# Few-Shot Learning for Fine-Grained Pill Classification

**Michael Gee**
Department of Computer Science
Carnegie Mellon University
mgee3@andrew.cmu.edu

**Ruhan Prasad**
Carnegie Mellon University
Department of Computer Science
rsprasad@andrew.cmu.edu

## 1 Introduction

**Dated 03 December 2021. Final project completed as a part of 10417/10617 - Intermediate Deep Learning.**

This project investigates the ePillID Dataset which consists of 13,000 pill images and 9804 appearance classes (front and back sides of 4,902 pill types) [27]. Many of these pills look the same, differing only by a small inscription. Moreover, for many pills only one image exists of each side. In this few and sometimes one-shot learning setting, we create an attention-based multi-headed architecture that achieves considerable fine-grained classification performance on the dataset. We come close to state-of-the-art performance utilizing attention-based techniques in an area of research currently dominated by ResNet architectures.

## 2 Background

### 2.1 Midway Report

Previously, we investigated the performance of the baseline model, a ResNet-152 with Compact Bilinear Pooling and a multi-head learning metric comprised of the weighted sum of several loss functions to learn an embedded space optimized for fine-grained classification.

The loss function used in this method is $L_{\text{final}} = \lambda_{SCE} L_{SCE} + \lambda_\eta L_\eta + \lambda_\rho L_\rho + \lambda_\Gamma L_\Gamma$. Where $L_{SCE}$ is the softmax cross-entropy loss, $L_\eta$ is cosine-softmax loss, $L_\rho$ is triplet loss, and $L_\Gamma$ is contrastive loss. In the benchmark model, the values for the loss weights $\lambda_{SCE}, \lambda_\eta, \lambda_\rho, \lambda_\Gamma$ were determined by doing cross validation on a ResNet-50 model to find favorable weight values.

### 2.2 Conclusions from Midway Report

Qualitatively, it was determined that the model struggles to distinguish pills of both similar shape and color where imprinted letters and numbers on the pill are the only distinguishing factor. We determine that in order to improve upon the benchmark, we must investigate an OCR multi-modal solution. Furthermore, we acknowledge the success of transformer architectures in image classification and determine that transformer architectures should also be investigated.

## 3 Related Work

**Pill Classification.** In 2016, the NIH held a pill image recognition challenge with a dataset of 1000 pills [38]. The challenge winner [36] utilized a deep-similarity model [31]. Following the competition, several papers have reported classification performance [7, 35], utilizing convolutional neural net architectures. Aside from convolutional neural net architectures, feature engineering approaches have also been successful, including Hu moment [1], color histogram decomposition [39], and recognition of text on pills [19].

**Few-Shot Learning.** Few-Shot Learning is the problem of making predictions based on a limited number of samples (1-2 samples in the case of the ePillID dataset). Neural network architectures like Siamese Convolutional Neural Networks [14] have been successfully used for few-shot learning. However, most recent state of the art performance [33] has been achieved by meta-learning [2]. Meta-learning is a class of models that become better at learning with more experience by learning at two levels: at a task-level where the base-model is required to acquire task-specific knowledge rapidly and at the meta-level

where the meta-model is required to slowly learn across-task knowledge [2]. It has most notably achieving state of the art performance for 1-shot learning on the Omniglot dataset [18, 20, 2].

**Fine-Grained Visual Classification.** Fine-grained visual categorization (FGVC) focuses on distinguishing subtle visual differences within a single type of image [40]; for example, classification of similarly shaped pills [27]. Approaches typically utilized for FGVC tasks [34] include Compact Bilinear Pooling (CBP) [12] and Metric Learning [26]. With the advent of attention [29] and Vision Transformers (ViT) [10], transformer architectures, such as TransFG [16], have been proposed for fine-grained classification and achieved state of the art performance for fine-grained classification on the NABirds dataset [28].

