Dataset Merging and Distillation for Text Classification

Sunny Son[1] **Albert Kong**[1] **Dennis Hu**[1] **Lake Wang** [1] **Haohai Pang**[1] **Kevin Hong**[1]

[1]Center of Data Science, New York University

Abstract

Large NLP models require extensive datasets and significant computational resources, making efficient training challenging. Dataset distillation mitigates this by creating smaller, synthetic datasets that retain essential task-specific information, allowing resource-efficient training with minimal performance loss. While prior research has focused on single-task distillation for image and basic NLP tasks, this project extends dataset distillation to support dataset merging for multi-task learning. Our methodology leverages pre-trained transformer models like BERT, frozen encoders, and innovative merging techniques, including synthetic and embedding-level dataset merging. Experiments on AG News and SST-2 benchmarks show slight performance improvements and significant resource savings, highlighting the potential of distilled dataset merging to enhance generalization across tasks. Challenges remain in refining merging techniques and expanding applications beyond classification, but the results underline the promise of this methodology for multi-task learning.

Introduction

The ever-growing size of NLP models and datasets has spurred interest in techniques to reduce computational costs while maintaining high performance. Dataset distillation, a method initially applied to computer vision, compresses datasets into smaller, synthetic representations that preserve essential task-related information. In recent years, this technique has been extended to NLP tasks, addressing challenges posed by the discrete nature of text and the diversity of tasks. However, existing work primarily focuses on single-task applications, leaving the potential for multi-task learning and dataset merging largely unexplored.

Our research builds upon single task dataset distillation and seeks to investigate the potential of dataset distillation in creating a more versatile and efficient fine-tuning process for NLP models. By combining knowledge distillation through transfer learning and label-based dataset distillation, we hypothesize that we can create a more compact, yet comprehensive dataset that enables efficient fine-tuning across multiple tasks.

Our work is structured into three phases: reproducing results for single-task distillation, during which time we also train source model on the first task for transfer learning, extending distillation to merged datasets, and evaluating the performance of multi-task learning models. Through this effort, we seek to uncover the trade-offs in performance, computational resources, and dataset size when combining transfer learning and distilled datasets techniques, and establish a foundation for further advancements in multi-task NLP models.

Our key hypothesis is that a distilled dataset created from a combination of task-specific datasets will enable more efficient and effective multi-task learning in NLP models, potentially achieving few-shot learning capabilities. Our contributions are summarized in two key points:

- *Multi-task learning.* We explore the feasibility of multi-task learning through dataset distillation– that is to train the model to do multiple tasks using merged distilled data.

- *Insights into distillation for discrete data.* We aim to derive more insights and findings through experiments on discrete data for further research and studies of dataset distillation in the natural language domain.

Related Works

Dataset Distillation

Inspired by the idea of network distillation (Hinton et al., 2015), which studies the methodology of distilling the knowledge of a larger neural network into a smaller one, Dataset distillation (DD), likewise, seeks to create smaller, representative datasets from large training data to reduce computational complexity and training time, while maintaining model performance. DD was first introduced in the paper *Dataset Distillation* (T. Wang et al., 2020), along with a proposed algorithm using backpropagation through optimization steps. This method optimizes a tiny set of synthetic images that can train a model to achieve performance similar to training on the full dataset. From then on, the majority of published studies on DD focus on distilling image data. One direction for distillation is gradient/trajectory matching (Zhao et al., 2021, Cazenavette et al., 2022,Shin et al., 2023), where the synthetic data is optimized by matching the gradients of the model trained on the synthetic data to the gradients of the model trained on the original data. The goal is to ensure that the distilled, smaller dataset captures the same learning signals as the full dataset, enabling the model to perform similarly. Another direction involves distribution and feature matching (K. Wang et al., 2022, Zhao et al., 2021). By matching features or distributions, the distilled dataset retains critical information necessary for training models effectively, enabling comparable model performance to be achieved with a much smaller dataset. There are also studies on kernel-based learning (Nguyen et al., 2021) that leverages kernel methods to approximate complex data distributions through kernel functions.

