LARGE LANGUAGE MODELS: COMPARISON OF CROSS-ENTROPY AND BINARY CROSS-ENTROPY LOSS

LIELIE VALODAS MODEĻI: ŠĶĒRSENTROPIJAS UN BINĀRĀS ŠĶĒRSENTROPIJAS ZUDUMU SALĪDZINĀJUMS

Author: **Ilmars Apeinans**, e-mail: ilmars.apeinans@rta.lv
Scientific supervisor: **Sergejs Kodors, Dr.sc.ing.,** e-mail: sergejs.kodors@rta.lv
Scientific supervisor: **Imants Zarembo, Dr.sc.ing.,** e-mail: imants.zarembo@rta.lv
Rezekne Academy of Technologies
Atbrīvošanas aleja 115, Rēzekne, LV-4601

Abstract. *The paper explores Large Language Model (LLM) training on custom datasets for classification microservice development. As training general purpose models for every possible situation is not feasible on smaller scale, because of limitations of computation power, usage of smaller model architectures, such as NanoGPT for training LLM model for specific use-case is a more cost-effective solution. In this article the dataset "Internet Movie Database (IMDB)" is applied in the experiment for LLM training. The dataset IMDB contains user comments about movies. Training criteria was Cross-entropy Loss (CELoss) and Binary Cross-entropy Loss (BCELoss), which were compared in the experiment. LLM training showed that validation accuracy for CELoss is 85.84% while validation accuracy for BCELoss is 86.1%. The biggest difference was in the consistency of results as distance between minimal and maximal accuracy for CELoss was 2.36%, but BCELoss distance between minimal and maximal accuracy was 1.04% providing more stable accuracy.*

Keywords: *artificial intellect, classifier, large language models, machine learning.*

Introduction

With advancement of artificial intelligence (AI) and popularity of Large Language Models (LLMs) through AI tools like ChatGPT [1] and Gemini [2] released by companies OpenAI and Google, the development of LLMs become strongly intensive after 2022 (see Fig. 1). ChatGPT and Gemini are general purpose tools and cover all possible topics that users may request within a communication. Meanwhile, LLM-based microservices with specific knowledge are required for less ambitious projects like internet shops, recommendation systems, etc.

The goal of this article is to experimentally compare two LLM architectures with training strategies: Cross-entropy Loss (CELoss) and Binary Cross-entropy Loss (BCELoss).

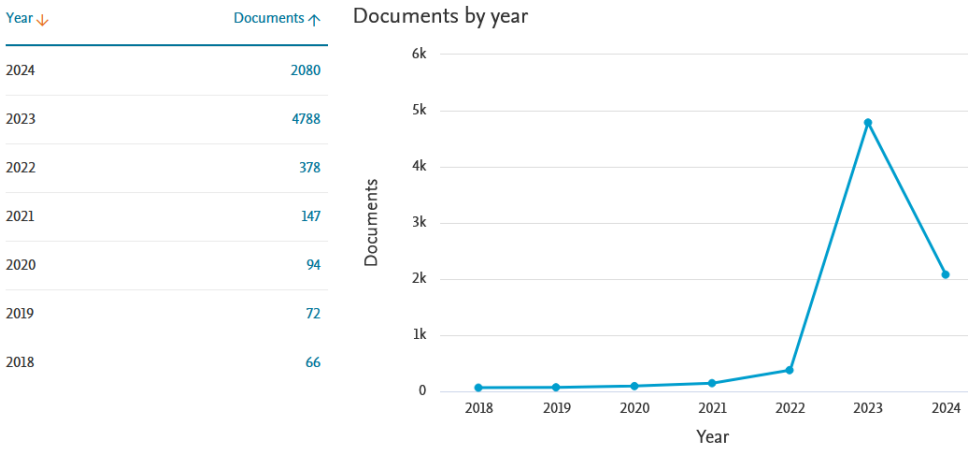

Fig. 1. Documents by year: large language models (Scopus)

https://doi.org/10.17770/het2024.28.8251

Materials and methods

Dataset

Dataset [3] used in this experiment contains 50000 reviews from Internet Movie Database (IMDB). The reviews are divided on 25000 “positive” and 25000 “negative” categories. Randomly selected 25000 training samples and 2500 validation samples were used for LLM training.

The dataset contains no more than 30 reviews per each movie. The training and testing datasets contained different movies to disclose the memorizing unique terms for specific movie, which could impact on approximation.

