Using Ordinal Labels for Text Augmentation and Simultaneous Contrastive Learning for DownStream Task

Anonymous ACL submission

Abstract

Leveraging large language models, advancements in text augmentation and embedding models for downstream tasks have shown promise, However yet challenges remain in distinguishing texts with similar meanings. The proposed scheme, incorporating ordered labels to enhance sequence information, employs an integrated technique combining Contrastive and Downstream Learning The proposed scheme outperforms Full Fine-Tuning methods using only classfication learning in text classification because it effectively uses ordered labels to train the model to distinguish similar texts with greater accuracy. our method boosts data diversity and model accuracy by refining the model's sensitivity to nuances, utilizing strongly hard-negative samples in generated texts to further enhance Contrastive Learning outcomes.

1 Introduction

In the field of Natural Language Processing (NLP), large language models (LLMs) have driven groundbreaking advancements, demonstrating superior performance over traditional techniques in various language tasks. Specifically, encoder-focused models developed using the Transformer architecture (1) have shown impressive performance in downstream tasks such as text classification, with notable examples including BERT (2), RoBERTa (3), and ELECTRA (4). These models have significantly improved the benchmarks for tasks such as sentiment analysis, sentence similarity evaluation, and document classification.

Text data generation and augmentation, as well as contrastive learning, are actively researched areas within NLP. Text data augmentation is an effective method for enhancing model performance by addressing data scarcity. Contrastive learning helps models learn similarities and differences between data, improving the representation of the embedding space.

Ordinal labels are used as key features in machine learning and deep learning. Zhu et al. (5) applied ordinal label relationships to regression problems, while Wen et al. (6) improved image model performance by learning the distribution of ordinal labels. In NLP tasks such as text classification, sentiment analysis, and similarity evaluation, ordinal labels can play a crucial role. For example, in sentiment analysis, the intensity of emotions can be distributed continuously from "very negative" to "very positive." Correctly learning these ordinal labels impacts the model's performance and generalization ability.

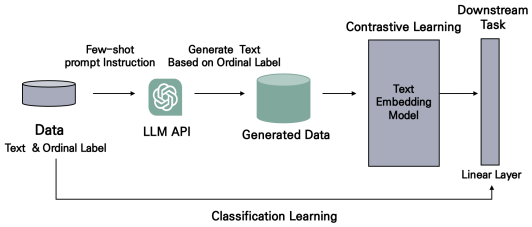

Figure 1: Ordinal Labels for Text Augmentation and Simultaneous Contrastive Learning for DownStream Task

We propose a method to augment text through ordinal label-based data generation, enhancing contrastive learning. This study aims to improve NLP task performance by integrating LLM-based text augmentation and contrastive learning using ordinal labels. We used LLM for data augmentation and prompt engineering to generate additional data based on the selected original labels and text(anchor). Each anchor text generated several new texts reflecting the differences in ordinal information. Additionally, we introduce a loss function that constructs n hard negatives based on the differences in ordinal labels from the augmented data. Our approach integrates contrastive learning and downstream task learning simultaneously to

improve model performance in various tasks involving ordinal labels.

2 RELATED WORK

2.1 Ordinal Labels for Downstream Task

Ordinal labels (Ordinal Label) are a type of data that lies between categorical and continuous data, where the order information between labels plays a crucial role. Downstream tasks utilizing ordinal labels are differentiated from general classification problems in that they must consider the order relationship between labels. To effectively handle ordinal labels, it is important to design the model to learn while maintaining the order information.

(Ganu et al., 2009)(7) introduced methods for predicting/classifying ratings based on text reviews using ordinal labels. They classified ratings, which are ordinal labels based on text reviews, into four stages: Positive, Negative, Neutral, Conflict, and examined Accuracy, Precision, and Recall, and then predicted ratings through a regression model. Verma et al., 2017)(8) used a parallel LSTM to pass text input through and obtain latent vectors, which were then passed to a GRNN to predict ratings. Chen & Hendry, 2019)(9) proposed methods for predicting and recommending ordinal labels through noise reduction processes by configuring Discriminative classifiers and Generative classifiers.