Text Dataset Distillation

Unlike image, text data presents unique challenges due to the discrete nature of language. As a result, there are not a lot of studies on text dataset distillation. The existing papers focus on label distillation, including soft label distillation (Sucholutsky and Schonlau, 2021) and attention label distillation (Maekawa et al., 2023). The labels are used as the supervision of probabilities optimized as a part of the distilled dataset, to enhance the effectiveness of the distilled dataset for training the transformer-based models.

Datasets

The AG News dataset (Zhang et al., 2016) is a widely used benchmark dataset for text classification tasks. It consists of over 120,000 news articles categorized into four distinct classes: World, Sports, Business, and Science/Technology, where each article includes a title and a short description. The dataset is balanced across categories, enabling researchers to evaluate the performance of various machine learning and deep learning models.

The SST-2 dataset (Socher et al., 2013), derived from the Stanford Sentiment Treebank, is a popular benchmark for sentiment analysis tasks. It consists of 11,855 of movie reviews annotated with binary sentiment labels: positive or negative. Note that SST-2 excludes neutral sentiment examples, making it focused for binary classification. Each example includes either a full sentence or a phrase extracted from a movie review, allowing models to learn and evaluate sentiment polarity effectively. The dataset is widely used in natural language processing research, and is a core component of the GLUE benchmark for evaluating language understanding systems.

Methodology

This project examines the application of dataset distillation for multi-task learning by leveraging techniques originally developed for single-task scenarios. The methodology begins with replicating prior results to establish a strong foundation, followed by exploring novel merging strategies for multi-task applications.

Replication of Single-Task Dataset Distillation Reproducing results from *Dataset Distillation with Attention Labels for fine-tuning BERT* (Maekawa et al., 2023) is a critical first step in this study. AG News and SST-2 are selected as benchmark datasets due to their well-defined tasks in text classification. AG News involves theme classification with balanced classes, while SST-2 focuses on binary sen-

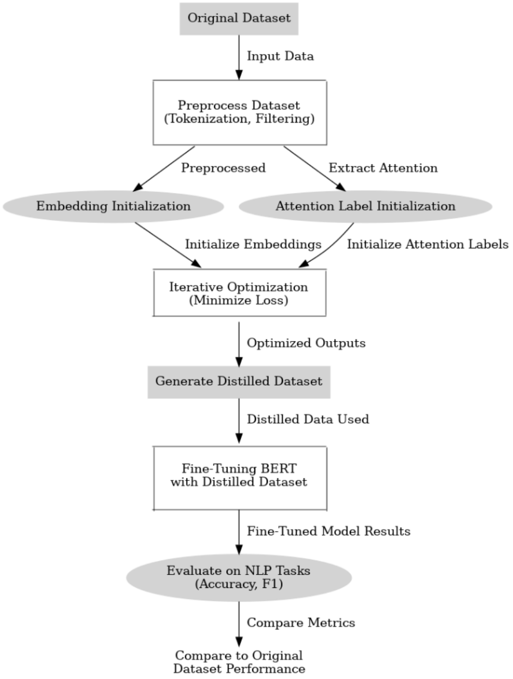

Figure 1: Flow chart of Dataset Distillation with Attention Labels

timent classification, making them ideal for evaluating the effectiveness of distilled datasets.

The replication process involves implementing the distillation algorithm as described in the original work. The key idea is to optimize a synthetic dataset that retains essential task-specific information while significantly reducing the dataset size. During this step, distilled datasets are evaluated using standard metrics such as accuracy and F1-score. These metrics provide a clear indication of how well the distilled datasets replicate the performance of full datasets, serving as a baseline for comparison in subsequent phases.

Text Classification Transfer Learning

Transfer learning leverages the knowledge gained from a pre-trained model on one task to enhance the performance of a different yet related task. In this project, we used the pre-trained AG News encoder as the foundation for transfer learning onto SST-2 dataset, which includes a new classification head tailored for the SST-2 dataset. Through this approach, We aim to eliminate the need for extensive retraining by freezing the encoder and fine-tuning only the classification head, thus significantly reducing computational overhead and training time. Through this process, we are able to achieve same level of accuracy in merely 6 epochs of training, compared to the training-from-scratch counterpart which trained for 30 epochs, thus the embeddings are generated with a consistent quality and scale, ensuring reliable performance and better efficiency.

