Dataset Distillation for Multimodal Data

Team Members: Cihan CALISIR
BBL 514E: Pattern Recognition & Analysis
Faculty of Computer and Informatics Engineering
Istanbul Technical University

***Abstract*—Training state-of-the-art deep learning models on large-scale datasets incurs significant computational costs and time. This project investigates Dataset Distillation (DD), a technique that synthesizes compact datasets preserving the essential knowledge of the original training data. We explore two distinct approaches: (1) *loss-based* dataset distillation for multimodal vision-language learning on the COCO dataset using a GPT2Vision-based architecture, and (2) coreset selection combined with Knowledge Distillation (KD) for unimodal classification on the CIFAR-10 dataset using ResNet-18 as the teacher model.**

For COCO, we pre-trained a vision-language model to identify challenging samples through per-sample cross-entropy loss. By selecting only the hardest examples — such as the top 5% or 30% of the dataset — we trained student models from scratch while maintaining strong performance on validation metrics. Encouragingly, a 5% distilled dataset trained for 220 epochs preserved most of the original model's accuracy despite drastically reduced training time.

For CIFAR-10, we applied K-Means clustering to extract 500 representative images per class, forming a 5,000-image coreset (90% reduction from the full training set). Training a student model solely on this coreset resulted in a performance drop (51.28% accuracy), but applying KD significantly improved accuracy to 70.21%, retaining over 84% of the teacher model's performance (83.58%) while accelerating per-epoch training by approximately 8-fold.

Our findings confirm that DD is a powerful strategy for reducing training cost without sacrificing performance, especially when augmented with KD techniques. The source code for this project is available at

***Index Terms*—Dataset Distillation, Knowledge Distillation, Model Compression, Efficient Deep Learning, Coreset, Multimodal Learning**

I. Problem Statement, Hypothesis, and Literature Survey

A. Problem Statement

The remarkable success of modern deep learning is intrinsically linked to the availability of massive datasets. However, this dependency has created a significant bottleneck: training models like Vision Transformers (ViTs) or large-scale ResNets on datasets such as COCO or ImageNet is a computationally intensive, time-consuming, and expensive process. This high barrier to entry limits the ability of researchers with constrained resources to participate in cutting-edge AI development. The core problem is, therefore, how to reduce the computational requirements of model training without catastrophically sacrificing performance.

B. Hypothesis

We hypothesize that the information within large datasets is highly redundant and can be "distilled" into a much smaller, yet highly informative, subset. Our primary hypotheses are:

1) A large dataset can be compressed into a core set representing 1-10% of its original size.
2) A model trained on this small, distilled dataset can achieve a significant fraction (e.g., 70-80%) of the test accuracy of a model trained on the full dataset.
3) The performance gap between models trained on full and distilled data can be further narrowed by employing Knowledge Distillation (KD).

C. Related Work

Dataset Distillation (DD) reduces large datasets into compact, informative subsets, enabling efficient training while preserving performance. While DD has been successful in unimodal tasks like image classification, its application to multimodal data, such as image-text pairs, remains largely unexplored. This gap is critical given the growing importance of multimodal models like CLIP and ALIGN, which rely on massive datasets. Addressing this requires new frameworks that preserve cross-modal relationships, scale effectively, and generalize across tasks. Below, we review unimodal DD, multimodal learning, and key research gaps.

1) Unimodal Dataset Distillation: The field of Dataset Distillation began with the foundational work of wang2020 who introduced the concept of synthesizing a small set of images and labels such that a model trained on this set would achieve performance comparable to training on the full dataset. Their key idea, Gradient Matching (GM), involves optimizing the synthetic data to match the gradients of a model's parameters computed on mini-batches of synthetic data with those computed on mini-batches of real data. This ensures the synthetic data points guide the model's parameter updates in a similar direction to the real data. Subsequent works have explored alternative matching criteria. zhao2021 proposed Dataset Condensation (a term sometimes used interchangeably with DD) focusing on aligning the feature distributions between real and synthetic data using metrics like Maximum Mean Discrepancy (MMD), demonstrating improved generalization across different network architectures. More recently, lee2022 advanced the state-of-the-art with Trajectory Matching (TM), which aims to match the entire trajectory of model parameters during training on synthetic data to that on real data, capturing finer details of the learning process.

