OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Improving Fine-Grained Medication Recognition with DINOv2 Projection Head Retrieval

Abhinav Desai
ad5103@icloud.com

ABSTRACT

ABSTRACT

Medication errors continue to plague the healthcare field, stemming from the misidentification of medications in pharmacies and patient care settings. Medications that look alike are often difficult to distinguish from one another, as they share identical geometries and colors. Meanwhile, individual pills and tablets often contain asymmetric imprints and scoring marks that make medications with the same identity look visually different. To counter this, machine learning models have been utilized to predict the identities of medications using visual recognition of various attributes, including shape, color, texture, and size. However, many models fail to perform at a high level, especially given the challenges of the datasets on which they are trained. This paper introduces how attaching a projection head on top of a pre-trained Vision Transformer (DINOv2) and using a 120 epoch cosine-annealing schedule, a low learning rate, and a multi-loss function allowed for the projection head to learn a better retrieval embedding space. During training, the optimal model was selected based on the highest validation two-sided mean top-10 accuracy to reflect real-world grid interfaces and to blend the multi-dimensional vector of a medication's front and back faces to eliminate noise. This model was thus transformed into a high-performing retrieval system with strong test accuracies including a top-10 accuracy surpassing 97%. These findings suggest that projection head attachment and training on top of the DINOv2 backbone can support a high-precision retrieval system whose end goal may be to potentially reduce cognitive fatigue for pharmacists and patients as a decision-support system.

INTRODUCTION

Medication Errors (MEs) are preventable events that occur during the "dispensing, monitoring, product labeling, distribution, compounding, administration, nomenclature and packaging, education, and use [of medications]" and are a leading cause of avoidable injury and death in US hospitals (Mutair et al., 2021). One major contributor to these medication errors is the misidentification and mislabeling of medications, specifically pills and tablets. This distribution of pills and tablets in pharmacies and hospitals often is

performed by pharmacists or pharmacy technicians, who are required to maintain their focus in a high stress, fast paced environment. However, this manual pill and tablet identification often falls short when human errors occur, leading to serious consequences. These medication errors "affect approximately 1 in every 10 hospitalized patients, with nearly 7% of these errors resulting in fatalities," causing "exorbitant costs on healthcare systems" and the consumption of "valuable resources that could be directed towards improving patient care" (Coelho et al., 2024).

In the face of these challenges, the use of artificial intelligence (AI) and machine learning (ML) to identify medications has emerged as a potential solution in recent years. This approach trains machine learning models on existing medication image data in verified existing databases and tunes them to identify pills or tablets based on various features of the pills/tablets, including color, shape, text or numbers displayed, and size. Then, the final, fully trained machine learning model is used on real world images to predict the true identity of a given pill or tablet. However, existing machine learning models for pill and tablet identification often fail to detect the fine visual differences between pills and the multiple angle features of a given pill or tablet that distinguish it from other pills or tablets. This paper proposes a deep learning architecture that combines a vision transformer model (DINOv2) with a custom projection head. The objective was to evaluate if combining the fine tuning of a projection head with the supervised learning of a DINOv2 model would improve the top-k retrieval accuracy for pills and tablets. The evaluation of this model revealed that the specialized metric learning framework substantially improves the performance of a frozen DINOv2 baseline on a highly challenging pill identification database, pushing top-10 two-sided mean retrieval accuracy above 97%.

DATA

Over the course of building this model, data was used from the ePillID dataset, a computer vision benchmark intended to be used to recognize pill and tablet images in various conditions. This dataset contains data integrated from two National Institutes of Health (NIH) National Library of Medicine (NLM) databases - the NIH NLM Pill Image Recognition Challenge dataset and the NIH NLM Pillbox dataset. The compiled ePillID dataset contained a total 13,532 images. It was built around 4,902 unique pill and tablet types. Due to the visual differences in the front and back sides of a pill or tablet, two classes were present for each pill type, splitting the 4,902 pill types into 9,804 separate visual classes. For most of these 4,902 pill types, there existed only a single ideal reference image for each side, with only 960 pill types including images beyond these baseline reference photos.