Exploration of Dataset Merging for Multi-Task Learning Building on the results from replication, the study explores merging distilled datasets to support multi-task learning. Two strategies are examined: synthetic data merging, which integrates distilled datasets at the text level to create a unified dataset, and embedding-level merging, which combines distilled representations to share features across tasks while preserving task-specific nuances.

To address challenges in merging, the distillation algorithm is modified to balance tasks and prevent dominance by one over another. Iterative experimentation refines these strategies, focusing on trade-offs between dataset compactness and task-specific performance.

A pre-trained BERT model is used to evaluate the merged datasets, with a frozen encoder leveraging pre-learned features and task-specific heads fine-tuned on AG News and SST-2. Metrics such as accuracy, F1-score, and evaluation loss assess the effectiveness of merging strategies compared to single-task baselines, highlighting their potential for multi-task learning.

Experiments and Results

To evaluate multitask learning through dataset distillation, we employed a structured approach utilizing the BERT Base model. The encoder obtained from transfer-learning during distillation was loaded and frozen to retain its learned representations, while task-specific classification heads were added for the two original tasks: AG News (theme classification) and SST-2 (sentiment classification).

These tasks were selected to test whether multitask learning could facilitate knowledge sharing across different text classification datasets. Baseline performance metrics, including accuracy and evaluation loss, were recorded for individual tasks using the distilled datasets to establish a reference point for comparison.

Replication Results

Figures 2 and 3 show the dimension-reduced embeddings of distilled data for AG News and SST2, colored by the distilled soft labels. For AG News, four distinct clusters emerge, each corresponding predominantly to a single class. In contrast, the SST2 embeddings obtained through transfer learning are less optimized, as positive and negative labels are intermixed and exhibit a similar distribution.

Table 1 presents the replication performance of training on distilled data for individual tasks without any merging. Both accuracy and evaluation loss are significantly improved after training on distilled data, even in a 1-shot, 1-step training setting. The improvement is more pronounced for AG News than for SST2, likely due to AG News having a greater number of classes. This difference is reflected in the quality of the distilled embedding clusters.

Experiment 1

In the first experiment, distilled datasets were combined along the batch dimension, essentially aggregating samples from different single-task datasets into a single batch for simultaneous multitask learning. The model was fine-tuned on this merged dataset in a one-shot, single-step training scheme. During training, the loss function consisted of two terms: the cross-entropy loss between model logits and distilled soft labels, and the KL divergence loss between the model's attention weights and distilled attention labels. Evaluation loss was computed using only the cross-entropy term to focus on classification performance.

$$\mathcal{L}_{\text{train}} = \text{lr} \cdot \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{L}_{\text{logits}}(y_i,\hat{y}_i) + \lambda \cdot \mathcal{L}_{\text{attention}}(a_i,\hat{a}_i)\right)$$

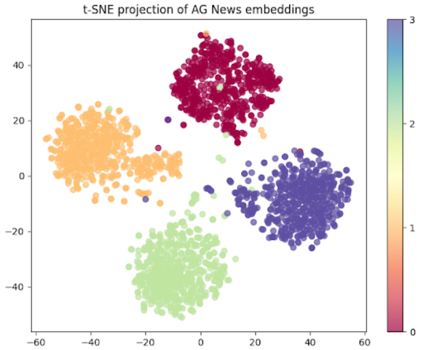

Figure 2: Dimensionally reduced t-SNE embeddings of distilled data for AG News

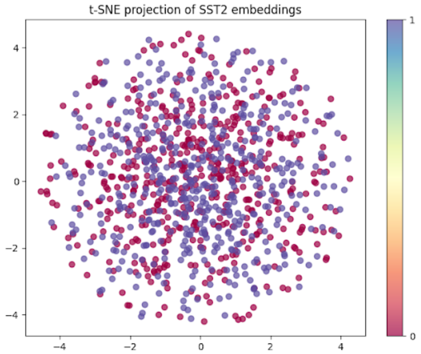

Figure 3: Dimensionally reduced t-SNE embeddings of distilled data for SST2

$$\mathcal{L}_{\text{evaluation}} = \frac{\sum_{j=1}^{B}\mathcal{L}_{,logits,j}\cdot \texttt{batch\_size}_j}{\texttt{total\_num\_samples}}$$