LLM architecture

CELoss and BSELoss training strategies require specific architecture for each case. BSELoss is related to binary classification – one output with two states “True” or “False”. Meanwhile, CELoss is used for multiple classification tasks with binary outputs.

LLM classifier is a lightweight architecture of design presented in the article “Attention Is All You Need” [4]. The LLM classifier contains only encoder part and classifier layer without decoder part (see Fig. 2). The LLM architecture without Position Encoding was applied in this experiment. The original architecture contains Softmax classifier, which is trained using CELoss strategy. If there is one output as in the case of dataset IMDB, it can be replaced by Sigmoid output with BSELoss.

PyTorch framework was applied in the experiment. Tiktokenizer was used for text embedding, cl100k_base version.

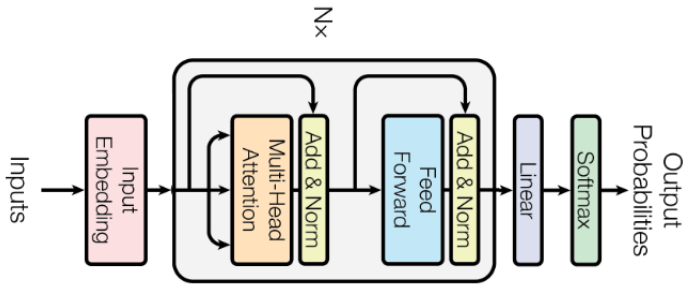

Fig. 2. LLM classifier with Softmax and CELoss

Experiment

Training was done using Nvidia GeForce RTX 4070Ti video card with support of CUDA technology [5]. 5 training runs were performed using each criterion to be able to calculate mean result. Each training run consisted of 10 epochs.

The first criteria that was used in the experiment was CELoss. CELoss calculates the loss by taking the negative log of the predicted probability assigned to the “True” class. If the model predicts a high probability for the true class, the loss is low. Conversely, if the model predicts a low probability for the “True” class, the loss is high.

Mathematically, for a single instance, the Cross-entropy Loss is defined as [6]:

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \quad (1)$$

where:

- L is the loss for one instance,
- M is number of classes,
- $y_{o,c}$ is a binary indicator of whether class c is the correct classification for the observation o,
- $p_{o,c}$ is the predicted probability that observation o belongs to class c.

Binary Cross-entropy Loss quantifies the difference between the "True" labels and the predicted probabilities of the "positive" class. For a model to perform well, it should predict probabilities close to 1 for the "True" class and close to 0 for the "False" class. The loss for each instance is calculated as [7]:

$$L = \sum_{i=1}^{N} [y_i \, log(p_i) + (1 - y_i) \, log(1 - p_i)] \quad (2)$$

where:

- L is the loss calculated over all instances,
- N is the number of instances,
- y_i is the "True" label for instance i, and
- p_i is the predicted probability that instance i is of the positive class.

The function penalizes the predictions that are confidently wrong more heavily, encouraging the model to be as accurate as possible.

Results

The training results are depicted in Table 1. The better results were achieved when BCELoss was used. The minimal BCELoss accuracy was 85.48% while the minimal accuracy for CELoss was 84.24%. The maximal accuracy for BCELoss was achieved at 86.52% and it was less than CELoss with accuracy of 86.6%.

Table 1.

LLM testing results

	min	mean	median	max
BCELoss	85.48	86.1	86.32	86.52
CrossEntropyLoss	84.24	85.84	86.36	86.6

Discussion

The achieved results show that there is no significant difference in accuracy between two LLMs with BCELoss and CELoss training strategies, but it is important to consider consistency of results (see Fig. 3). BCELoss distance between the minimal and maximal accuracy is only 1.04% while at the same time distance between the minimal and maximal when using CELoss is 2.36%. In situations where multiple models are trained and finetuned, smaller distance between the minimal and maximal accuracy will result in better overall accuracy of the final model.

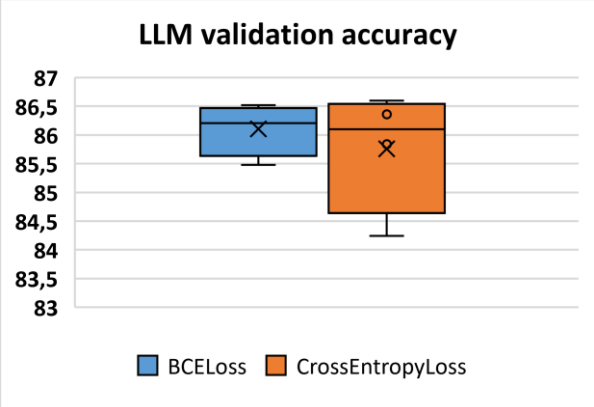

Fig.3. **LLM validation accuracy results**

If results are examined by each epoch, then progress of training can be discussed. If we look at the best run of BCELoss criteria in Fig. 4, the 3rd epoch showed the best improvement of accuracy, but later epochs provided less implementation in comparison. Important to take

note, that training and validation graphs show differences in values, which are relatively close to each other, and it may be wise to try training again, but increase the number of epochs.