To optimize pill and tablet recognition using this challenging dataset, the data was partitioned across a training, reference, and query distribution in accordance with the predefined split files in the original ePillID dataset benchmark (Usuyama et al., 2020). The training set contained 12,042 images used for the model to learn projection weights and pill and tablet features. This set excluded all images present in the validation and test sets. The reference set consisted of 9,804 images which acted as a gallery against which the unknown queries were matched. This set was created from all training images that were marked is_ref = True in the dataset. The query set consisted of 1,490 total images created from non-reference

images (is_ref = False) and was split into 2 subsets: the validation query set and the test query set. The validation query set contained 745 evaluation queries which represented previously unseen image inputs, which were used to monitor validation accuracy and loss. Lastly, the test query set contained 745 unique evaluation queries which were used to obtain the final test baseline retrieval accuracies.

Within the dataset, various lighting conditions, background settings, and distortions complicated consumer/query images, which allowed for a simulation of real world clinical inputs to predict against idealized reference images.

METHODS

This section covers the methods used over the course of the research. The methods of the research span from initializing pill labels to testing a frozen DINOv2 backbone to the inclusion of projection heads, their training, and their evaluation using the test query set.

1. Label Encoding

1.1 Overview

Each pill identity in the dataset was represented by a string label whose characters included letters, numbers, and dashes. In order to adapt these pill identities for the models used during this research, a Label Encoder was used to convert, or map, each of these string labels to a unique integer index. This label mapping was used for the training, validation, and evaluation of the model, as well as the final retrieval of reference images. Using this initial label mapping, retrieval scores were aggregated across the 4,902 pill identity labels.

2. Frozen DINOv2 Baseline Retrieval Model

2.1 Baseline Feature Extraction

As a first step, a pretrained DINOv2-large model was used as a visual backbone and a baseline evaluation. This pretrained model was adapted from the facebook/dinov2-large Vision Transformer (ViT) model, which was already trained on a "large collection of images in a self-supervised fashion" and had a frozen backbone without "any fine-tuned heads" (Oquab et al., 2024). After being adapted to the ePillID dataset, this backbone took on a feature dimension of 1,024 while the cached training, reference, and test query features took on shapes of 12,042x1024, 9,804x1,024, and 745x1,024, respectively. Then, each image in the training set was passed through the DINOv2 image processor and frozen backbone, yielding an extracted image embedding. These embeddings were then cached to the disk for training, reference, validation query, and test query sets. However, for minority classes that did not meet the requirement of 4 images per batch during PK sampling (detailed below), the raw images were instead augmented with random rotations and cropping before being passed through the DINOv2 backbone dynamically during

the construction of batches.

2.2 Label-Level Retrieval Scoring

After the processing of images and the extraction of their embeddings, the evaluation of how accurately the query images were matched to the corresponding pill identities occurred. Prior to the evaluation, the extracted query and reference embeddings were L2-normalized, which scaled each image vector to the value of 1, allowing for the equal comparison of the resulting vectors onto a single unit. This also allowed for the use of the metric cosine similarity, which was able to produce a numerical measure of similarity between image embeddings by calculating the angle between the image vectors, as illustrated in the equation below, in which A and B are the vectors.

$$cos(\theta) = \frac{A \bullet B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^{\,2}}\sqrt{\sum_{i=1}^{n} B_i^{\,2}}}$$

This cosine similarity metric was then computed between each query embedding and all reference embeddings, and for each pill label, the final evaluation score was defined by the maximum similarity between the query and any reference image that belonged to that pill label. However, as aforementioned, each pill type's reference image set consists of both a front and back image, meaning multiple reference images may correspond to the same pill identity. To account for this, reference-image scores were converted into label-level scores, allowing front and back pill images corresponding to the same pill to be categorized under the same label. This final pill label score computation resulted in a query by class score matrix that would be used for an evaluation of the model's performance.

3. DINOv2 Projection Head Model

3.1 Motivation for Using Projection Heads

While the frozen backbone of the DINOv2 model provided a baseline on how well a machine learning model could predict pill identities, the frozen embeddings of this backbone were limited to general-purpose visual features. Meanwhile, pill identification - especially using the low-shot, fine-grained images of the ePillID database - requires isolating the small differences in the imprint, shape, color, scoring, and texture of pills. Thus, it was decided to attach a small neural network module, or projection head, to the DINOv2 frozen backbone. The projection head was trained to modify the embeddings of the DINOv2 model into a metric space specific to pills. Thus, while the DINOv2 backbone remained frozen, the projection head attached to the backbone learned the specific embedding transformations needed to more accurately distinguish the fine features of pills.

