dedicated machine translation task, especially in situations where a high degree of contextual preservation and grammatical accuracy is required. On the other hand, the versatility of mT5 indicates potential utility for tasks beyond translation, such as summarization or question answering, although it needs further fine-tuning to enhance its translation performance.

Future Research Work

Considering the observed challenges and in a quest to enhance performance, we recommend the following:

Improved Dataset Curation: More complex sentence structures and diverse linguistic expressions might help in enhancing their robustness.

Transfer Learning and Domain Adaptation: Fine-tuning these models on domain-specific subsets, such as legal or medical translations, may allow better performance for specialized applications.

Advanced Evaluation Metrics: BERT Score could be supplemented by grammar-focused metrics or human evaluations to provide more comprehensive translation quality assessment.

Hybrid Models: Bringing together the powers of mBART and mT5 into one hybrid model could extract the contextual precision of mBART with the flexibility of mT5.

Overall, the present study showcases the potential for neural machine translation models while also pointing toward areas needing further research and development beyond existing limitations in performing English-to-Hindi translation tasks.

CONCLUSION

This projects mainly focuses on the use of mBART and mT5 neural machine translation models, applied to the English-Hindi and Hindi-English translation tasks, respectively. The models have been implemented using the IIT Bombay English-Hindi Parallel Corpus. The results proved that, in both translation directions, the mBART model performed better than mT5. The mBART model was task-specifically pre-trained and fine-tuned to output translations that were grammatically correct and semantically packed, with preservation of nuance and better handling of linguistic challenges.

In contrast, the mT5 model, while superior for a wide variety of text-to-text applications, had relatively lower scores on especially complex sentences. Though mT5 gave the best results in simple translations, its general training made it less capable of grasping language-specific nuances as effectively as mBART did. These findings emphasize the importance of task-specific pre-training and high-quality datasets for good results in translation.

Despite the promising result, the following are those areas which must be improved, including handling syntactically complicated structures, idiomatic expressions, and inconsistencies of the dataset. BERT Score used as the major evaluation metric shed light on semantic similarity and contextual alignment but demonstrated that some complementary ways of evaluation must be created to fully capture grammatical and stylistic accuracy.

Overall, this work shows the possibility of neural machine translation models, especially mBART, in enhancing the works of English-Hindi translation. Future work may develop better datasets, use newer evaluation metrics, and explore hybrid architectures to overcome the noticed limitations and further improve translation quality.

References

[1] IIT Bombay English-Hindi Parallel Corpus: https://www.cfilt.iitb.ac.in/iitb_parallel/
[2] https://www.cse.iitb.ac.in/~pb/papers/lrec18-iitbparallel.pdf
[3] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
[4] https://huggingface.co/google/mt5-small
[5] https://discuss.huggingface.co/t/t5-variants-return-training-loss-0-and-validation-loss-nan-while-fine-tuning/30839
[6] https://discuss.pytorch.org/t/training-loss-0-0-validation-loss-nan/168467