**Attention & Vision Transformers.** Self-attention and transformer architectures [30] have facilitated research in natural language processing (NLP) and machine translation, in many cases achiveing state of the art performance [6, 8]. Following its success in NLP, researchers have endeavored to apply transformers to computer vision. Initially, this materialized in the form of a transformer learning sequential embeddings extracted by a CNN [13]. Following this, the transformer architecture has been utlized in computer vision applications of object detection [5], image segmentation [37], and object tracking [25]. Most recently, the advent of ViT [11] has spurred many pure transformer models that achieve state of the art performance on tasks such as image classification [41], object detection [17], and fine-grained classification. In this paper, among other approaches, we attempt to extend the work done by He et. al which achieves state of the art performance for fine-grained classification [16].

# 4 Methods

In this section, we first elaborate on our over-arching experimental method, including how we consistently split the data across all experiments. We then discuss each model architecture that we experiment on in detail; in particular, the Benchmark ResNet Model, simple ensemble classifers, attempts at OCR-based models, and finally attention-based transformer models.

## 4.1 Experimental Method

The 3728 consumer images (for 1920 pill classes, two sides for 960 pill types) from the ePillID dataset were split into train (80%) and test (20%) sets with each pill class belonging to only one test set. The train set was then further split on pill type for 4-fold cross-validation. All 9804 reference images (two-sided for 4902 pill classes) are available during training. However, only the train set of consumer images is available during training. All models are trained on an NVIDIA T4 GPU 16 GB GDDR6.

## 4.2 Evaluation Metrics

Given some double-sided image query containing an image of both the front and back of a pill, the model returns a confidence score for each pill class.

When evaluating performance, we consider the following:

**1. Mean Average Precision (MAP)**. The average precision score is calculated separately for each query and then the mean is found across the dataset. MAP is utilized to measure the ability of the model to predict the correct pill type given queries [27]. **2. Global Average Precision (GAP)**. The precision score for each query is treated independently and averaged globally. GAP is utilized to measure the ranking performance [27] and the ability to use some common threshold across the model when identifying pill class [22]. **3. MAP@1 and GAP@1**. Consistent with the benchmark [27], we also consider MAP@1 and GAP@1 whereby only the top predicted pill class from each query is considered.

## 4.3 Benchmark Model: ResNet152 CBP

The current state-of-the-art model on the ePillID dataset is ResNet152 CBP, a 152-layer residual neural network pretrained on the ImageNet dataset. This model also uses a technique called compact bilinear pooling to capture higher order interactions between different feature channels. This model performs remarkably well and achieves a MAP@1 and GAP@1 of 92.01% and 91.19% respectively. This served as the baseline on which to improve off in this project.

## 4.4 Ensemble Classification Methods

A simple, but often effective, method of improving classification accuracy is through ensemble averaging. Ensemble averaging is a technique in which outputs or predictions from multiple machine learning models are combined to potentially make a more accurate prediction. [15][9]

**Model Averaging.** In this method, the confidence scores from two or more different machine learning models are averaged with equal weighting to make a prediction. Though a naive and limited combining scheme, there was a possibility of this

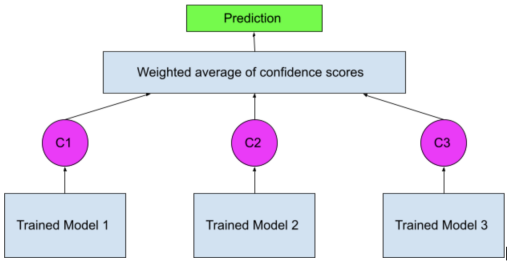


Figure 1: **Weighted Ensemble Averaging with Three Models**. After training each of the three classifiers separately, the confidence scores $C_1, C_2, C_3$ for a particular query is computed and are combined via a weighted average of the form $w_1C_1 + w_2C_2 + w_3C_3$ where each weight is non-negative and all weights sum to 1. The averaged confidence score is then used to make a prediction

method performing well, as there were several different deep learning architectures mentioned in the paper that achieve high accuracy on the ePillID dataset, and we observed during testing that on $94\%$ of queries, at least one of the top 5 best models outputed a correct prediction compared to $89\%$ for the best single model.

