TITLE
...
...

Carmen Armenti

October 2022

Supervised by
Prof. Rocco Oliveto

Co-Supervised by

Contents

1 Introduction **1**
1.1 Application context . . . 1
1.2 Motivations and Objectives . . . 1
1.3 Results . . . 1
1.4 Document Structure . . . 1

2 Deep Learning and Curriculum Learning **3**
2.1 State of the art . . . 3
2.2 Curriculum Learning related works . . . 3
2.3 Curriculum Learning application in Deep Learning tasks . . . 3

3 Deep Learning Applications to Software Engineering tasks **5**
3.1 State of the art . . . 5
3.2 Software Engineering related works . . . 5
3.2.1 Bug fixing task . . . 5
3.2.2 Code summarization task . . . 5
3.2.3 Log generation task . . . 5
3.3 . . . 5

4 Evaluating Curriculum Learning in Software Engineering tasks **7**
4.1 Neural Machine Translation . . . 9
4.2 Canonical Training . . . 10
4.2.1 Bug-fixing . . . 10
4.2.2 Code summarization . . . 10
4.2.3 Log generation . . . 11
4.3 Training with Curriculum Learning . . . 11
4.3.1 Bug fixing task . . . 11
4.3.2 Code summarization task . . . 12
4.3.3 Log generation task . . . 13

5 Analysis of results **15**
5.1 Bug fixing task . . . 15
5.2 Code summarization task . . . 16

5.3 Log generation task 17

6 Conclusion **19**

A Appendix **21**

List of Figures

4.1 Predefined Curriculum Design. 7
4.2 Code summarization task: data distribution. 13

List of Tables

5.1 Models' performances 15
5.2 Overlapping metrics: baseline and curriculum learning models. 16
5.3 BLEU values summary 16
5.4 Perfect predictions summary. 17

Chapter 1

Introduction

1.1 Application context

1.2 Motivations and Objectives

1.3 Results

1.4 Document Structure

This document is organized as follows:

- In Chapter 2,
- In Chapter 3,
- In Chapter 4,
- In Chapter 5,

Chapter 2

Deep Learning and Curriculum Learning

Inspired by human learning, curriculum learning is an algorithm that emphasizes the order of training instances in a computational learning setup. As a feature of human learning, curriculum, or even better learning in a meaningful way, has been transferred to machine learning, thus creating the subdiscipline named *curriculum learning* [7].
Essentially, human education is organized as curricula, by starting small indeed, and gradually presenting more complex concepts. The paramount hypothesis is that simpler instances should be learned during the first steps as building blocks to then learn more complex ones. Several experiments on sentiment analysis task and tasks similar to sequence prediction tasks in NLP carried on by Cirick *et al* [1] prove that curriculum learning has positive effects on LSTM's internal states, by biasing the model through building constructive representations. Specifically, the internal representation at the previous timestep is used as building block for the next one, thus contributing at the final prediction.

2.1 State of the art

2.2 Curriculum Learning related works

2.3 Curriculum Learning application in Deep Learning tasks

Chapter 3

Deep Learning Applications to Software Engineering tasks

3.1 State of the art

3.2 Software Engineering related works

3.2.1 Bug fixing task

Each instance of the dataset is a pair (m_b, m_f), where m_b is a buggy code component and m_f is the corresponding fixed code. These BFPs were used to train the NMT model, allowing it to learn the translation from the buggy to the fixed method, thus being able to generate fixing patches.

3.2.2 Code summarization task

ù Automatic source code summarization is the task of generating short natural language description for source code [3]. The idea is that a brief description allows programmers to understand what a chunk of code does and what is the purpose of the program by and large, without necessarily read the code itself.

