

NLP for Health-care Data

Daniel Persson *

March 2024

Abstract

In this lab, I dove into the world of advanced natural language processing (NLP), focusing on the Flan-T5 model for its ability to handle a variety of text processing challenges. My main focus was the MIMIC-IV dataset, aiming to extract patient diagnoses from clinical notes. This task showed me just how much potential NLP has in revolutionizing healthcare data handling. The lab was structured around a few key skills in NLP: getting to grips with pre-trained Large Language Models (LLMs), getting creative with prompt-based In-Context Learning (ICL) for directing the model on specific tasks, fine-tuning LLMs for targeted uses, and getting my hands dirty with Fine-tuning all weights as well as Parameter-Efficient Fine-Tuning (PEFT) methods. By applying what I learned to the MIMIC-IV dataset, I managed to fine-tune the Flan-T5 model, significantly boosting its performance in diagnosing extraction. I also compared the traditional fine-tuning approach to the newer PEFT Lora method. Through my experiments and comparing ROUGE scores, I highlighted the benefits of thorough model training and the strategic use of LoRA for improving performance on specialized tasks. This lab didn't just deepen my understanding of both the basics and more complex parts of NLP; it also showed me how these methods could come together to tackle real-world problems, opening doors to future projects in this field.

*daniel.p-00@hotmail.com

Contents

1	Introduction	3
2	Part 1 (Pre-process clinical text)	3
3	Part 2 (ICL for extracting the diagnosis)	4
4	Part 3 (Fine-tuning the Flan-T5 base)	4
5	Part 4 (LORA PEFT on the Flan-T5)	6
6	Conclusion	8

1 Introduction


In this laboratory session, we confront several prevalent challenges in the field of advanced natural language processing (NLP) by focusing on Flan-T5, a model distinguished by its robust performance across a wide array of text processing tasks. Our primary case study involves the MIMIC-IV dataset, a comprehensive collection of clinical notes, from which we aim to accurately extract patient diagnoses. This endeavor not only showcases Flan-T5's adeptness at handling complex information but also underscores the potential of NLP to revolutionize data processing in healthcare.

The goals for this lab are designed to equip participants with a solid foundation in several key areas of NLP:

- **Engaging with Pre-trained Large Language Models (LLMs):** I will learn to understand how to utilize their extensive pre-trained knowledge.
- **Implementing Prompt-based In-Context Learning (ICL):** I will get introduced to ICL, how to design effective prompts that guide the model in performing specific tasks.
- **Fine-tuning LLMs for Targeted Applications:** I will delve into the fine-tuning process, demonstrating how to specialize LLMs for particular tasks.
- **Exploring Parameter-Efficient Fine-Tuning (PEFT) Techniques:** I will also be introduced to PEFT strategies, highlighting how they allow for the efficient adaptation of LLMs to new tasks with minimal modifications.

2 Part 1 (Pre-process clinical text)

In the first part, the task was to exclude all non-formative tokens from the text column of the data. This part was very straightforward, and I executed this relatively fast and easy, please see Figures 1 and 2 for a screenshot of the same text with, and without unnecessary tokens, respectively.



7 Admission Date: [**2112-12-8**] Discharge Date...

Figure 1: An example of row 7, column "text", before the removal of unnecessary tokens.

7 admission date discharge date service medic...

Figure 2: An example of row 7, column "text", after the removal of unnecessary tokens.

0-shot result: 76yearold female with metastatic ovarian carcinoma who underwent total abdominal hysterectomy omentectomy sigmoid resection with end colostomy bleeding duodenal ulcer and fungal duodenal mass status post exploratory laparotomy and oversewing of duodenal ulcer and pyloroplasty status post pleurodesis times two right groin pseudoaneurysm status post injection of the ombin
1-shot result: adolescent ovarian carcinoma : a case report from a u.s. hospital
2-shot result: radionecrosis ct head postop findings there is a leftsided frontal craniotomy with surgical resection of the enhancing mass noted in previous studies there is extensive confluent hypodensity in the left frontoparietal region that likely represents postradiation vasogenic edema with resultant 85mm shift of normallymidline structures not significantly changed from the previous mr there is a tiny focus of hyperdensity in the postoperative field which may represent a small amount of hemorrhage there is expostoperative pneumocephalus and subcutaneous air noted impression status post left craniotomy and resection of av
3-shot result: radionecrosis ct head postop findings there is a leftsided frontal craniotomy with surgical resection of the enhancing mass noted in previous studies there is extensive confluent hypodensity in the left frontoparietal region that likely represents postradiation vasogenic edema with resultant 85mm shift of normallymidline structures not significantly changed from the previous mr there is a tiny focus of hyperdensity in the postoperative field which may represent a small amount of hemorrhage there is expostoperative pneumocephalus and subcutaneous air noted impression status post left craniotomy and resection of av
4-shot result: radionecrosis ct head postop findings there is a leftsided frontal craniotomy with surgical resection of the enhancing mass noted in previous studies there is extensive confluent hypodensity in the left frontoparietal region that likely represents postradiation vasogenic edema with resultant 85mm shift of normallymidline structures not significantly changed from the previous mr there is a tiny focus of hyperdensity in the postoperative field which may represent a small amount of hemorrhage there is expostoperative pneumocephalus and subcutaneous air noted impression status post left craniotomy and resection of av