These methods have been successful on benchmark image datasets like MNIST and CIFAR, demonstrating significant compression (e.g., condensing CIFAR-10 into 10-50 images).

2) Multimodal Learning and Efficiency: Simultaneously, significant progress has been made in multimodal representation learning, particularly in vision-language understanding. Models like CLIP radford2021 and ALIGN jia2021 utilize dual encoders (one for images, one for text) trained with contrastive objectives on massive datasets (millions or billions of image-text pairs from the web). The core idea is to pull embeddings of paired images and texts closer in a shared latent space while pushing away embeddings of unpaired data. This contrastive approach has proven highly effective at learning cross-modal alignment and generating powerful zero-shot capabilities. However, the reliance on gargantuan datasets for training these models underscores the scalability issue mentioned earlier. While distillation has been explored for other data types like text sucholutsky2021, focusing on distilling text classification datasets using techniques like matching soft labels, there is no established framework for distilling coupled multimodal datasets like image-text pairs, which require preserving both intra- and inter-modal relationships.

3) Research Gaps: Based on this survey, I identify the following key research gaps that this proposal aims to address:

- **Lack of Cross-Modal DD Frameworks:** To the best of my knowledge, no existing work proposes a general Dataset Distillation framework capable of synthesizing multimodal datasets (like image-text pairs) while explicitly preserving the crucial inter-modal semantic relationships and modality-specific features.
- **Scalability for Large Multimodal Data:** Applying existing unimodal DD methods, which are already computationally intensive, to the scale of multimodal datasets like COCO is a significant technical challenge that requires efficient optimization strategies.
- **Generalization of Synthetic Multimodal Data:** Ensuring that synthetic multimodal data generalizes well across diverse downstream tasks (retrieval, captioning) and different model architectures is critical for the utility of distilled multimodal datasets.

II. Methodology

A. Multimodal Distillation (for COCO)

The second method implements a loss-based dataset distillation strategy, where the goal is to identify and retain only the most informative samples from the full training dataset for subsequent model training. Unlike the earlier gradient-matching approach, this technique leverages the pre-trained teacher model's loss per sample to rank the difficulty of each training example.

1) Step-by-Step Explanation:

a) Pre-training the Teacher Model: Initially, a GPT2Vision-based vision-language model was trained on the full dataset to learn effective image captioning capabilities. This step ensured that the model developed a strong understanding of both visual features and language generation, which is crucial for meaningful loss estimation during distillation.

b) Loss Calculation for Distillation: After the model reached a satisfactory performance level (measured using validation metrics), it was used as a *teacher model* to compute the cross-entropy loss per sample across the entire training dataset. The loss values were calculated without updating the model parameters (`@torch.no_grad()`), ensuring stable and deterministic results.

c) Selection of Hard Examples: Each sample in the training set was paired with its corresponding loss value. These pairs were then sorted in descending order by loss — higher loss indicating harder or more challenging examples. A subset of the data was selected based on a `distillation_ratio`, such as 5% or 20%, effectively creating a distilled dataset containing only the most difficult samples.

d) Retraining on the Distilled Dataset: A new student model was trained from scratch using only the distilled dataset. Despite being significantly smaller (e.g., 5% of the original size), this dataset retained high-quality, informative samples that enabled the student model to achieve competitive performance compared to one trained on the full dataset.

The process is formalized in Algorithm 1 and consists of the following steps:

This methodology allows for the creation of a dense, information-rich dataset that retains the most critical learning signals from the original, much larger COCO training set. A new model is then trained from scratch exclusively on this distilled core-set.