3.2.3 Log generation task

3.3

Chapter 4

Evaluating Curriculum Learning in Software Engineering tasks

In a nutshell, curriculum learning means "training from easier data to harder data" [7]. More specifically the core idea is to "start small" [2], train the machine learning model with easier subtasks, to then gradually increase the difficulty level of subtasks until the whole training dataset is used.
Bearing in mind the strategy of training from easier to harder data, to design such a curriculum idea (i), what kind of training data is supposed to be easier than other data, (ii) and when is appropriate to present more harder data for training - and how much more - must necessarily be decided. Technically, those 2 issues can be abstracted in the concepts of a Difficulty Measurer, that decides the "easiness" of each data instance to start the training process from, and a Training Scheduler, that rules the sequence of data subsets during the whole training process [7]. Therefore, Difficulty Measurer together with Training Scheduler constitute a general framework for curriculum design, as illustrated in Figure 4.1. First of all, the Difficulty

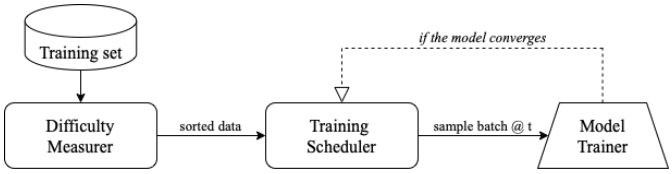

FIGURE 4.1: Predefined Curriculum Design.

Measurer sorts all the training examples from the easiest to the hardest and passes them to the Training Scheduler. Then, at each training epoch t, the Training Scheduler samples a batch of training instances from the easier subset and gives them to the Model Trainer for training.
As training epochs increase, the Scheduler decide when to sample from more harder data, generally until uniform sampling from the whole training set. This schedule either depends on the training loss feedback from the Model Trainer, or on some other parameters that implies that the model would deverge, if left to training for more epochs.
A distinction between **predefined CL** and **automatic CL** must be clarified. The first refers to the

framework where both the Difficulty Measurer and Training Scheduler are defined by human prior knowledge, thus with no data-driven algorithms involved; the latter instead, if any - or both - of the components are designed by data-driven algorithms.
Usually, the power of introducting Curriculum into Machine Learning depends on how the curriculum for specific applications and dataset is designed. Due to this, Difficulty Measurers often relie on the data characteristics of specific tasks, and most of them are defined by complexity, diversity, or noise estimation definitions.

We adopted the most popular discrete scheduler, known as *Baby Step*, where the complexity of the training data needs to be gradually increasing. Due to this, a reliable metric to split the initial dataset of each task needed to be defined. This approach distributes the sorted data into buckets, from easy to hard according to the metric, and starts training with the easiest bucket. Starting from the easiest bucket, after a fixed number of training epochs or convergence, the subsequent bucket is merged into the current training subset - main characteristic of *Baby Step* approach. Finally, once all the buckets are merged and used, the training process either stops or continues several extra epochs.
Note that, at each epoch the scheduler shuffles the current bucket and samples mini-batches for training.

However, before testing curriculum learning approach, we thougth it was strictly necessary reproducing - in the case of bug-fixing task - or conducting - within the framework of the other 2 tasks - the experiment on the modelbase.
We worked with a neural network whose configuration is composed by 1-layer bidirectional Encoder, 2-layer Attention Decoder both with 256 units, embedding size of 512, and LSTM RNN cells.

In this chapter, we present how we applied Curriculum Learning to some of Software Engineering tasks, i.e. *bug-fixing*, *code summarization*, and *log generation* tasks.
For each of the tasks considered, the following aspects are described thoroughly:

- **Difficulty Measurer**: a measure to sort the datasets' instances is needed, each task was experimented with a different metric;
- **Training Scheduler**: sequence of data subsets are presented to the model following a training schedule.

The goal of this study is to asses whether Neural Machine Translation, combined with the Curriculum Learning approach, can be used for the tasks experimented. In the following section, the design of our study is described in detail.

4.1 Neural Machine Translation

The experimented models are based on an Recurrent Neural Network (RNN) Encoder-Decoder architecture with attention mechanism, frequently used in Neural Machine Translation. This kind of model is composed by two dominant components:

- a RNN Encoder, which encodes a sequence of terms **x** into a vector representation;
- a RNN Decoder, which decodes the vector representation into another sequence of terms **y**.