**Weighted Model Averaging.** Rather than taking a simple arithmetic mean of confidence scores, this method involves taking a weighted average of confidence scores from two or more different classifiers as outlined in Figure 1. This is often a more powerful ensemble averaging method as some models perform more accurately than others and deserve to be weighted higher in the averaging step. Weights for each model in the ensemble were decided based on performance in 4-fold cross-validation on the ePillID dataset. We then evaluated performance of the ensembles on the test dataset (avoiding fine tuning the weights on the test dataset as this would constitute "training" on the test data).

### 4.5 Optical Character Recognition

As mentioned in the midway report and in the original paper, optical character recognition (OCR) can be utilized to improve fine-grained visual classification performance [4]. However, authors of the paper caution that current OCR models would likely perform unreliably on the pills due to unpredictable orientation of the images, varying textures and backgrounds, and poor contrast of the text. Nonetheless, we applied two of the best open source OCR models on the ePillID dataset; Tesseract and EasyOCR. With one of these OCR models, we hoped to create an ensemble classifier that combined OCR outputs with the outputs of a standard image classifier to make a more powerful prediction. See Figure 2 for a high level architecture we proposed inspired by Bai et al's paper in scene text recognition. However, as we will later discuss, the open source character recognition models attempted were too unreliable to have been helpful in an ensemble classifier. Tesseract is an open source OCR engine developed by Google that uses a recurrent neural network architecture called long short-term memory architecture. While known to perform well on document text recognition, it is slightly less accurate on scene-text recognition tasks (text recognition in complex backgrounds). EasyOCR is a less popular but still robust open source OCR engine that uses the CRAFT (Character-Region Awareness For Text detection) model from Pytorch. This OCR model performs better than Tesseract on images with noisy backgrounds.

We examined the accuracy of these two models by running all reference and consumer pill images through them and performing side by side comparisons of pills and their predicted text. Since the pill texts were unlabeled in the dataset, performance of the OCR models were assessed qualitatively by manually scanning and comparing pill images with their predicted text.

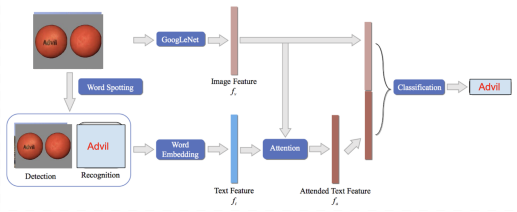


Figure 2: **High-level method of integrating OCR with an image classifier**

### 4.6 Warm-Started Dual Transformer Architecture for Few-Shot, Fine-Grained Classifcation.

We shall briefly review vision transformers and how they are extended for fine-grained classification by He et. al's proposed TransFG architecture [16] in Section 4.5.1. We shall then discuss the process through which we extended these architectures to fine-grained, few shot pill classification.

#### 4.6.1 ViT and TransFG

The standard ViT (Dosovitskiy et. al) splits an image into fixed-size patches, linearly embeds each of them (flattens and multiplies with a pre-trained or selected embedding matrix), adds position embeddings, and feeds the resultant sequence of vectors into a standard Transformer encoder originally outlined by Vaswani et. al [11, 30]. Position embeddings utilized by ViT are the absolute positional encoding originally proposed by Vaswani et. al's standard Transformer architecture [30].

He et. al extend ViT in their proposed TransFG architecture by proposing the following:
**1. Image Sequentialization**. The original ViT architecture splits the image into overlapping patches $x_p$. This degrades the local neighboring structures, especially when discriminating regions are split. Instead, He et. al propose a sliding window approach whereby each two adjacent patches share an overlapping area of a fixed size, better preserving local information. Specifically, for some image of size $H \times W$, image patch size $P$ and step size of the sliding window $S$, the input image will be split into $N$ patches where:

$$N = N_H \times N_W = \lfloor \frac{H - P + S}{S} \rfloor \times \lfloor \frac{W - P + S}{S} \rfloor \tag{1}$$

**2. Patch embedding**. Vectorized patches $x_p$ are mapped into a latent D-dimensional embedding space using a trainable linear projection. A learnable position embedding is added to the patch embeddings to retain their positional information. This is demonstrated in Equation 2 where $N$ is the number of image patches, $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \cdot D}$ is the patch embedding projection and $\mathbf{E}_{pos} \in \mathbb{R}^{N \cdot D}$ is the position embedding.