3.2 Projection Head Architecture

Over the course of this attachment of a projection head to the DINOv2 frozen backbone, the architecture

of the projection head was made of several layers. First, the projection head began with the input, which consisted of cached DINOv2 embeddings with a dimension of 1,024 and acted as the general feature extractor. The input was followed by a Linear layer, which was a fully connected layer also with a dimension of 1,024 that allowed the model to adapt general DINOv2 features to the pill space. Next, the LayerNorm layer performed normalization and stabilized training of the resulting model by ensuring feature scaling. This was followed by GELU activation, which introduced nonlinearity by allowing the projection head to learn more complex transformations of the DINOv2 embeddings. Next, a Dropout layer was added with a dropout probability of 0.10 to randomly deactivate features of the model during training to reduce overfitting. The Dropout layer was followed by a second Linear layer compressing the original 1,024 dimension embedding into a 512 dimension embedding. The final embedding was L2 normalized to scale all pill image vectors, allowing the cosine similarity function to be performed.

Finally, ArcFace class weights were learned over the course of the projection head, allowing the resulting model to perform metric based retrieval by forcing a greater degree of separation between the 4,902 pill classes in the vector space.

4. Projection Head Training

4.1 Training Inputs

Once the architecture of the projection head was finalized, training of the projection head began. Instead of training the projection head on raw images, the cached DINOv2 embeddings were used. This was performed by extracting DINOv2 features once and saving them to the disk and then training the projection head on these cached training embeddings. This minimized computational cost, as the large DINOv2 frozen backbone did not need to be repeatedly run during training.

4.2 Balanced PK Batch Sampling

To improve the metric learning occurring during training, a PK Batch Sampler was used. This resulted in the selection of pill images for training using batches, where each batched sampled P pill classes and K examples per pill class, yielding a total batch size of $B = P \times K$. This batch sampling allowed each batch to contain positive pairs and negative pairs to train the model to accurately match positive pairs to the pill class. During the course of research, a sample size of $P = 12$ pill classes and $K = 4$ examples per class was used, producing batches of up to 48 samples. Because most of the 9,804 visual classes only contained a single baseline reference image, the sampler operated with replacement at the individual pill side level. To successfully sample the $K = 4$ examples per class for data scarce classes, the single available reference photo was duplicated within the batch, and the aforementioned data augmentations such as random rotation and cropping were applied independently to each duplicate image.

4.3 Training Configuration

After the initialization of training inputs and sampling methods, the projection heads were trained. The

training took place with the AdamW optimizer, with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-4}. The training proceeded for 120 epochs, with a schedule of cosine annealing and a gradient clipping norm of 1.0. Mixed precision training was used when CUDA was available. All training and evaluation experiments were performed on a system equipped with a NVIDIA Tesla T4 GPU and an Intel Xeon CPU. During training, the best training checkpoint was selected using validation two-sided mean top-10 accuracy (explained below)

5. Multi-Loss Training

5.1 Overview

The training of the projection head was performed by combining retrieval and metric learning loss. This occurred through four main channels of loss: cross entropy, ArcFace, supervised contrastive, and triplet loss. During training, cross entropy loss put a heavy penalty on confident but incorrect predictions of the model, encouraging correct pill retrieval. ArcFace loss maximized the distance between different classes and minimized the distance between identical classes in the embedding space using a margin of 0.25 with cosine similarity as the metric. To improve the projection head's learning of gradients, a scale of 30.0 was used to scale up the cosine values. Supervised contrastive loss used overall distance between batch pairs to pull embeddings from the same class closer together. The temperature - the variable affecting the uniformity of the similarity score distribution - was set to 0.08. Lastly, batch-hard triplet loss found the single pair representing the hardest positive and negative images to the anchor image, pushing same class embeddings closer than different class embeddings by a set margin of 0.16.

Thus, cross entropy and ArcFace losses were implemented in order to enforce class level separation across global features (Terven et al., 2025). Simultaneously, supervised contrastive and triplet loss were utilized to enhance the fine grained features of the embedding space, allowing for metric visual retrieval. These different losses collectively trained the projection head to create an embedding space that was adapted to retrieval during the evaluation of the full model's performance. The relative weights for each of these losses were chosen based on preliminary experiments on the validation query set, allowing the model to balance broad class separation with fine grained visual details of medication images. Total loss was computed as follows, where $\mathcal{L}$ represents the loss value for each type of loss listed above:

$$\text{Total Loss} = (0.5 \times \mathcal{L}_{\text{CrossEntropy}}) + (0.3 \times \mathcal{L}_{\text{ArcFace}}) + (0.7 \times \mathcal{L}_{\text{SuperContrastive}}) + (0.7 \times \mathcal{L}_{\text{Triplet}})$$

6. Retrieval Evaluation

6.1 Score Matrix Construction

After training was complete, the frozen DINOv2 backbone with the attached projection head was evaluated. First, images in both the reference and test query sets were passed through the trained projection head. The resulting projected embeddings were L2 normalized. Then, cosine similarity was

computed between the projected embeddings of the test query and reference set images, and these similarities were reduced to label-level scores using the maximum reference similarity for each pill class. The matrix containing each of these scores was used for final top-k evaluation.