The model learning is based on a conditional distribution, where the output sequence of terms is conditioned by the input sequence: $P(y_1, ..., y_m|x_1, ..., x_n)$, where m and n not necessarly have to have the same length. The Encoder takes as input a sequence $\mathbf{x}= (x_1, ..., x_n)$ and produces a sequence of states $\mathbf{h}= (h_1, ..., h_n)$. The framework relies on a bi-directional RNN Encoder, which is composed by a backward and a forward RNN, where both are able to create representations taking into account past and future inputs. Specifically, each state h_i is the concatenation of the states produced by the two RNNs when reading the sequence not only in a forward but also in a backward manner.
The RNN Decoder computes the probability of a target sequence $\mathbf{y}= (y_1, ..., y_n)$ given **h**. The probability of each output term h_i is computed based on:

- the recurrent state s_i in the Decoder;
- the previous $i - 1$ terms $(y_1, ..., y_i - 1)$;
- a context vector c_i, which constitutes the attention mechanism.

The vector c_i is a weighted average of the states in **h**, where the weights associated to each state allow the model to pay more attention to some parts of the input sequence than to others:

$$c_i = \sum_{t=1}^{n} a_{it} h_t$$

Precisely, the weigth a_{it} defines how much the model should take into consideration the term of the sequence in input x_i when predicting the target term y_t. Encoder and Decoder are simultaneously trained - instead of sequentially - by minimizing the negative log likelihood of the target terms, using stochastic gradient descent. The configuration used by the neural network is composed by 1-layer bidirectional Encoder, 2-layer Attention Decoder both with 256 units, embedding size of 512, and LSTM RNN cells. Bucketing and padding was used to deal with the variable length of the sequences.

4.2 Canonical Training

Before experiment the curriculum learning approach, we reproduced the baseline approaches for each task experimented that include the conventional technique of randomly ordering the training samples with the aim to investigate how the performance of the original models are, compared to the experiment affected by curriculum learning.

4.2.1 Bug-fixing

As for the bug-fixing task, the datasets used to train the NMT model is the union between *small* and *medium* method-level datasets used by Tufano *et al.* [6]. So, we worked with the standard training, evaluation and test set splits of respectively 99.044, 12.381, and 12.380 instances.
It is vital to use new data when evaluating a model to prevent the likelihood of overfitting to the training set. However, we decided to evaluate our model as we were building it to find the best parameters of a model. To evaluate the model while still building and tuning the model, we used the evaluation set. The neural network was set up to evaluate the model every 1.000 steps. The training was performed for 60k steps as upper bound, out of which we considered the best model, early stopping the model before overfitting considering the loss values computed every 1.000 steps. The best model configuration was then used to run the inference. Indeed, after the model was trained, it was evaluated on the test set of unseen buggy code. Instead of using the classic greeding decoding that selects the output term y_i with the highest probability, however, we used another decoding strategy known as Beam Search. The key idea is that the decoding process keeps track of k hypotheses - being k the beam size or width.
The beam search algorithm selects multiple tokens for a position in a given sequence based on conditional probability. Moreover, the algorithm can take any number of N best alternatives through a hyperparameter known as beam size or width indeed. Conversely to what greedy search does, Beam search broad the search to include other words that might fit better apart from the best word for each position in the sequence. If Greedy search looks at each position in the output sequence in isolation, deciding the word based on highest probability and then moving down to the resto of the sentence, Beam search also takes the N best output sequences and look at the current preceding words and probabilities compared to the current position that is being decoded in the sequence.
We considered the following sizes: 1 - that corresponds to the Greedy Naïve Approach -, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50.