$$\mathbf{z_0} = (x_p^1\mathbf{E}, x_p^2\mathbf{E}, ..., x_p^N\mathbf{E}) + \mathbf{E}_{pos} \tag{2}$$

**3. Part Selection Module**. To fully exploit information obtained via attention, they alter the input to the last Transformer Layer. Suppose the model has $L$ layers and $K$ self-attention heads and the hidden features input to the last layer is denoted as $\mathbf{z_{L-1}} = (z_{L-1}^0; z_{L-1}^1, z_{L-1}^2, ..., z_{L-1}^N)$. The attention weights of the previous layers can be written as:

$$\mathbf{a_l} = (a_l^0, a_l^1, ..., a_l^K) \qquad \mathbf{a_l}^i = (a_l^{i_0}, a_l^{i_1}, ..., a_l^{i_K}) \qquad l \in 1, 2, ..., L-1 \tag{3}$$

Previous works suggested that raw attention weights do not necessarily correspond to the relative importance of input tokens due to lack of token identifiability of the embeddings. So, He et. al recursively apply matrix multiplication to the raw attention weights of all previous layers:

$$\mathbf{a_{final}} = \prod_{i=0}^{L-1} \mathbf{a_l} \tag{4}$$

As $\mathbf{a_{final}}$ captures how information propagates from the input layer to the embeddings in higher layers, it serves as a better choice for selecting discriminative regions compared to the single layer raw attention weights $a_{L-1}$. We then choose the index of the maximum value $A_1, A_2, ..., A_K$ with respect to the $K$ different attention heads in $\mathbf{a_{final}}$. These positions are used as index for our model to extract the corresponding tokens in $\mathbf{z_{L-1}}$. Finally, they concatenate the selected tokens along with the classification token as the input sequence which is denoted as:

$$\mathbf{z_{local}} = (z_{L-1}^0; z_{L-1}^{A_1}, z_{L-1}^{A_2}, ..., z_{L-1}^{A_K}) \tag{5}$$

By replacing the original entire input sequence with tokens corresponding to informative regions they not only keep the global information but also force the last Transformer Layer to focus on the subtle differences between different sub-categories while abandoning less discriminative regions such as background or common features among classes.

**4. Contrastive Feature Learning**. He et. al observe that simple cross-entropy loss is not enough to supervise the learning of features since the difference between subcategories may be extremely small. To ameliorate this, they utilize contrastive loss $\mathcal{L}_{con}$ which minimizes the similarity of classification tokens corresponding to different labels and maximizes the similarity of classification tokens of samples with the same label. In order to prevent loss being dominated by easy negative samples, they also introduce constant parameter $\alpha$ such that only negative pairs with similarity larger than $\alpha$ can contribute to the loss $\mathcal{L}_{con}$. For a batch size of $N$, contrastive loss is calculated with

$$\mathcal{L}_{con} = \frac{1}{N^2} \sum_i^N ( \sum_{j; y_i = y_j}^N (1 - \text{cosine\_similarity}(z_i, z_j) + \sum_{j; y_i \neq y_j}^N max((\text{cosine\_similarity}(z_i, z_j) - \alpha), 0)) \tag{6}$$

where $z_i$ and $z_j$ are pre-processed with L2 normalization. The TransFG architecture is trained with the loss function $\mathcal{L} = \mathcal{L}_{CE}(y, y') + \mathcal{L}_{con}(z)$ where $\mathcal{L}_{CE}(y, y')$ is the cross entropy loss between predicted label $y'$ and ground-truth label $y$.

#### 4.6.2 Transformer Architecture for Few-Shot, Fine-Grained Classification

We propose the final architecture outlined in Figure 3. The process of developing the final architecture was an iterative process governed by the scientific method. Models and the rationale for their construction shall be discussed in this section. Results of final and intermediate models shall be discussed in the following section, *5. Results*.