6.2 Single-Sided Evaluation

The first evaluation method was single-sided evaluation, in which each image from the test query set was evaluated independently. This meant the front-side and back-side query images were treated as separate query examples. Thus, a prediction by the model was counted correct for top-k only if the true identity of the pill appeared within the top k ranked labels for that individual test query image.

6.3 Two-Sided Evaluation

The second method of evaluation was two-sided evaluation. In this method, the front and back test query images that corresponded to the same pill identity were paired. Then, the model returned a label-score vector for each side, and the scores for the paired sides were combined before computing final top-k accuracy. Furthermore, mean aggregation averaged these front and back score vectors while max aggregation used the stronger score for each class across the two sides. Both of these techniques corresponded with a distinct evaluation metric. Overall, this two-sided method of evaluation more accurately represented a real-world setting in which both sides of the pill are likely available to use.

7. Top-K Accuracy Metrics

7.1 Top-K Definition

For each test query image, the model ranked the 4,902 pill identities present in the ePillID dataset based on top-k accuracy. Top-k accuracy was defined as the proportion of queries for which the true pill identity appeared within the top k ranked predictions. This was the primary evaluation metric for the model as it measured if the true pill identity was present in the highest scoring predicted labels. While the model was evaluated for top-1, top-3, top-5, top-10, top-20, and top-50 accuracies, the top-5 and top-10 accuracies were emphasized as the pill recognition model was intended to be used for retrieval or decision support.

8. Model Selection and Final Evaluation

8.1 Checkpoint Selection

After each epoch during training, the projection head was evaluated on the validation query set images. The checkpoint with the highest validation two-sided mean top-10 accuracy was saved, preventing the selection of the final model from test performance.

8.2 Test Set Evaluation

Lastly, an evaluation on the test query set was performed. The best saved projection head was loaded, and projected reference and test query embeddings were computed. Then, label-level retrieval scores were generated, and single-sided and two-sided top-k metrics were calculated for 745 test query evaluation images. The final reported model was the validation based checkpoint model consisting of the DINOv2 frozen backbone and projection head.

RESULTS

This section summarizes the results of the model evaluation. This includes the comparison of the DINOv2 backbone model's top-k accuracies with those of the backbone attached to the projection head, as well as the final metrics of retrieval performance.

The initial evaluation of the standalone DINOv2 frozen backbone without the projection head is presented in Table 1.

Accuracy Type	Single-sided	Two-sided	
		Mean Aggregated	Max Aggregated
Top-1	26.17%	37.06%	33.23%
Top-3	41.21%	54.47%	49.20%
Top-5	46.85%	62.94%	54.95%
Top-10	55.84%	72.20%	63.42%
Top-20	64.56%	80.03%	73.48%
Top-50	75.84%	89.62%	83.39%

Table 1: Top-K Accuracies for DINOv2 Frozen Backbone Model

Then, the projection head was attached to this backbone and trained as outlined in the Methods section over 120 epochs using a cosine annealing schedule and a learning rate of 3×10^{-4}. After the optimal projection head was selected using validation two-sided mean-aggregated top-10 accuracy, the evaluation of the final model took place. The results of this evaluation are shown in Table 2.

Accuracy Type	Single-sided	Two-sided

		Mean Aggregated	Max Aggregated
Top-1	52.75%	66.77%	61.02%
Top-3	75.84%	86.42%	84.66%
Top-5	82.68%	93.45%	90.58%
Top-10	90.47%	97.28%	95.53%
Top-20	94.77%	98.88%	98.40%
Top-50	98.79%	100.00%	100.00%

Table 2: Top-K Accuracies for DINOv2 Backbone Attached to Projection Head

Table 2 shows that the DINOv2 backbone attached to the projection head outperformed the backbone alone in all three aggregation techniques (single side, two-sided mean aggregated, and two-sided max aggregated), and across all top-k accuracy types. Notably, the combined model outperformed the standalone backbone's two-sided mean-aggregated accuracies. Top-1 two-sided mean-aggregated accuracy increased from 37.06% to 66.77%, a gain of 29.71 percentage points, or an 80.17% relative improvement. Similarly, the top-5 and top-10 accuracies experienced a 48.47% and 34.74% relative improvement, respectively.