Figure 3: 1 out of the 5 examples about the ICL

3 Part 2 (ICL for extracting the diagnosis)

As for task 2, an appropriate prompt was created to extract diagnoses from clinical text using something called in-context-learning (ICL). Which is when you provide a text instructing the model what to do, as well as examples. I chose to do 5-shot ICL, which is when you do each prediction individually with 0,1,2,3, and 4 examples shown. Again, for visualization purposes, only 1 example is shown in the report, please watch the notebook for further examples. Nevertheless, by inspecting Figure 3, this example is made with a maximum length of 2000 input tokens, which yielded the result in the image, where the predictions differ the more examples the model is given. Furthermore, when playing around with the maximum length of the input tokens, I found out that depending on this number, the predictions will vary a lot, especially in how many examples are needed. When using a low value of about 512, the predictions were the same for the 1-4 examples given to the model, while for a value of 2000, the predictions were only the same for the 3-4 examples given, as seen in Figure 3.

4 Part 3 (Fine-tuning the Flan-T5 base)

For this part, the Flan-T5 was retrained on our personal training dataset in order to relearn all the models weights so that they reflect this task more. In order to analyse and see if the training actually had any results at all, a list of the column "labels" for the first 50 rows was converted in order to create the "ground truth". Then the model was utilized to do predictions of these labels based on an input text. These 50 predictions and their corresponding labels are

printed out in a cell within the notebook as ordered by the teacher, and i chose not to include them all here due to the limitation of the report.

At first i trained the model with my own implementation, and not the trainer library, which made the training take a lot of time, namely about 45 min per epoch. However i later switched to using the trainer due to it only taking about 10-15 min per epoch, which helped a lot during the laboration.

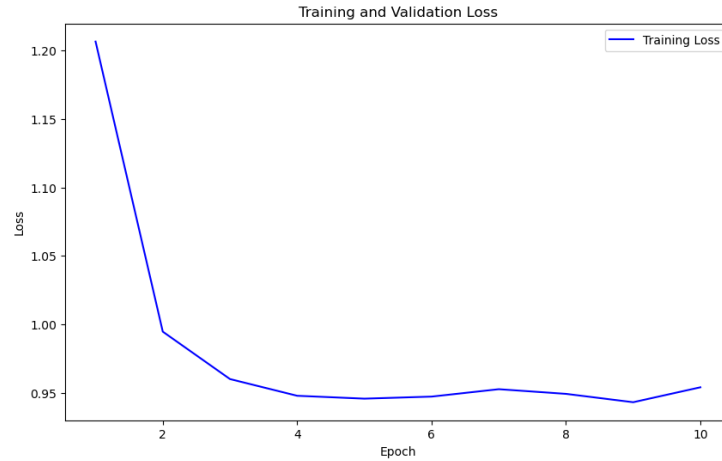


Figure 4: A plot representing the training losses over 10 epochs for task 3.

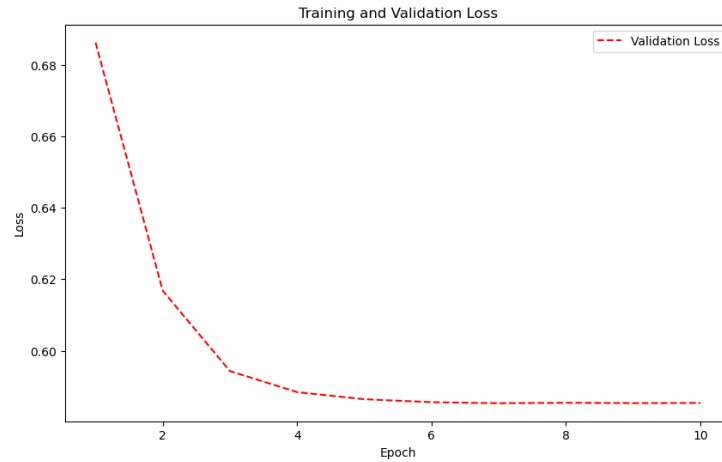


Figure 5: A plot representing the Validation losses over 10 epochs for task 3.

The training of all weights on the Flan-T5 model can be seen in Figures 4 and 5, where the losses of the training as well as validation is plotted for every epoch (10). it is clearly seen that both the training loss as well as the validation

loss is converging very nicely with no overfitting. I also printed the validation loss of the model after the random weights were assigned, which was *43.4698*, which is a lot higher than the first validation loss of *1.20* in Figure 5, resulting in positive training.

In order to evaluate the model, a metric called rouge was calculated. And in Table 1 you can see the original rouge score without any training, the score when all weights were trained(task 3), and at last the score where PEFT lora was used (task 4).