2.2 Text Data Augmentation

Text data augmentation is crucial for enhancing model performance in NLP by increasing data diversity and helping models generalize better. Traditional methods include synonym replacement, random insertion, random deletion, and sentence shuffling, transforming original data to expose models to various scenarios.

Adding noise to text data, similar to image noise in computer vision, has also been explored. Xie et al. (10) introduced unigram and blank noising techniques, where tokens are randomly replaced or removed to create augmented data.

Recent advancements use LLMs like GPT-4(11) and LLaMA(12) to generate datasets through prompt engineering. Kieser et al. (13) used ChatGPT to simulate diverse groups, validating synthetic data through physical tests. Chowdhury et al. (14) generated context and question-answer pairs to improve reading comprehension model robustness.

Research for specific downstream tasks includes simple data augmentation techniques for text classification, such as randomly adding punctuation (15), and generating synthetic datasets using pretrained language models and task-specific prompts (16). These methods highlight the effectiveness of synthetic data in various NLP tasks, particularly sentiment analysis, text classification, and machine translation.

2.3 Contrastive Learning

Contrastive learning extracts useful features by bringing similar samples closer and pushing dissimilar ones apart in the embedding space. Initially used in unsupervised settings, it has recently been applied in supervised environments, effectively leveraging data structure even with limited labels.

Contrastive learning's core is to maximize similarity and differences between data samples using a margin-based loss function or Noise Contrastive Estimation (NCE). The InfoNCE loss function, a representative of NCE, is defined as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\log\left(\frac{\exp\left(z_i \cdot z_j/\tau\right)}{\sum_{k=1}^{N}\exp\left(z_i \cdot z_k/\tau\right)}\right) \quad (1)$$

where z_i and z_j are positive pair embeddings, z_k are other batch sample embeddings, and τ is a scaling parameter.

Contrastive learning is advantageous as it effectively uses unlabeled data, clusters similar samples, and improves model generalization. It is widely used in image processing, NLP, and speech recognition.

SimCSE (17) is a notable text contrastive learning study, presenting both unsupervised and supervised methods, achieving high performance in benchmarks like STS-B. Unsupervised SimCSE uses dropout as noise, encoding the same sentence twice with different dropout masks. Supervised SimCSE leverages Natural Language Inference (NLI) datasets, using entailment and contradiction pairs as positive and hard-negative pairs.

DiffCSE (18) extends SimCSE by adjusting sensitivity to variations, extracting richer contextual representations through masked language model (MLM) methods and discriminator training for replaced token detection.

SupCon (19) applies supervised learning to image classification using contrastive loss, treating

same-class samples as positive pairs and different-class samples as negative pairs, learning feature differences for classification.

3 Proposed Method

3.1 Process

Our proposed method comprises the following two key processes: 1) generating augmented text data using an LLM with ordinal label information 2) leveraging this data for contrastive learning alongside the original data for classification.

3.1.1 Text Augmentation with Ordinal Labels

We use an LLM API to generate augmented text data. The augmentation process involves selecting anchor texts from the original dataset and generating new texts that reflect different ordinal rating levels. The steps are as follows:

1. **Anchor Text Selection**: Sample anchor texts from the original dataset.
2. **Prompt Engineering**: Each prompt includes examples that demonstrate how to transform a text to reflect different rating levels.xw
3. **Text Generation**: Use the LLM API with a temperature setting to generate N variations of each anchor text, corresponding to ordinal levels.

Result Example: 2 Review(Anchor): "Arrived late, sound quality below expectations, short battery life, slow customer service."

- 1 review: "Very disappointed overall. Never buy these headphones again."
- 2 review: "Delivery was delayed, and customer service was slow to respond. Not bad sound."
- 4 review: "Late delivery but good sound and decent battery life. Satisfied overall."
- 5 review: "Quick delivery, excellent sound, long battery, helpful customer service. Highly recommend!"