The changes in the cross entropy, ArcFace, supervised contrastive, and triplet losses of the model were also recorded over the 120 epochs the combined model underwent during training. The results are shown in Figure 1.

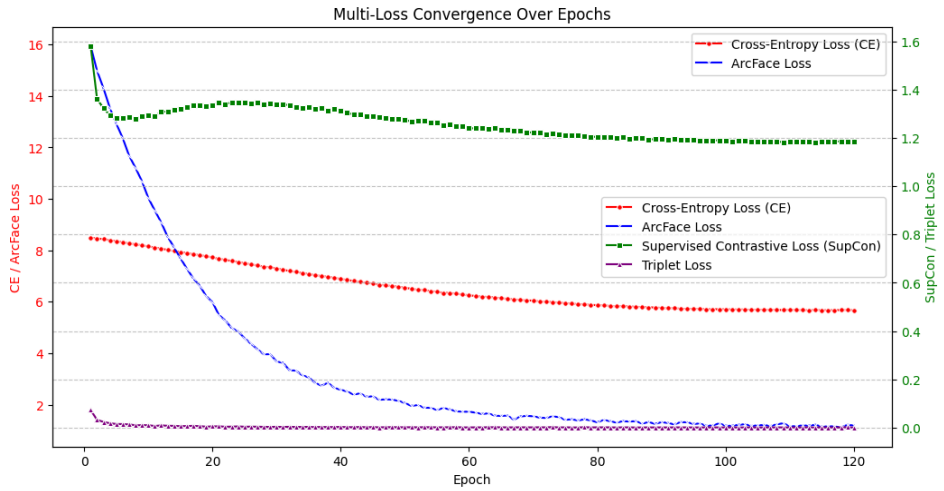

Figure 1: Training Epoch Trends in Cross-Entropy, ArcFace, SupCon, and Triplet Loss

Figure 1 shows that each type of loss plateaued near epoch 80-90, and that each loss behaved differently while still adhering to the overall decrease. Cross entropy and triplet loss both experienced a slight decline over the course of training, with triplet loss specifically smoothly decaying toward zero. Meanwhile, supervised contrastive loss was initially unstable before beginning a permanent decline around epoch 30. Lastly, ArcFace loss experienced a sharp decrease in the first 40 epochs before slowly decaying toward zero.

Next, the single query top-k accuracy and two-sided mean top-k accuracy were graphed using the value of k as the independent variable. This showed how the model performed as the search window for pill images was widened from an absolute match (top-1) to a meaningful result (top-5) to the extensive result bank of a large system (top-50). The results are shown below in Figure 2.

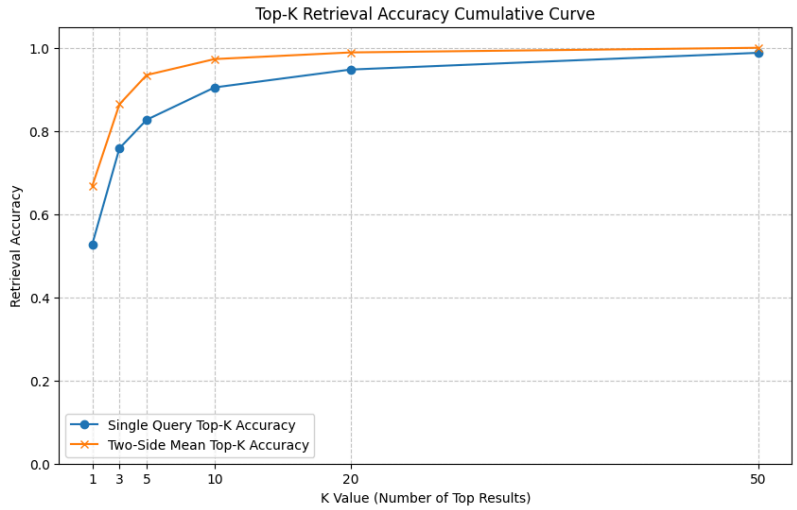

Figure 2: Changes in Single Query vs Two-Sided Mean Top-K Accuracy as a Function of K

As illustrated in Figure 2, the combined model's performance using aggregated dual image testing was always greater than when only single query images were used. This gap is especially seen at the k values of 1, 5, and 10, where the two-sided mean top-k accuracy outperformed the single query top-k accuracy by a relative improvement of +26.58%, +13.03%, and +7.53%.