4.2.2 Code summarization

Regarding the task of code summarization, we selected a random sample of 200k instances from a dataset of 2.1 million items. Also here we considered the canonical division between training, test and evaluation of respectively 200.000, 105.832, and 106.153 instances. We trained

the model on the training set for 20 epochs, that corresponds to 125.000 steps, and we took the best model over the 20 model configuration evaluated after every epoch, thus using the best checkpoint to run the inference step. This time, to guide the selection of the best configuration, we used the BLEU values computed on the validation set instead of using the loss function. Conversely, the results are computed on the test set. Indeed, the same way it was done for the previous task, after the training step the model is evaluated over the test set of unseen functions to generate code summarizations of such methods.
The scope of the training here was to make the model able to learn how to summarize pieces of code, starting from a sequence of tokenized methods.

4.2.3 Log generation

As well as happened in both the previous task experimented, also in log generation we considered the well-known split between training set, test set, and evaluation set. The sets in question are - following the same order - of the amount of 106.382, 12.020, and 13.260. The model was trained for 45 epochs, i.e. 149.625 steps, and it must be specified that it was evaluated every epoch. The best configuration in this case was guided on an early-stopping criterion based on BLEU values. Once the model was trained, it was evaluated against the test set of unseen methods without log statements and messages in order to assess its ability to correctly inject log statements and messages within methods.

4.3 Training with Curriculum Learning

As stated in the Curriculum Learning literature, the training instances need to be ordered based on the curriculum. Each of the task considered is related to a different curriculum, specifically the datasets were ordered following 3 different complexity ideas. Since a curriculum approach requires the definition of a **Difficulty Measurer** and a **Training Scheduler**, we defined those for each task. In the following sections details are explained.

4.3.1 Bug fixing task

Based on the entire training dataset at our dispose of couple of buggy-fixed methods, we defined a Difficulty Measurer to divide the dataset and a Training Scheduler to rule the order in which the instances were given to the model. Since both the source and the target were tokenized, we thought that computing the number of changes from the buggy version to the fixed one of each method was a quite reliable way to define the complexity for the type of instances used in this task.

Difficulty Measurer. As stated above, a difficulty measurer definition is the core element of the approch implemented. For the task at issue here the **Levenshtain distance** was used as metric. The *Levenshtein distance*, also known as the *edit distance*, was introduced by Vladimir Levenshtein in 1965 [4]. It is a string metric for measuring difference between two sequences, specifically the number of insertions, deletions, and substitutions required to transform one string into the other. However, the distance we used for our experiments was token based, thus the granularity is word based instead of being single-character based.
Defined the measure, the distance between buggy and fixed method for each of the BFPs was computed and we observed the data distribution; then we used the quartiles' values to divide the initial dataset in multiples smaller datasets, where each of these represents a different level of difficulty.
By doing so, we clearly obtained 4 level of difficulty; thus we defined an incremental difficulty criteria to feed the model with: the bug-fixing training scheduler described as follows.

Training Scheduler. The training scheduler decides the sequence of data subsets throughout the training process based on the judgment from the difficulty measurer. As mentioned in the introduction of this chapter, Baby Step approach switches to the next difficulty level as soon as the model converges on the previous one. In the bug-fixing task, the scheduler adjusts the training data subsets based on an early stopping criterion: after a fixed number of steps early stopping is run and it considers the model as being converging if the loss value does not improve after 5K steps.

4.3.2 Code summarization task

As well as happened for the training on the baseline model, a random sample of 200K method-comment pairs was picked from the initial dataset; evaluation and test datasets, however, were used as-is. This initial dataset was then divided in 4 sub-datasets, each representing a different level of difficulty.

Difficulty Measurer. In Natural Language Processing tasks, the **sentence lenght** intuitively expresses the complexity of a sentence, thus we used the instances' lenghts as measure of complexity. However, instead of focusing on the source of the couple, for this task we decided to sort out the subdatasets computing the difficulty measure on the target, namely the lenght of each method's comment. Once obtained the lenghts, accordingly to the data distribution we took advantage of the quartiles obtained to split the intial dataset in 4 smaller buckets. Accordingly to the training set, the evaluation set was splitted in 4 sub-datasets as well, because the model trained on a defined level (or levels) of difficulty needs to be evaluated on instances with the same complexity, according to the curriculum learning idea.