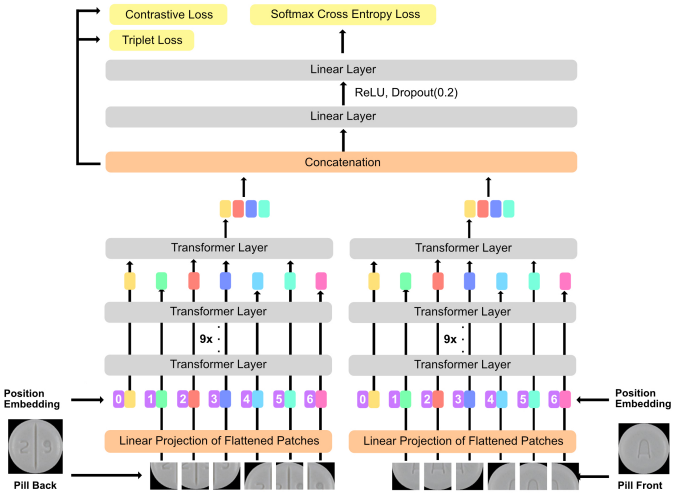


Figure 3: **Framework of Proposed Dual Transformer Model**. Images are split into non-overlapping patches and projected into the learned embedding space. The input consists of patch embeddings along with learnable position embeddings. Before the last Transformer Layer, the Part Selection Module [16] is applied to select tokens that correspond to the discriminative image patches and only use these selected tokens as input. Figure design inspired by He et. al [16]

**Iteration 1. TransFG with Single-Sided Pill Queries**. Due to the wide success of TransFG at achieving state of the art results across several FGVC datasets (CUB-200-2011, NABirds) [16], we initially hypothesized that a single-sided query would be valuable enough for the TransFG model to accurately distinguish between image classes. Results of this model are elaborated upon in section *5. Results*. We train this model utilizing a learning rate of $0.1$, after testing learning rates of $0.001$ and $0.0001$. To update model parameters we utilize stochastic gradient descent with momentum = 0.9. The model tends to converge after 50 - 100 epochs.

**Iteration 2. TransFG with Double-Sided Pill Queries**. Following minimal success of single-sided queries with TransFG, two approaches were construed in order to generalize the TransFG to accept two images. The first architecture (Iteration 2.1) entailed naively stitching the front and back images together, passing in rectangular inputs to the original TransFG model. Due to the rectangular inputs, we first manually warm-start this network with the ImageNet21K dataset unlike other models that take square inputs where `ViT-B_16` pretrained model weights [11] could be leveraged. The second architecture (Iteration 2.2) entailed training two TransFG models with identical architecture, one trained on the front pill image and one trained on the back pill image. We then utilize a fully-connected dense layer with softmax cross-entropy loss in order to train both models simultaneously and make predictions.

We train both of these model utilizing a learning rate of $0.1$. To update model parameters we utilize stochastic gradient descent with momentum = 0.9. Iteration 2.1 tends to converge after 50 - 100 epochs after 500 epochs of warm-started training on the ImageNet21K dataset [23]. Iteration 2.2 tends to converge after 100 - 200 epochs. Results of these models are elaborated upon in section *5. Results* and Table 2.

**Iteration 3. Dual Transformer Architecture with Multi-head Metric Learning (Ours)**. Following poor performance and minimal success of attempting to get the TransFG model to generalize to a low-shot setting, the following alterations to the TransFG model are proposed based on observations from the experiments in previous iterations to create the model proposed in Figure 3.

**1. Dual Architecture.** In order to accommodate the two-sided pill queries, two distinct modified TransFG models are trained. The modified TransFG models utilize Image Sequentialization, Patch Embedding, and the Part selection model originally proposed by He et. al. However, utilize a different MLP head and training procedure in order to be effective in a few-shot setting. Each individual transformer architecture is warm-started [3] utilizing the ImageNet21K dataset [23]. The final output tokens are concatenated, passed through two fully-connected layers, one with 20% Dropout a ReLU activation function, and the other with a Softmax activation function to make predictions.