The aggregation techniques of single side, two-sided mean, and two-sided max aggregation were also graphed for the accuracy types of top-1, top-5, and top-10. A direct grouped comparison of these three distinct types of handling pill queries is shown below in Figure 3. All of these three types are represented in each set of bars, with the k value plotted on the horizontal axis to show accuracy in each aggregation technique as a function of k.

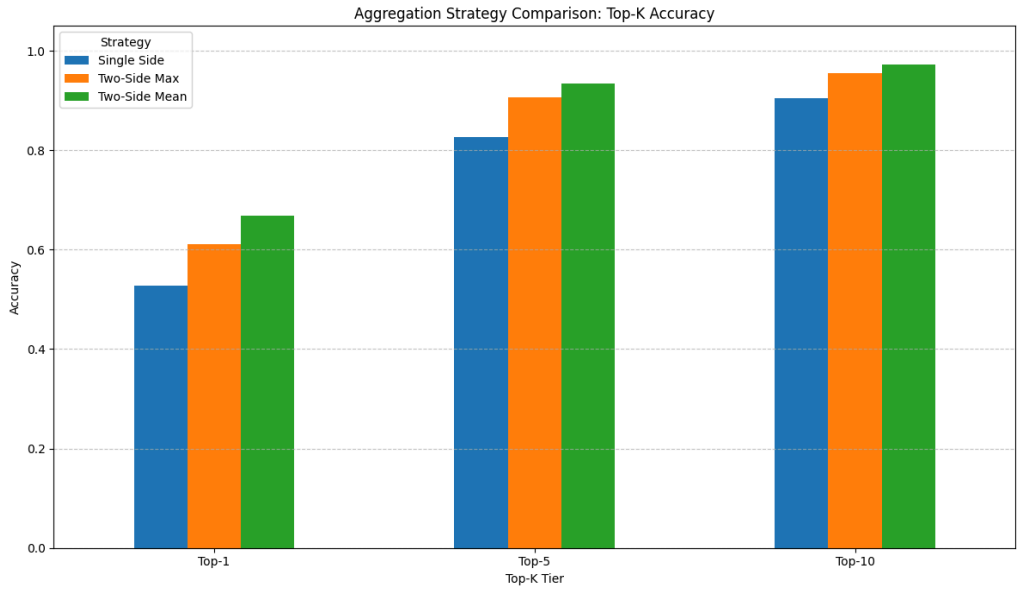

Figure 3: Grouped Comparison of Single-Sided, Two-Sided Max, and Two-Sided Mean Aggregation

As already represented in Figure 2, two-sided mean aggregation strongly outperformed the single-sided aggregation technique across all tiers of top-k. However, although two-sided mean aggregation still outperformed the two-sided max aggregation technique across all tiers, the gap was much smaller, boosting accuracy by +9.42%, +3.17%, and +1.83% at the top-1, top-5, and top-10 tiers, respectively.

In light of these results, the model was run as a local application on CPU. This produced strong results, as the local run of the application was able to achieve sub-second inference speed.

DISCUSSION

This research investigated how attaching a projection head to a pre-existing, pre-trained DINOv2 frozen backbone could greatly boost accuracy metrics, which would lead to more effective pill and tablet recognition. Initially, a test run of the DINOv2 frozen backbone revealed substantially lower baseline performance - including top-1, top-5, and top-10 accuracies of 37.06%, 62.94%, and 72.20%, respectively.

Related Work

This low baseline behavior was contextualized by Usuyama et al. in the original ePillID dataset study, where the “plain classification” method using a standard ResNet-18 architecture collapsed, yielding a top-1 accuracy of 33.13% ± 2.34% (Usuyama et al., 2020). Even after scaling up to the massive classification networks combined with specialized additions in the form of DenseNet161 BCP and ResNet152 B-CNN, top-1 global accuracy remained relatively low at 48.72% ± 1.92% and 55.14% ±

3.67%, respectively. To boost performance on known classes, Usuyama et al. moved to multi-head metric learning, achieving strong top-1 global accuracies of 88.86% ± 2.88% and 89.61% ± 1.83% for DenseNet161 BCP and ResNet152 B-CNN, respectively. However, although these setups achieved high benchmarks, they required end to end training of millions of hyperparameters from scratch and created feature matrices of over 250,000 dimensions. Attempts were made to mitigate these issues, such as Zeng et al.'s engineering of multi CNN (Convolutional Neural Network) pipelines to train independent structures for color, shape, and imprints. Despite this, the underlying model "contains about 40 million parameters and consumes over 3.2 billion FLOPS," requiring aggressive compression frameworks to shrink the architecture into a manageable size for deployment (Zeng et al., 2017).