3.1.2 Simultaneous Learning

The generated text data, along with the original data, is used to perform simultaneous contrastive learning and classification. Positive pairs consist of anchor texts and their corresponding generated texts with the same ordinal label, while negative pairs consist of anchor texts and generated texts with different ordinal labels.

Additionally, we proposed a modified contrastive loss function, named noc (N-ordinal contrastive loss). This loss function includes hard negative samples to emphasize the differences between similar but differently labeled texts. It allows for the adaptive selection of N hard negatives based on ordinal labels for batch training.

$$\mathcal{L}_{\text{noc}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{\text{sim}(z_{a_i},z_{p_i})/\tau}}{e^{\text{sim}(z_{a_i},z_{p_i})/\tau}+\sum_{j=1}^{M}e^{\text{sim}(z_{a_i},z_{\text{hn}_{j_i}})/\tau}}\right) \tag{2}$$

where z_{a_i} is the anchor embedding, z_{p_i} is the positive embedding, and $z_{\text{hn}_{j_i}}$ are the hard negative embeddings. The similarity function sim uses cosine similarity, and τ is the scaling parameter(0, 1). The variable j indicates the number of hard negative samples included in the calculation, and it can be adaptively selected based on the difference in ordinal labels.

Train the model using both contrastive learning and downstream objectives. Use cross-entropy loss for classification(L1 loss for regression) and combine it with the contrastive loss. The total loss is calculated as:

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{classification}} + (1-\lambda)\mathcal{L}_{\text{noc}} \tag{3}$$

where λ is a weight parameter to balance the two losses.

4 Experiments

4.1 Dataset Description

The dataset used in this study consists of 200,000 text reviews and ratings from Naver, a major online shopping platform in Korea. The dataset includes text reviews and corresponding ratings, which are used as ordinal labels ranging from 1 to 5(except 3). We selected 4,000 anchor data per label for text generation.We used the ChatGPT API to generate four different texts for each anchor text, corresponding to different labels. A total of 64,000 generated data were used for contrastive learning.

4.2 Experimental Setup

We conducted experiments to evaluate the effectiveness of the proposed N-ordinal contrastive loss

(Noc) method. The embedding model used was 'klue-roberta-small' (20) , a pre-trained Korean language model. The dataset was split into training, validation, and test sets with a ratio of 0.6:0.2:0.2, and stratified sampling was used to maintain an even distribution of classes across the sets.the best value for λ weight parameter to balance the two losses was found to be 0.7. All experiments were performed on a single NVIDIA A100 GPU.

We fixed the contrastive loss and changed the downstream loss to L1 Loss to measure MAE, and to CrossEntropy Loss to measure Accuracy.

4.3 Baseline Models

We compared the proposed Noc method with two baseline models:

- **Baseline(Full)**: A full fine-tuning model that adjusts all weights of the embedding model for classification tasks using only the original dataset. This model is trained exclusively with classification training.
- **Baseline_Gen(Full)**: A full fine-tuning model that includes generated data for training, using the same augmentation method. This model employs the same methods as the Baseline but leverages generated data.
- **Not Simultaneous Learning**: A model where the embedding model is first trained using contrastive learning, and subsequently, the downstream task is trained. This approach separates the contrastive learning phase from the downstream learning phase, unlike the proposed method which integrates both simultaneously. This method follows the approach used by Setfit.(21)

4.4 Results and Discussion

The performance of the models was evaluated using Accuracy and MAE(Mean Absolute Error) as the primary metric. To ensure robust evaluation, we performed cross-validation by varying the random state for dataset splitting and reported the average accuracy across multiple runs.

Table 1 demonstrate that the proposed model can more effectively improve the model's performance compared to 3 other baselines.

4.5 Model Size Comparison

Additionally, we investigated the effect of model size by comparing 'klue-roberta-small' and 'klue-roberta-base' models. The results showed that increasing the model size led to improved performance, highlighting the benefits of using larger models for capturing more complex patterns in the data.