Algorithm 1 Loss-Based Core-Set Selection for Data Distillation

Require: Full training dataset D_{full}, a pre-trained reference model M_{ref}, a distillation ratio r.
Ensure: A distilled dataset subset $D_{distilled}$.
function CreateDistilledDataset(D_{full}, M_{ref}, r)
  losses_with_indices $\leftarrow$ empty list
  Set M_{ref} to evaluation mode.
  for each sample (x_i, y_i) with index i in D_{full} **do**
    logits$_i \leftarrow M_{ref}(x_i)$ {Perform a forward pass}
    loss$_i \leftarrow$ CrossEntropyLoss(logits$_i, y_i$) {Calculate loss for the sample}
    Append (loss$_i, i$) to *losses_with_indices*
  end for
  Sort *losses_with_indices* in descending order based on loss values.
  $N_{total} \leftarrow$ size of D_{full}
  $N_{keep} \leftarrow \lfloor N_{total} \times r \rfloor$
  top_samples $\leftarrow$ the first N_{keep} elements from *losses_with_indices*
  distilled_indices $\leftarrow$ extract indices from *top_samples*
  $D_{distilled} \leftarrow$ CreateSubset(D_{full}, *distilled_indices*)
  return $D_{distilled}$
end function=0

B. Unimodal Distillation via Coreset Selection (for CIFAR-10)

For CIFAR-10, we implemented a Coreset selection algorithm:

1) **Train Teacher Model:** A ResNet-18 model was trained on the full 50,000-image CIFAR-10 training set.
2) **Extract Features:** We extracted feature vectors for every training image from the teacher's penultimate layer.
3) **Cluster Features & Select Prototypes:** For each class, we applied K-Means clustering with 'k=500'. The original image closest to each of the 500 cluster centroids was selected as a prototype.
4) **Create Coreset:** This process resulted in a 5,000-image coreset (500 images x 10 classes).

1) Knowledge Distillation (KD): To enhance training on the coreset, we applied KD. The student model's loss function combines standard cross-entropy ($\mathcal{L}_{\text{CE}}$) with a KL-Divergence loss ($\mathcal{L}_{\text{KL}}$) against the teacher's softened predictions:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{CE}} + (1-\alpha)T^2\mathcal{L}_{\text{KL}} \tag{1}$$

where temperature $T = 4.0$ and weighting factor $\alpha = 0.1$.

III. Datasets

- **COCO:** A benchmark dataset with over 250,000 images and corresponding captions.
- **CIFAR-10:** A standard classification dataset with 50,000 training and 10,000 testing 32x32 color images in 10 classes.

IV. Results

A. COCO Distillation

For first distillation strategy, a gradient matching-based dataset distillation method is implemented to synthesize a small set of images that capture the essential knowledge from real data by aligning the parameter gradients of a student model with those of a pre-trained teacher model. During local experimentation, however, it was observed that the student model failed to learn effective image captioning capabilities through this method alone with any data distillation ratio. As a result, a GPT2Vision-based architecture was trained separately to achieve better performance in vision-language tasks. Subsequently, an alternative distillation approach was applied, tailored specifically to transfer knowledge from the teacher to the student in a more structured and task-aware manner.

a) Results and Efficiency: Encouragingly, when distilled with just 5% of the data and trained for 20 epochs, the student model maintained a large portion of the original model's performance on the validation set. This result highlights the effectiveness of selecting hard examples through loss ranking and demonstrates that meaningful knowledge can be transferred using a much smaller, curated dataset. Results is given Table I

Table I
SECONDTH DISTILLATION STRATEGY DISTILLATION RATIO, EPOCH AND LOSS VALUES. %5 DISTILLATION AND 20 EPOCH IS BEST AND ENOUGH TO ACHIEVE OUR GOALS. WE USE 100K TRAINING SAMPLES FOR FAST EXPERIMENT.

Dis. Ratio	Epochs	Loss	Training Time
Full Dataset (100%)	20	3.0	68 min.
30% Distilled	10	3.5	30 min.
5% Distilled (20 epochs)	20	2.8	25 min.
5% Distilled (10 epochs)	10	3.2	13 min.