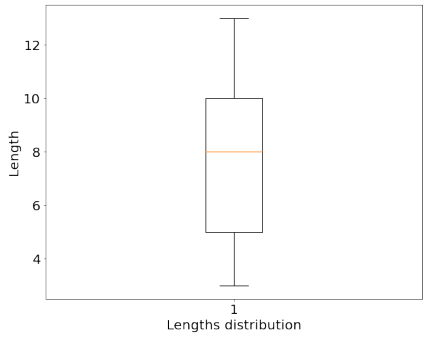

FIGURE 4.2: Code summarization task: data distribution.

Training Scheduler. Similarly to what happens in bug-fixing task, the training scheduler decides to sample from more harder data only when the model converges on the previous easier bucket. The convergence is assessed after a fixed number of epochs, after which early stopping with patience of 5 epochs is run. Once the model converges on the first bucket, the training proceeds on the second and so forth.

4.3.3 Log generation task

Inserting log messages is a practice broadly used and decide where to inject log statements, what information to report through it, and at which log level is as hard as it is crucial [5]. In the following section Curriculum Learning approach is applied at log generation task. The dataset for this task is composed by couple of methods, where the source is the method without the log statement, the target instead has not only the log statement but also the log message.
Those couples where used to train the model to generate and inject log statements in Java code. If the training set for the training on the baseline model was composed by 106.382, the training set for the experiment with curriculum learning is of 105.985 instances. This was because of the Difficulty Measurer we choose to use: according to the latter, 397 instances had a null difficulty, hence were excluded from the set. The same happend in the case of the evaluation set, 13.197 instances were used instead of 13.260. The test set ???

Difficulty Measurer. Given the dataset at our dispose we choose as difficulty measurer the complexity of the instruction set in program execution tasks. More specifically we computed **cyclomatic complexity** through Lizard tool [8]. Cyclomatic complexity is a quantitative software metric developed by Thomas J. McCabe in 1976, used to indicate the complexity of a program. It measures the number of linearly-independent paths through a program module. The measure is computed using the control-flow graph of the program where the nodes of the graph correspond to sets of commands of a program, and a directed edge connects 2 nodes if

the second command might be executed immediately after the first command. It can be applied to individual functions, modules, methods or classes within a program. We decided to compute it on each of the source methods.
It must be said that from the initial dataset were removed 397 instances whose cyclomatic complexity was 0. Decided and computed the difficulty measure for each instance, the dataset was ready to be used by the training scheduler.

Training Scheduler. Once again we observed the data distribution and considered the 3 quartiles to divide the whole dataset in 4 subsets. Each of those represented a different level of difficulty.
As soon as the model converged on the current bucket, BLEU values after each epoch were assessed, and early stopping was applied. We took the best model before divergence to restart the training from the following bucket.

Chapter 5

Analysis of results

In the following section the results achieved by NMT Encoder-Decoder with curriculum learning and by the baselines for the three tasks we considered are reported. We used different metrics for each of the tasks, depending on the metric used in the works that introduced the baselines.

5.1 Bug fixing task

To begin with, as assessed by Tufano *et al.* [6] we used perfect predictions as a metric to evaluate the model performances. In the paper taken as reference to test curriculum learning on, the authors differentiated 2 types of datasets; (i) the first one constituted by methods whose length was up to 50 tokens and (ii) methods with length between 50 and 100 tokens. However, as stated in the previous sections we considered the merging of the two datasets, therefore firstly we reproduced the canonical learning with the merged dataset and then we tested curriculum learning approach. Table 5.1 reports the percentage of bug fixing pairs correctly predicted by the models for different beam sizes. Increasing the beam size, and generating more candidate

