Replication of Krishna's paper

Terry Cruz Melo

Supervision

– Eduardo Blanco

Abstract

This is a replication of the paper *Disentangling Indirect Answers to Yes-No Questions in Real Conversations* by Krishna et al. (2022)[1]. The goal of this replication is to fine-tune the model and the evaluation metrics to better fit the context of the project. In the replacation, the same models, datasets, and evaluation metrics are used. The aim is to get similar results to the original paper.

Steps for replication

We need to get the same results as the original paper. To do so, we need to follow the same steps as the original.

1. Setup the environment and install the dependencies.
2. Download the datasets located at conversations *repository*.
3. Import required libraries, load pre-trained RoBERTa tokenizer and model, and define label mappings.
4. Define a function to compute metrics (accuracy, precision, recall, F1-score, classification report) and tokenize input text.
5. Load train, validation, and test datasets, and convert labels to numerical format.
6. Create tokenized train, validation, and test datasets using the tokenizer.
7. Define a custom PyTorch Dataset class to convert tokenized inputs and labels into a format suitable for model training.
8. Instantiate the custom Dataset class with train, validation, and test encodings and labels.

9. Set up training arguments with appropriate hyperparameters and instantiate the Trainer class.
10. Train, evaluate, and save the fine-tuned model, tokenizer, and evaluation metrics.
11. Predict labels for the test dataset, and calculate the classification report.

Criteria for replication

The stopping criteria and hyperparameter settings for the model training are summarized below. These settings determine the training duration, evaluation frequency, and how the best model is selected during the training process.

1. **num_train_epochs**: Total training epochs (32).
2. **per_device_train_batch_size**: Training batch size (16).
3. **per_device_eval_batch_size**: Evaluation batch size (16).
4. **warmup_steps**: Warmup steps for learning rate (200).
5. **weight_decay**: Weight decay strength (0.01).
6. **load_best_model_at_end**: Load the best model at the end (True).
7. **metric_for_best_model**: Use F1 score to determine the best model.
8. **evaluation_strategy**: Evaluate at the end of each epoch (epoch).
9. **save_strategy**: Save at the end of each epoch (epoch).

Training stops after 32 epochs, and the best model based on F1 score is used for evaluation and prediction.

Results

SwDA-IA_Q

Experiment	All Labels			Yes	ProbYes	Middle	ProbNo	No
	P	R	F1	F1	F1	F1	F1	F1
Krishna et al. (2022)	0.18	0.43	0.26	0.45	0.18	0.22	0.22	0.18
Terry	0.30	0.29	0.28	0.55	0.13	0.47	0.12	0.15

Table 1: Comparison of Performance Metrics for Sequence Classification Models: The table shows the performance of two models, Krishna et al. (2022) and Terry, on a classification task. The models were trained on the SwDA-IA_Q dataset.

SwDA-IA_A

Experiment	All Labels			Yes	ProbYes	Middle	ProbNo	No
	P	R	F1	F1	F1	F1	F1	F1
Krishna et al. (2022)	0.42	0.45	0.43	0.60	0.29	0.40	0.09	0.31
Terry	0.51	0.49	0.50	0.66	0.38	0.67	0.29	0.50

Table 2: Comparison of Performance Metrics for Sequence Classification Models: The table shows the performance of two models, Krishna et al. (2022) and Terry, on a classification task. The models were trained on the SwDA-IA_A dataset.

SwDA-IA_Q_A

Experiment	All Labels			Yes	ProbYes	Middle	ProbNo	No
	P	R	F1	F1	F1	F1	F1	F1
Krishna et al. (2022)	0.44	0.45	0.44	0.62	0.31	0.33	0.08	0.44
Terry	0.54	0.53	0.53	0.71	0.39	0.69	0.27	0.59

Table 3: Comparison of Performance Metrics for Sequence Classification Models: The table shows the performance of two models, Krishna et al. (2022) and Terry, on a classification task. The models were trained on the SwDA-IA_Q_A dataset.

Takeaways

The main takeaway from this replication report is the valuable experience gained in finetuning a RoBERTa-based model using the Hugging Face tutorial [2]. By comparing the performance of the original Krishna et al. (2022) model with my own replication model, Terry, and analyzing metrics like precision, recall, and F1 scores, I've gained a deeper understanding of how well both models predict different labels.

This experience has been instrumental in learning how to finetune RoBERTa models for classification tasks, allowing me to explore ways to improve model performance and apply these techniques to future projects.

References

[1] K. Sanagavarapu, J. Singaraju, A. Kakileti, A. Kaza, A. Mathews, H. Li, N. Brito, and E. Blanco, "Disentangling indirect answers to yes-no questions in real conversations," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 4677–4695. [Online]. Available: https://aclanthology.org/2022.naacl-main.345

[2] "Introduction - hugging face course." [Online]. Available: https://huggingface.co/course/