Table2 shows results from Experiment 1. The performance showed a slight improvement in accuracy for both original tasks. Training loss decreased as the model successfully optimized over the merged distilled data, but evaluation loss remained consistent. This suggests that the observed performance gains were primarily confined to training and did not translate into substantial generalization improvements. Increasing the number of training steps did not systematically enhance single-task performance. These findings highlight the feasibil-

Task	Pre Acc.	Pre Eval. Loss	Trained Acc.	Trained Eval. Loss	# Distilled Samples
AG News	0.24	1.46	0.89	0.56	4
SST2	0.51	0.69	0.86	0.34	2

Table 1: Distilled Baseline Performance for Individual Tasks. (Pre Acc. = Pretrain Accuracy, Pre EvalL. = Pretrain Evaluation Loss, Trained Acc. = Trained Accuracy, Trained EvalL. = Trained Evaluation Loss)

Task	Pre Acc.	Pre Eval Loss	Trained Acc.	Trained Eval Loss	Training Loss
AG News	0.225	1.43	0.327	1.41	1.01 → 0.28
SST2	0.478	0.73	0.531	0.69	0.13 → 0.07

Table 2: Experiment 1 Performance on Individual Tasks

ity of multitask learning even with limited data and computational resources but also expose challenges in maintaining task-specific performance.

Experiment 2

Building on multitask-learning in Experiment 1, the second experiment explored creating a new, joint task by combining datasets along the sequence length dimension. Sentences from AG News and SST-2 were paired to form combined inputs, encoding both theme and sentiment information. Hard labels from the original datasets were mapped to a cross-label space, effectively doubling the number of classes (from 4 to 8). To accommodate this increase, the distilled soft labels were generated through an outer product computation, capturing the joint distribution of task-specific features. This merging process increased the sample size to preserve one-shot training, allowing the model to learn from a richer representation of the tasks. We apply the same loss functions and training scheme as in experiment 1.

Table3 shows results from Experiment 2. They were consistent with those of Experiment 1. There was a slight accuracy improvement for the new task, alongside a noticeable reduction in training loss. However, evaluation loss remained stable, suggesting that improvements observed during training primarily arose from alignment with distilled attention labels rather than enhanced generalization. The merging process, while effective in expanding the label space and incorporating richer task representations, also introduced challenges. The outer product computation allowed for joint feature encoding, but the complexity of the merged labels made it difficult to ensure interpretability and efficient knowledge transfer. Nonetheless, generating distilled samples in a customized manner enabled the use of smaller batch sizes and extended training steps, contributing to slight performance gains.

Discussion

The improvements observed in both Experiment 1 and Experiment 2 are relatively small compared to the replication results. This indicates that while multitask learning is feasible with distilled data, effectively combining information from the distilled data remains challenging due to significant information loss. We also experimented with using a neural network to merge embeddings, but our ability to optimize this network was limited under the constraints of our training schedule (1-shot, 1-step) and the absence of a tailored loss function. Furthermore, using the same training loss for both the language model and the merge-embedding network did not yield any consistent improvements.

Embedding-to-Text Reconstruction Experiment

To explore the possibility of having better interpretation of the distilled dataset and create the merged dataset manually, we conducted an experiment to reconstruct human-readable text from embeddings

Task	Pretrain Acc.	Pretrain Eval. Loss	Trained Acc.	Trained Eval. Loss	Training Loss
Combined	0.118	2.14	0.142	2.12	1.78 → 0.84

Table 3: Experiment 2 Performance on Combined New Task

generated.

We implemented a token-by-token reconstruction approach, where each embedding vector in the distilled dataset was matched back to the most similar token in the same model vocabulary used using cosine similarity. This method processes embeddings independently and maps each embedding to the token with the highest similarity score. We used the distilled dataset generated for the AG News task, containing embeddings and corresponding classification labels. For each embedding vector, the similarity between the embedding and all tokens in the BERT embedding matrix was computed. The token corresponding to the highest similarity score was selected for reconstruction. BERT-based tokenizer for vocabulary and token handling. Cosine similarity was used to compute token similarity. There is an excerpt of the reconstructed text from the business class embeddings in the AG News distilled dataset in the Appendix.