**2. Learned Embedding of Concatenated Output Tokens for Classification.** The rationale behind concatenating the output tokens of each modified TransFG Transformer model and passing them through the two fully-connected layers is to allow trained relationships between the front and back-side of the pills that distinguish the pill class. Dropout [24] is utilized to ensure both sides of the pill are used in classification instead of lazily becoming dependent on the front pill image which can be used to distinguish a large number (but not all) pill classes. We train this classification head with plain cross-entropy loss $\mathcal{L}_{CE}$ as outlined in Figure 3.

**3. Triplet Loss on Concatenated Output Tokens.** After observing poor performance utilizing only contrastive loss in TransFG and the wide success of multi-headed loss functions in training the embedding layer of the ResNet-152 CBP baseline model, metric learning techniques were investigated for training the concatenated final outputs of modified TransFG models.

We propose a multi-headed loss function $\mathcal{L}_{dual}$ for training the concatenated output tokens of each modified TransFG model.

$$\mathcal{L}_{dual} = \mathcal{L}_{con}(z) + \mathcal{L}_{trip}(z) \tag{7}$$

Where $\mathcal{L}_{trip}(z)$ is the triplet loss of batch $z$. The triplet loss function uses three images during evaluation. The anchor is an arbitrary data point, the positive image has the same class as the anchor and the negative image is a different image class. The triplet loss reduces the distance between the anchor and the positive image while increasing the distance between the anchor and the negative image. The Pytorch Metric Learning implementation of Triplet Loss [21] was utilized and can be calculated with Equation 8 where $x_i^a$ is the anchor sample, $x_i^p$ is the positive sample, $x_i^n$ is the negative sample, and $\alpha$ is the margin between positive and negative pairs. We set $\alpha = 0.05$.

$$\mathcal{L}_{trip} =_i^N \left[||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha\right] \tag{8}$$

**4. Image Augmentation.** In order to increase the amount of data available in this few-shot learning problem and improve the generalizability of our model, we utilize image augmentation. Image augmentation was particularly useful for using only one pair of reference images to classify consumer images which vary in lighting, angle, size, and rotation. In particular, we utilized rotation (due to the rotation invariance of pills), perspective transformations, scaling, and changes in brightness, contrast, saturation and hue. Perspective transformations and changes in brightness, contrast, saturation and hue were only applied to reference images.

**5. Training.** We train this model utilizing an initial learning rate of $0.1$ with a cosine annealing scheduler that decreases the learning rate from $0.1$ to $0$ every 50 epochs. To update model parameters we utilize SGD with momentum = 0.9. The model tends to converge after 150 - 250 epochs.

# 5 Results

## 5.1 Ensemble Classification

Averaging Ensemble Architectures
1. Top 2: ResNet152 CBP + Resnet 101 CBP
2. Top 3: ResNet152 CBP + ResNet101 CBP + ResNet50 CBP
3. Top 4: ResNet152 CBP + ResNet101 CBP + ResNet50 CBP + DenseNet161 CBP
4. Top 5: ResNet152 CBP + ResNet101 CBP + ResNet50 CBP + DenseNet161 CBP + ResNet34 CBP

Weighted Averaging Ensemble Architectures
1. Top 2: ResNet152 CBP (70%) + ResNet101 CBP (30%)
2. Top 3: ResNet152 CBP (50%) + ResNet101 CBP (30%) + ResNet50 CBP (20%)
3. Top 4: ResNet152 CBP (40%) + ResNet101 CBP (25%) + ResNet50 CBP (15%) + DenseNet161 CBP (10%)
4. Top 5: ResNet152 CBP (35%) + ResNet101 CBP (25%) + ResNet50 CBP (20%) + DenseNet161 CBP (10%) + ResNet34 CBP (10%)