B. CIFAR-10 Results

Four controlled experiments were conducted, with results derived directly from our execution logs. The performance of each scenario on the test set is detailed in Table II.

Table II
CIFAR-10 TEST ACCURACY FOR DIFFERENT TRAINING SCENARIOS

Scenario	Training Data Details	Accuracy (%)
Teacher (Baseline)	Full Data (50k)	83.58
Student (No Aug)	Coreset (5k), No Augment	51.28
Student (With Aug)	Coreset (5k), With Augment	56.89
Student (KD)	**Coreset (5k), With Aug + KD**	**70.21**

The logs clearly show the impact of each technique. Training on the raw coreset results in poor performance (51.28%). Data augmentation provides a modest improvement (56.89%). The most significant gain comes from Knowledge Distillation, which boosts accuracy to 70.21%, demonstrating its effectiveness in compensating for the reduced data size.

The training efficiency gains are summarized in Table III.

Table III
TRAINING EFFICIENCY COMPARISON ON CIFAR-10

Scenario	Data Size	Time / Epoch	Speedup
Full Data (Teacher)	50,000	$\sim$7-8s	1x (Baseline)
Coreset (Student)	5,000	$\sim$1s	$\sim$**8x**

By reducing the data size by 90%, we achieved an approximately 8-fold speedup in per-epoch training time.

V. Discussion and Conclusions

This study investigated dataset distillation (DD) as a strategy to reduce training costs while preserving model performance across both multimodal and unimodal learning tasks. Our findings demonstrate that DD is a promising approach for compressing large-scale datasets into compact, informative subsets without significantly sacrificing accuracy.

For the multimodal setting on COCO, we employed a loss-based sample selection method using a pre-trained GPT2Vision-based teacher model. By selecting only the hardest examples — specifically 5% or 30% of the full dataset — we trained student models from scratch. Encouragingly, a distilled dataset containing just 5% of the original data, when trained for 220 epochs, preserved most of the teacher model's performance on validation metrics. This result highlights the

effectiveness of using per-sample loss values to identify high-information samples and demonstrates that meaningful knowledge can be transferred using a much smaller dataset.

On CIFAR-10, we implemented a coreset selection strategy based on K-Means clustering of penultimate layer features. The resulting 5,000-image coreset (a 90% reduction from the original dataset) alone led to a significant drop in performance (51.28% accuracy). However, combining this coreset with Knowledge Distillation (KD) significantly improved student model accuracy to 70.21%, retaining over 84% of the teacher model's performance (83.58%). Furthermore, this approach accelerated training by approximately 8x per epoch.

These results confirm our hypothesis: a small, well-curated dataset — especially when paired with KD — can train models to achieve competitive performance with a fraction of the computational cost. Importantly, the synergy between coreset selection and KD enables not only faster training but also better generalization through soft-label guidance from the teacher.

In conclusion, dataset distillation — whether applied via loss-based filtering for multimodal tasks or coreset + KD for unimodal classification — presents a powerful framework for efficient deep learning. This approach supports faster experimentation cycles, reduces energy consumption, and makes advanced AI training more accessible, particularly in resource-constrained environments.

REFERENCES

[1] T. Wang, J. Zhu, A. Torralba, and A. Efros, "Dataset distillation," *arXiv preprint arXiv:1811.10959*, 2018.

[2] G. C. G. Zhao and V. B. G. B. B., "Dataset Condensation with Gradient Matching," in *Proc. ICLR*, 2021.

[3] B. Kim, J. H. Bae, and M. K. Kim, "Dataset condensation with distribution matching," in *Proc. ICML*, 2022.

[4] S. A. Coleman, et al., "Selection via proxy: Efficient data selection for deep learning," in *Proc. ICLR*, 2020.

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[7] A. Dosovitskiy, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[8] T. Lin, et al., "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[9] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.