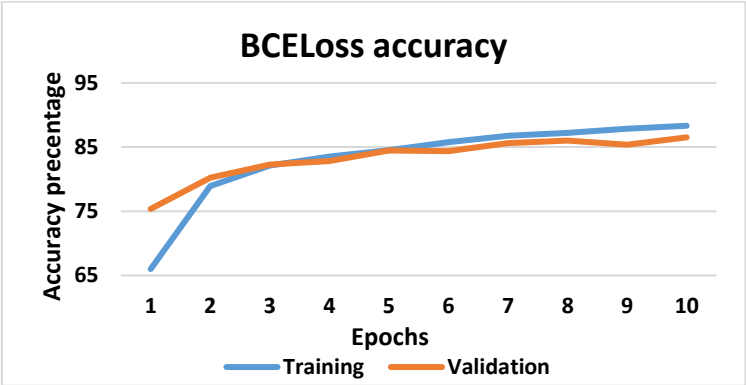

Fig. 4. **BCELoss training accuracy with 10 epochs**

If we look at the CELoss accuracy graph in Fig. 5, we can see that increase in accuracy in the 3rd epoch is great as well, but the validation accuracy after the 3rd epoch stats to slow down and at the end of graph at 10th epoch stated to drop down. The distance between validation and training accuracy is growing for each epoch. It may result in situation where validation may not increase and stagnates at 85%, but further testing is required.

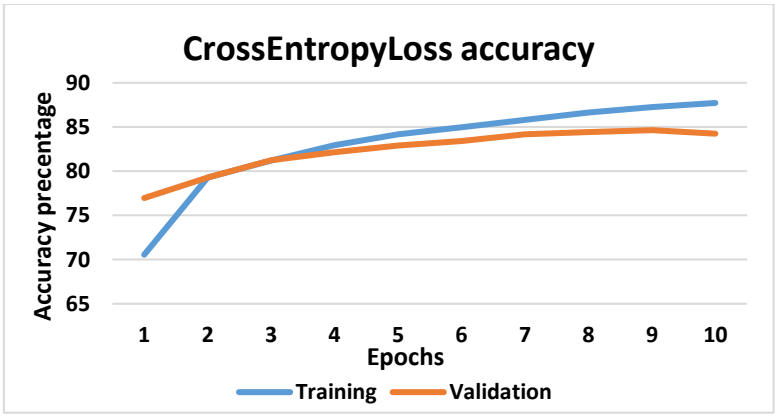

Fig.5. **CrossEntropyLoss training accuracy with 10 epochs**

Acknowledgment

This research is funded by the Latvian Council of Science, project "Testing Interventions and Developing a Knowledge-based Recommendation System to Reduce Plate Waste in School Catering in Latvia", project No. lzp-2022/1-0492.

Bibliography

1. ChatGPT by OpenAI [Online] [Reference to 10.04.2024.]. Available: https://chat.openai.com/
2. Gemini by Google [Online] [Reference to 10.04.2024.]. Available: https://gemini.google.com/
3. Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. Learning word vectors for sentiment analysis. ACL Anthology. June 2011. [Online] Available: https://aclanthology.org/P11-1015/
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. Attention is all you need. June 2017. [Online] Available: https://arxiv.org/abs/1706.03762
5. CUDA toolkit [Online] [Reference to 10.04.2024.]. Available: https://developer.nvidia.com/cuda-toolkit
6. Cross-Entropy Loss and Its Applications in Deep Learning [Online] [Reference to 04.04.2024.]. Available:https://neptune.ai/blog/cross-entropy-loss-and-its-applications-in-deep-learning#:~:text=Cross%2Dentropy%20loss%20is%20the,how%20effective%20each%20model%20is
7. Shalev-Shwartz, S., & Ben-David, S. Understanding machine learning. 2014. [Online] Available: https://doi.org/10.1017/cbo9781107298019