| Model | GAP | GAP@1 | MAP | MAP@1 |
|---|---|---|---|---|
| ResNet152 CBP | 37.91 | 82.37 | 92.11 | 85.30 |
| ResNet101 CBP | 40.45 | 86.20 | 93.44 | 87.86 |
| ResNet50 CBP | 35.10 | 78.30 | 89.68 | 81.30 |
| DenseNet161 CBP | 30.33 | 73.44 | 85.88 | 76.28 |
| ResNet34 CBP | 29.24 | 69.64 | 86.48 | 76.11 |
| Top 2 | 40.04 | 85.44 | 93.89 | 88.33 |
| Top 3 | 39.12 | 83.92 | 93.47 | 87.70 |
| Top 4 | 38.06 | 82.02 | 92.22 | 85.46 |
| Top 5 | 37.03 | 80.72 | 92.00 | 84.99 |
| Top 2 (weighted) | 39.43 | 84.93 | 93.64 | 87.94 |
| Top 3 (weighted) | 39.36 | 84.25 | 93.52 | 87.78 |
| Top 4 (weighted) | 39.00 | 83.81 | 93.36 | 87.54 |
| Top 5 (weighted) | 38.43 | 82.75 | 92.96 | 86.74 |

Table 1: **Results of baseline models, averaging ensembles, and weighted averaging ensembles**. Performance metrics of the top five baseline architectures and eight averaging ensemble architectures on the test dataset are reported in percentages.

### 5.2 OCR

For the Tesseract OCR engine, the noisy backgrounds of the pill images prevented accurate text detection on nearly all images. There were only a handful of images where Tesseract succeeded, typically when the text in the image was clear and high resolution. The EasyOCR engine was more adept at detecting and correctly recognizing text in both reference and consumer pill images. However, the model still made mistakes on a vast majority of pills; on images with higher resolution, the predicted text from EasyOCR would often be one or more characters off from correct and in images with low resolution the model would fail to detect any text. Though more promising than Tesseract, it was clear that neither model was reliable enough to be effective in an ensemble classifier.

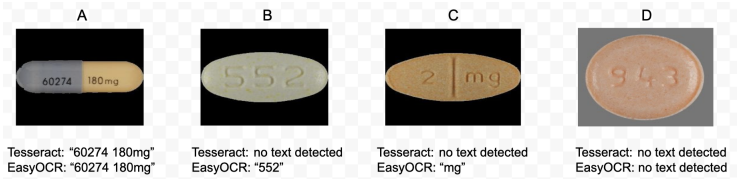


Figure 4: **Example outputs of OCR Models**. Outputs of Tesseract and EasyOCR is listed below the image. In image D, both models fail to identify text (which occurred in the majority of images in the dataset)

### 5.3 Attention-Based Approaches

The results of attention-based approaches are outlined in Table 2. We observed very low performance of plain TransFG architectures on the ePill-ID dataset, achieving maximally a mean MAP of 26.39% across 4-fold cross-validation (Iterations 1 2). Ultimately, studying the results of Iteration 1, 2.1 and 2.2, the poor performance of TransFG in a low-shot setting surprised us. TransFG is the state of the art approach for many fine-grained classification datasets (CUB-200-2011, NABirds) [16] and outperforms all RNN architectures significantly. We hypothesized that these results would generalize to a few-shot setting; however, they did not.

However, our Dual Transformer Architecture (Iteration 3) outperforms the current state-of-the-art attention-based model (TransFG) by nearly 300% across all metrics. Our Dual Transformer Architecture achieves 68.93% GAP, 75.50% MAP, 80.70% GAP@1, and 72.29% MAP@1. However, we do not manage to outperform the current state-of-the-art model on the ePill-ID dataset, falling short by approximately 10-15% across all metrics, as elaborated upon in Table 2.