Motivation for Multi-Loss Structure

To reduce the computational cost of full-network retraining and the fragile nature of multi model pipelines, this research introduced a single model alternative using DINOv2. Despite this, a gap still existed between the high potential of the DINOv2 backbone and the challenging conditions of the ePillID dataset. To bridge this gap, a multi loss structure was used to train a projection head on top of this pre-trained backbone. As outlined in the Methods section, this loss structure was composed of cross entropy, ArcFace, supervised contrastive, and triplet losses, where each loss interacted with the others. While the cross entropy and ArcFace losses solidified the boundaries of pill retrieval, supervised contrastive and triplet loss forced variants of pills in the same class towards each other and similar looking negatives away.

Justification for Selected Accuracy Type

During the training of the model, validation accuracy metrics were monitored in order to select the projection head most likely to yield the highest accuracy on the test query set. However, various accuracy metrics were available, including various accuracy types (top-1, top-3, top-5, top-10, top-20, and top-50) and aggregation techniques (single-sided, two-sided mean, two-sided max). Of these, two-sided mean top-10 accuracy was selected as the checkpoint criteria due to its value in creating an easily scannable grid to support the recognition of medications. Meanwhile, two-sided mean aggregation was selected as the aggregation technique as it mathematically averaged the 512-dimensional vectors of both sides of a medication, which is effective for the filtering out of noise, lighting differences, and asymmetric imprints.

Final Outcome and Application

The final run of the combined model consisted of testing the model on 745 previously unseen test query images. This combined model, which was chosen using validation accuracy based checkpointing, yielded final two-sided mean aggregated accuracies of 66.77%, 86.42%, 93.45%, and 97.28% for top-1, top-3, top-5, and top-10 accuracies, respectively, showing strong improvements over the corresponding accuracies of the standalone backbone based model.

Thus, in a real world medication verification system, the top medication match will be the true medication

over ⅔ of the time. Additionally, the final top-10 accuracy surpassing 97% means that a pharmacist or patient is highly likely to have the correct match visible on the screen in an easily scannable grid. In practice, when a user scans an unknown medication image using their device, the system would handle the prediction by sorting the reference image database and rendering the top 5 or 10 top matches using cosine similarity onto a digital interface. These results would each be accompanied by the clinical metadata, allowing the pharmacist or patient to visually compare the unknown medication with the top results displayed on screen to confirm the medication's identity. Lastly, the top-5 accuracy resting comfortably above 90% shows that if utilized as a decision support system or application, the system will, in practice, reduce cognitive strain due to the reduced number of candidates for review.

CONCLUSION

This research focused on whether the use of a vision transformer model (DINOv2) could be improved upon to create a more accurate machine learning retrieval of medications. The research relied on the ePillID dataset, whose lack of query images, background and imprint variations, and wide range of lighting conditions created a highly challenging setting for medication recognition. However, in the results, the attachment of a projection head to the DINOv2 frozen backbone greatly boosted accuracy, demonstrating the impact of improving DINOv2 to identify medications.

Key Achievements

While direct comparison to prior ePillID systems is difficult because of differences in architecture, metrics, and evaluation setup, the proposed method shows that strong retrieval performances can be achieved without end-to-end retraining of an incredibly large model size. This was demonstrated as the attachment of the projection head on the DINOv2 backbone transformed the previously inadequate baseline accuracies into strong two-sided mean aggregated accuracies, including the 66.77%, 93.45%, and 97.28% for top-1, top-5, and top-10 respectively. The combination of a low learning rate with a 120-epoch training budget with cosine annealing and two-sided mean aggregation allowed the model to produce strong results in visual medication retrieval. Not only this, using the top-10 validation metric to evaluate and utilize the model allows for the decision support for a pharmacist, while using the top-5 metric decreases the cognitive strain on vulnerable populations while fostering independence. Furthermore, the strong results produced by the initial test application confirms the model's computational efficiency and potential to be fine tuned into a deployable application.