BEAM	Baseline	Baby-step
1	5.22 %	5.34 %
5	13.00 %	19.41 %
10	16.55 %	25.16 %
15	18.38 %	28.64 %
20	19.60 %	30.84 %
25	20.51 %	32.39 %
30	21.29 %	33.69 %
35	21.93 %	34.93 %
40	22.26 %	35.72 %
45	22.73 %	36.47 %
50	23.02 %	37.39 %

TABLE 5.1: Models' performances

patches accordingly, the percentages of BFPs for which the models can perfectly generate the

corresponding fixed code - starting from the input buggy code - increases. If the baseline model can predict the fixed code of 5.22% of the BFPs with only one attempt, the same model together with curriculum learning approach performs predicting 5.34% of the same BFPs. It is a decent better result, but looking at bigger beam sizes, the model with curriculum learning performs definetly better, almost triplicating the percentage of perfect predictions when 5 patches are generated, reaching 37.39% when 50 candidate patches are considered. Overall, it can be seen that the improvement margin is constant.
On a second evaluation, we carried on a complementary analysis based on perfect predictions obtained when the model generated 10 candidate patches for each prediction. As can be seen in Table 5.2, the combination between the two models leads to a reasonable percentage of perfect predictions, i.e. 42.38%. However, the model with curriculum learning approach performs even better and on its own. Indeed, 44.63% of BFPs are correctly predicted only by the model with baby-step. On the other hand, only 12.97% of perfect predictions are from the baseline. This result not only indicates that the two approaches are complementary for the bug-fixing

$Dataset(d)$	$Shared_d$	$OnlyBL_d$	$OnlyCL_d$
BF_{all}	42.38 %	12.97 %	44.63 %

TABLE 5.2: Overlapping metrics: baseline and curriculum learning models.

task, but also that the model with curriculum learning reports even better outcomes. Considering that the approach implemented is one of the easiest one, those results recall the need to better tuning the approach with the aim of other improvements in the ability of such a model to exploit the knowledge acquired on this specific task.

5.2 Code summarization task

	BLEU-A	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Baseline	13.70	38.8	18.4	10.6	7.1
Baby-step	13.29	38.0	17.8	10.2	7.0

TABLE 5.3: BLEU values summary .

	Baseline	Baby-step
BEAM 1	4.66 %	4.64 %
BEAM 5	6.68 %	6.53 %
BEAM 10	7.75 %	7.52 %
BEAM 15	8.36 %	8.07 %
BEAM 20	8.84 %	8.51 %
BEAM 25	9.17 %	8.83 %
BEAM 30	9.47 %	9.06 %
BEAM 35	9.74 %	9.29 %
BEAM 40	9.90 %	9.49 %
BEAM 45	10.06 %	9.67%
BEAM 50	%	9.83%

TABLE 5.4: Perfect predictions summary.

5.3 Log generation task

Chapter 6

Conclusion

Appendix A

Appendix

Bibliography

[1] V. Cirik, E. H. Hovy, and L.-P. Morency. Visualizing and understanding curriculum learning for long short-term memory networks. *ArXiv*, abs/1611.06204, 2016.

[2] J. L. Elman. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99, 1993.

[3] A. LeClair, S. Haque, L. Wu, and C. McMillan. Improved code summarization via a graph neural network. *CoRR*, abs/2004.02843, 2020.

[4] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, feb 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[5] A. Mastropaolo, L. Pascarella, and G. Bavota. Using deep learning to generate complete log statements. *CoRR*, abs/2201.04837, 2022.

[6] M. Tufano, C. Watson, G. Bavota, M. D. Penta, M. White, and D. Poshyvanyk. An empirical study on learning bug-fixing patches in the wild via neural machine translation. *ACM Trans. Softw. Eng. Methodol.*, 28(4), sep 2019.

[7] X. Wang, Y. Chen, and W. Zhu. A comprehensive survey on curriculum learning. *CoRR*, abs/2010.13166, 2020.

[8] T. Yin.