Team 23: Evaluating Self-Supervised ViT Pretraining on Synthetic vs. Real Data

Apoorva Batham
Immatrikulation: 7032927

Ahmad Hatam
Immatrikulation: 7048335

1. Task and Motivation

1.1. Task Statement and Definitions

We aim to evaluate the effectiveness of self-supervised pretraining using synthetic versus real-world data. Using DINO (self-distillation with no labels) with a ViT-Small backbone, we pretrain two models on synthetic and real versions of the same animal dataset (AFHQ). We then assess the learned representations on two downstream tasks:

- Image classification
- Image retrieval

Both tasks are performed on the Oxford-IIIT Pet dataset.

1.2. Motivation

While self-supervised learning (SSL) has advanced general-purpose representation learning, its application to synthetic data is still debated. Many studies evaluate only classification performance and ignore broader representational capabilities such as those tested via retrieval. Furthermore, synthetic vs. real comparisons often suffer from mismatch in resolution, dataset size, or class composition, making conclusions difficult to generalize. Our goal is to conduct a fair comparison with tightly controlled dataset characteristics and use a secondary task (retrieval) to probe the structure of learned representations.

1.3. Related Work

- Caron et al., introduced DINO, which learns robust visual representations using ViT without labels [1].
- Azizi et al. showed that synthetic data generated via diffusion models can rival real data for large-scale classification when training from scratch [2].
- Zhang et al. found that segmentation models trained on synthetic datasets are outperformed by those using retrieved real images; however, they did not evaluate retrieval tasks directly [3].

1.4. Challenges

- Designing a dual-task pipeline (classification + retrieval) within resource constraints
- Ensuring fairness by matching resolution, image count, and label distribution across synthetic and real datasets
- Measuring subtle differences in representation quality, not just top-line classification accuracy

2. Goals

2.1. Challenges Addressed

- Does SSL benefit equally from synthetic and real data given identical conditions?
- How well do the learned representations generalize across tasks and domains?

2.2. Mid-Term Goals

- Complete DINO + ViT-Small pretraining on synthetic and real AFHQ (15k images each)
- Prepare Oxford-IIIT Pet dataset (resize, class filtering if needed)
- Build evaluation pipelines for classification and image retrieval using frozen features

3. Methods

3.1. Models and Frameworks

- **Pretraining:** DINO (ViT-Small)
- **Downstream:**
  - Classification: Linear classifier on frozen features
  - Retrieval: Nearest neighbor search using cosine similarity

3.2. Justification

ViT-Small is computationally tractable and has been shown to produce quality representations in DINO. DINO's self-distillation encourages strong semantic grouping, which is ideal for both retrieval and classification. Including a retrieval task adds depth to the evaluation by testing representation geometry, not just decision boundaries.

3.3. Differences from Prior Work

- Inclusion of retrieval task as a diagnostic tool
- Real vs. synthetic data compared fairly using tightly controlled datasets (same classes, resolution, size)
- Pretraining and downstream datasets are different but related, adding a realistic domain transfer test

3.4. Compute Budget

- Pretraining time: ∼2–3 days per model on a single A100 or V100 GPU
- Downstream: Linear classifier and retrieval are lightweight

4. Datasets

4.1. Pretraining

- **Real:** AFHQ — 15,000 real animal images (cats, dogs, wild) at 256×256 resolution
- **Synthetic:** AFHQ (SDXL-generated) — 15,000 synthetic images matching resolution and classes

4.2. Downstream Evaluation

- **Dataset:** Oxford-IIIT Pet dataset (∼7,000 images across 37 pet breeds; cats and dogs)
- **Tasks:** Used for both classification and retrieval
- **Retrieval Setup:** Treated as a nearest-neighbor problem in embedding space

5. Evaluation

5.1. Metrics

- **Classification:** Top-1 accuracy (pet breed prediction)
- **Retrieval:** Mean Average Precision (mAP), using cosine similarity on embeddings

5.2. Evaluation Strategy

- Pretrain DINO on real and synthetic datasets
- Freeze the backbone
- Train a linear classifier for breed classification
- Perform retrieval by computing nearest neighbors over validation set embeddings
- Compare real- and synthetic-pretrained models on both tasks to measure performance gap

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, *Emerging Properties in Self-Supervised Vision Transformers*, In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 1

[2] S. Azizi, M. R. Abid, Y. Zhang, et al., *Synthetic Data from Diffusion Models Is All You Need for ImageNet Classification*, arXiv preprint arXiv:2306.00988, 2023. 1

[3] Y. Zhang, A. T. Ahuja, R. Kohli, and C. Guestrin, *Synthetic Data is Outperformed by Retrieved Real Images*, arXiv preprint arXiv:2404.07503, 2024. 1