Table 1: Experimental Results(Average for 5 time cv)

Model	Acc(%)	MAE
Ours	**72.74**	**0.356**
Baseline(Full)	72.03	0.3752
Baseline_Gen(Full)	71.76	0.379
Not Simultaneous (21)	67.25	N/A

Table 2: Impact of Model Size on Performance

Embedding	param	Avg Acc(%)
klue-roberta-small	68.1M	72.74
klue-roberta-base	111M	73.10

Table 2 shows the impact of model size on performance. The use of larger models, such as klue-roberta-base, also contributed to better performance, indicating the advantage of increased model capacity in our method.

5 Conclusion

This study proposed a novel method for text augmentation and contrastive learning using ordinal labels to improve the performance of NLP models in downstream tasks. The method leverages large language models (LLMs) for text generation and incorporates ordinal labels to generate diverse datasets. By integrating simultaneous contrastive learning and classification learning, the method effectively captures the nuanced differences between texts with different ordinal labels, leading to enhanced model performance. The proposed scheme provides a robust framework for leveraging ordinal labels in text augmentation and simultaneous contrastive learning, offering significant potential for advancing NLP downstream task performance in diverse applications.

Limitation

Limitation of our study is that we did not perform downstream task experiments on additional ordinal datasets. Our experiments were conducted using a dataset. Future work should include testing the proposed model across multiple languages and diverse

datasets to validate its robustness. Furthermore, a direction for future work is to develop and integrate methods for embedding data more finely in the latent space based on differences in ordinal labels. This could potentially enhance the model's ability to capture subtle differences between data points with similar texts with ordinal labels.

Ethics Statement

The study was conducted in accordance with the ACL Ethics Policy.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-December.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1.

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[4] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. 8th International Conference on Learning Representations, ICLR 2020.

[5] Zhu, H., Zhang, Y., Li, G., Zhang, J., & Shan, H. (2020). Ordinal distribution regression for gait-based age estimation. Science China Information Sciences, 63(2).

[6] Wen, C., Zhang, X., Yao, X., & Yang, J. (2023). Ordinal Label Distribution Learning. Proceedings of the IEEE International Conference on Computer Vision.

[7] Ganu, G., Elhadad, N., & Marian, A. (2009, June). Beyond the stars: Improving rating predictions using review text content. In WebDB (Vol. 9, pp. 1-6).

[8] Verma, S., Saini, M., & Sharan, A. (2017, August). Deep sequential model for review rating prediction. In 2017 Tenth international conference on contemporary computing (IC3) (pp. 1-6). IEEE.

[9] Chen, R. C. (2019). User rating classification via deep belief network learning and sentiment analysis. IEEE Transactions on Computational Social Systems, 6(3), 535-546.

[10] Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., & Ng, A. Y. (2017). Data noising as smoothing in neural network language models. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings.

[11] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

[12] Meta. (2024, April 18). Introducing Meta Llama 3: The most capable openly available LLM to date. Meta AI. https://ai.meta.com/blog/meta-llama-3/

[13] Kieser, F., Wulff, P., Kuhn, J., & Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. Physical Review Physics Education Research, 19(2).

[14] Chowdhury, A. G., & Chadha, A. (2023). Generative data augmentation using LLMs improves distributional robustness in question answering. arXiv preprint arXiv:2309.06358.

[15] Karimi, A., Rossi, L., & Prati, A. (2021). AEDA: An easier data augmentation technique for text classification. Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021. https://doi.org/10.18653/v1/2021.findings-emnlp.234

[16] Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., & Kong, L. (2022). ZEROGEN: Efficient zero-shot learning via dataset generation. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022.

[17] Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings.

[18] Chuang, Y. S., Dangovski, R., Luo, H., Zhang, Y., Chang, S., Soljačić, M., Li, S. W., Yih, W. T., Kim, Y., & Glass, J. (2022). DiffCSE: Difference-based contrastive learning for sentence embeddings. NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference.

[19] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning. Advances in Neural Information Processing Systems, 33, 18661-18673.

[20] Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.

[21] Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., Pereg, O. (2022). Efficient few-shot learning without prompts. arXiv preprint arXiv:2209.11055.