Diverse models Rouge scores				
Model	rouge1	rouge2	rougeL	rougeLsum
Original	0.0323	0.0059	0.0296	0.0301
Fine-Tuned	0.1259	0.0126	0.0957	0.0957
PEFT-Lora	0.100	0.020	0.080	0.085

Table 1: ROUGE scores for different models.

5 Part 4 (LORA PEFT on the Flan-T5)

When Fine-Tuning the Flan-T5 model using PEFT LORA, i chose to go with the following parameters seen in Table 2, where the name, value, and explanaiton is given.

LoRA Parameter List		
Parameter	Value	Description
r	8	Specifies complexity
lora_alpha	16	Scaling factor
target_modules	['q', 'v']	Applied to Query & Value
lora_dropout	0.01	Dropout rate for the layers
bias	None	No adjusted/added bias
task_type	SEQ_2_SEQ_LM	Specifying the model's task

Table 2: Overview of LoRA Parameters.

By inspecting the Table with all Rouge scores, namely Table 1, we can see that the LORA also improved the original models predictions etc for this specific data task. For this task as well as task 3, the same method was used in order to convert the first 50 rows of the test dataset's column "labels" into labels, as well as making predictions on the first 50 input texts and comparing against the labels. Just as for task 3, these predictions and their corresponding labels are printed out in the notebook.

When training the model, i printed out the first evaluation loss with the newly randomized weights with no training at all, and i got the same value, *43.4698*, as i got for task 3. This is probably because when initializing the

random weights, the same random seed is being used and therefore the randomly initialized weights in the beginning for both models in task 3 and 4 are the same. So they have the same "starting point".

In addition, by inspecting Figures 6 and 7, we see again that the training losses as well as the validation losses, respectively, steadily converge at around the same time without any real overfitting.

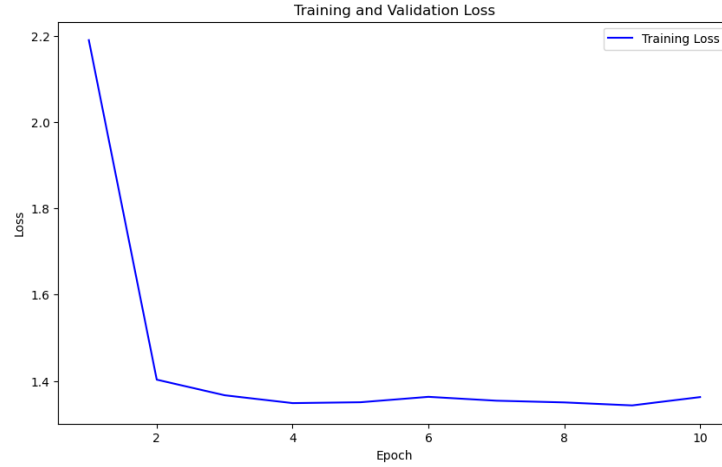


Figure 6: A plot representing the training losses over 10 epochs for task 4.

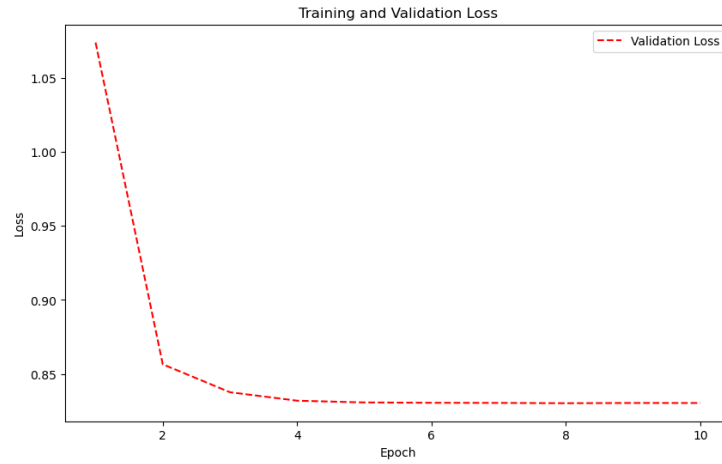


Figure 7: A plot representing the Validation losses over 10 epochs for task 4.

6 Conclusion

Regarding the conclusions for this laboration, the instructions were very clear and the notebook as well, giving a good red thread to follow. I feel like the lab was not too hard, but not too easy either. I think it was just enough of both. In addition I also think that the tasks we were given did their job, by introducing us to diverse fragments within Natural language processing, and I think that I learned a lot from this lab about NLP models actually.

As for the technical outcomes of this lab, I feel like the models adapted pretty well and converged, both for the model in task 3 but also for the model in task 4 even though it converged at a higher rate and gave a lower Rouge score. These results are probably because when training all the weights, we make task-specific adaptations to all the different levels of the Flan-T5's architecture, resulting in a better adaptation than when using Lora, which focuses on adapting only a small subset of the levels.