Conclusion

In this project, we explored the feasibility of extending dataset distillation to merge datasets for multi-task text classification, using AG News and SST-2 as benchmarks. Experiments with embedding-level merging of synthetic data showed that distilled datasets retain essential task-specific information, achieving training outcomes comparable to full datasets. The results indicate the feasibility of multi-task learning with merged datasets, though gains in generalization and accuracy are modest. These findings advance dataset distillation research by exploring its potential for multi-task and cross-

Limitations

This study identified several limitations despite promising results. Embedding-level merging poses interpretability challenges, as converting high-dimensional embeddings back to text remains underexplored, limiting practical applications. The experiments focused solely on text classification, leaving the applicability of dataset distillation to tasks like summarization or question answering unaddressed. While distillation accelerates training by reducing sample size, it remains computationally intensive, with over 30 hours of GPU computation required for AG News on three NVIDIA A5000 GPUs.

The token-by-token reconstruction method for embeddings also has notable shortcomings. Subword prefixes from BERT's WordPiece tokenization reduce readability, contextual coherence is often lacking, and nonsensical or redundant phrases like ”between between” appear due to the absence of sequence-level attention.

Future Work

Future research should focus on advanced merging strategies that preserve semantic integrity and improve task performance, potentially using LLMs for embedding interpretation. Expanding dataset distillation to diverse NLP tasks, such as entity recognition and dialogue systems, would evaluate broader applicability. Incorporating metrics like computational efficiency and robustness would better assess real-world utility. Leveraging LLMs can also enhance embedding interpretability and enable generation of human-readable synthetic datasets.

To address reconstruction issues, context-aware methods using self-attention or decoder-based models like T5 or GPT could improve text coherence and fluency. Post-processing pipelines to merge subwords and remove redundancies, combined with metrics like BLEU scores, would further enhance readability.

References

Cazenavette, G., Wang, T., Torralba, A., Efros, A. A., & Zhu, J.-Y. (2022). Dataset distillation by matching training trajectories.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network.

Maekawa, A., Kobayashi, N., Funakoshi, K., & Okumura, M. (2023). Dataset distillation with attention labels for fine-tuning BERT. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 119–127.

Nguyen, T., Chen, Z., & Lee, J. (2021). Dataset meta-learning from kernel ridge-regression.

Shin, S., Bae, H., Shin, D., Joo, W., & Moon, I.-C. (2023). Loss-curvature matching for dataset selection and condensation.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642). Association for Computational Linguistics.

Sucholutsky, I., & Schonlau, M. (2021). Soft-label dataset distillation and text dataset distillation. *2021 International Joint Conference on Neural Networks (IJCNN)*.

Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., & You, Y. (2022). Cafe: Learning to condense dataset by aligning features.

Wang, T., Zhu, J.-Y., Torralba, A., & Efros, A. A. (2020). Dataset distillation.

Zhang, X., Zhao, J., & LeCun, Y. (2016). Character-level convolutional networks for text classification.

Zhao, B., Mopuri, K. R., & Bilen, H. (2021). Dataset condensation with gradient matching.

Appendix

Class	**Reconstructed Text (Excerpt)**
World (Class 1)	renewable depression turkey hell before mused vo d tle ship california sville ious tomatoes serves person worked miniature tension withdrew thought in gence armstrong classical beginning different a tead make polly doorbell blur belgian abolished margaret lacey...
Sports (Class 2)	amongst tv even ponytail gardner scout imagining wondered wondered fy arrow almost ten thought though function june shoe foster partly insisted trading reelection five salt baseball revenues hazel village wireless july saturday football scout nordic success company perspectives football...
Business (Class 3)	size directors logan look my behind my our came the decades four noctuidae mono many briggs precious moment joyah roosevelt kg daniel tully investigation between only imating da ignant connect war length september cabin orchestra christmas could disorder roughly bennett...
Sci/Tech (Class 4)	nivorous ap ison obe e revolves focus comic or renaissance album economy usually consists scandal whoever matching game divine excuse vowel substitute hips forget circulation game issued fetch career solar titles georgie admire...

Table 4: Examples of reconstructed text from distilled embeddings for the AG News task, categorized by class.