Limitations

The results of this research suggest that the model may be useful as a decision-support retrieval system; however, it faces key limitations. First, this model contains various features, including the projection head, each of the four types of losses, PK batch sampling, cosine annealing, and the two-sided aggregation method. This variety of features makes it difficult to isolate which component or components were the most important in driving the successful performance of the model. In the future, an ablation

study must be performed, as this would systematically remove or alter each of these components to mitigate this lack of clarity.

Secondly, and more significantly, this model faces several limitations in a real-world setting. The system was evaluated on a benchmark dataset, which although offered various lighting and background conditions, was not truly equivalent to a live hospital setting. In a real-world clinical setting, the system may be more sensitive to image quality, occlusion, shadows, broken tablets, packaging, or unusual camera angles, disrupting the high performance achieved on the ePillID dataset images. Additionally, the ePillID dataset is a closed set, meaning that all candidate pill and tablet identities are known beforehand, and that the model cannot detect whether a medication is outside this reference database. Lastly, the model can only be used as a decision support system in a clinical setting, as retrieval relies on top-k ranking, which requires a human to visually select the correct medication. Thus, further validation under clinical imaging conditions and with human-in-the-loop workflows is needed before clinical deployment.

Future Steps

While the proposed final model consisting of the DINOv2 backbone and the projection head greatly boosts retrieval performance, various future steps can be taken to transform this model into a clinically ready system. First, an end to end training structure could be implemented, shifting the frozen DINOv2 embeddings into an active training pipeline. This would allow for updates to the gradients of the DINOv2 backbone to flow to the initial layers of the vision transformer instead of stopping before the projection head, potentially boosting accuracy. Not only this, the model could be expanded to multi-view dynamics, which would include harnessing the two-sided mean aggregation to process varying numbers of images per medication. This would make the model adaptable to a clinical setting, as degraded tablets, shifting lighting angles, and reflections on capsules are some of the most common challenges (My et al., 2025). Other techniques, such as Optical Character Recognition (OCR) and imprint parsing, could also be used to explicitly extract the imprints on pills and tablets and verify the medication's name, dosage, and effects using metadata, also potentially boosting accuracy (Jo et al., 2026).

Overall, the use of machine learning for medication retrieval remains an intriguing field, and further advances in this field must be accompanied by a mindset geared towards both efficiency in the pharmacy and sensitivity to the needs of vulnerable populations at home.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my mentor, Ronil Synghal, for his invaluable guidance that allowed me to pursue this project and his support throughout my research.

REFERENCES

Coelho, F., Furtado, L., Mendonça, N., Soares, H., Duarte, H., Costeira, C., Santos, C., & Sousa, J. P. (2024). Predisposing factors to medication errors by nurses and prevention strategies: A scoping review of recent literature. *Nursing Reports, 14*(3), 1553-1569. https://doi.org/10.3390/nursrep14030117

Jo, J., Yoon, S., Cho, J. (2026, January 21). GO-PILL: A geometry-aware OCR pipeline for reliable recognition of debossed and curved pill imprints. *Mathematics, 14*(2), 356. https://doi.org/10.3390/math14020356

Mutair, A. A., Alhumaid, S., Shamsan, A., Zaidi, A. R. Z., Mohaini, M. A., Al Mutairi, A., Rabaan, A. A., Awad, M., & Al-Omari, A. (2021). The effective strategies to avoid medication errors and improving reporting systems. *Medicines, 8*(9), 46. https://doi.org/10.3390/medicines8090046

My, L., Le, V., Vo, T., Hoang, V. (2025, March 6). A comprehensive review of pill image recognition. *Computers, Materials & Continua, 82*(3), 3693-3740. https://doi.org/10.32604/cmc.2025.060793

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P. Y., Li, S. W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., & Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2304.07193

Tariq, R., Vashisht, R., Sinha, A., & Scherbak, Y. (2024, February 12). Medication dispensing errors and prevention. *StatPearls*. https://www.ncbi.nlm.nih.gov/books/NBK519065/

Terven, J., Cordova-Esparza, D., Romero-González, J., Ramírez-Pedraza, A., Chávez-Urbiola, E. (2025, April 11). A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review, 58*(195), 78. https://doi.org/10.1007/s10462-025-11198-7

Usuyama, N., Delgado, N., & Hall, A. (2020). ePillID dataset: A low-shot fine-grained benchmark for pill identification. *arXiv.* https://doi.org/10.48550/arXiv.2005.14288

Zeng, X., Cao, K., Zhang, M. (2017, June 16). MobileDeepPill: A small-footprint mobile deep learning system for recognizing unconstrained pill images. *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services,* 56-68. https://doi.org/10.1145/3081333.3081336