| Model | GAP | GAP@1 | MAP | MAP@1 |
|---|---|---|---|---|
| **Baseline Methods** | | | | |
| ResNet 152 | $39.57 \pm 1.23$ | $49.97 \pm 1.58$ | $68.51 \pm 1.09$ | $51.64 \pm 1.23$ |
| ResNet 152 CBP Cross-Entropy | $41.45 \pm 2.15$ | $52.43 \pm 2.07$ | $69.12 \pm 2.10$ | $52.64 \pm 2.29$ |
| ResNet 152 CBP Multi-head Metric Learning | $81.20 \pm 1.47$ | $91.19 \pm 0.28$ | $95.76 \pm 0.40$ | $92.01 \pm 0.67$ |
| **Attention-Based Methods** | | | | |
| TransFG (Single-Image Query) (Iter. 1) | $9.08 \pm 2.02$ | $11.71 \pm 1.00$ | $9.04 \pm 0.82$ | $8.72 \pm 1.70$ |
| TransFG (Stitched-Together Query) (Iter. 2.1) | $9.88 \pm 1.32$ | $12.72 \pm 1.66$ | $11.47 \pm 2.63$ | $11.00 \pm 0.82$ |
| Dual TransFG (Iter. 2.2) | $24.19 \pm 2.19$ | $29.72 \pm 2.84$ | $26.39 \pm 3.69$ | $24.47 \pm 3.06$ |
| Dual Transformer Architecture (Ours) (Iter. 3) | $68.93 \pm 2.72$ | $80.70 \pm 2.52$ | $75.50 \pm 5.38$ | $72.29 \pm 2.24$ |

Table 2: **Results of models on ePillID dataset**. Mean and standard deviations of holdout metrics from the 4-fold cross-validation are reported in percentages.

## 6 Discussion and Qualitative Analysis

### 6.1 Ensemble Classifiers

For the simple model averaging ensemble classifiers, we can see that the best performing models were the Top 2 averaging model and the Top 3 averaging model. These performed better than the baseline on the test dataset on all four metrics. However, we observe that the ensemble classifiers become slightly less accurate every time an additional weaker model is introduced. One major limitation of this method is the fact that each learner in the ensemble contributes equally to the final prediction; obviously this is not favorable this task as some architectures were more accurate than others.

The weighted averaging ensemble method provided an extra layer of flexibility as it allowed the predictions of a strong model to have more authority over the predictions of a weaker model. Surprisingly, despite weighting each model in the ensembles based on 4-fold cross validation performance, these models did not perform significantly better than the simple averaging ensemble classifiers. There are numerous more complex ensemble methods that could possibly lead to better performance than these two models such as stacking, boosting, and Bayes-model averaging.

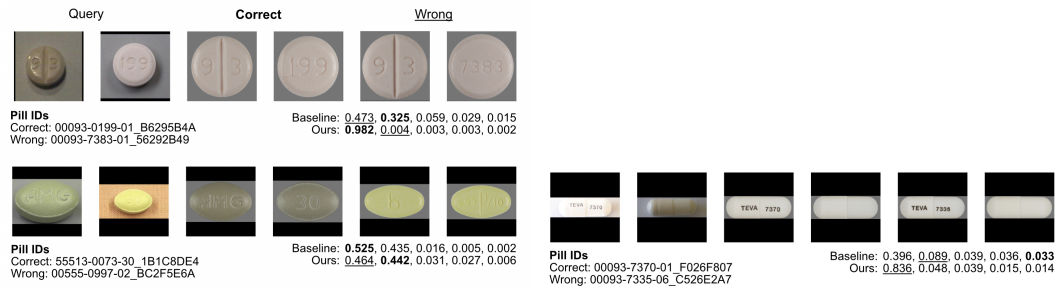


Figure 5: **Qualitative results from the ePillID holdout dataset**. For each query, top confidence scores (softmax layer output) are shown. from the state-of-the-art ResNet-152 CBP + Metric Learning baseline and our Dual Transformer model.

### 6.2 OCR

Despite the authors of the original paper recommending OCR integration as a next step, the current open source OCR models available are not robust enough to be involved in a classifier that utilized OCR. Even the widely used deep learning based engines Tesseract and EasyOCR failed to consistently identify the text imprinted on pills. This is likely a result of many factors such as poor image quality in the dataset, unpredictable orientation of images, varied backgrounds and fonts, and different textures of the pills that make text recognition extremely challenging. As a result, more sophisticated methods for OCR are likely required to achieve consistent performance on pills, such as training a deep learning based OCR on labeled pills. The texts on the pills were likely unusual for the pretrained Tesseract and EasyOCR models as they may have had different training data.

### 6.3 Attention-Based Approaches

**Iteration 1.** Analysing performance across the holdout validation set showed confusion across pills of similar color and shape. For example, the model failed to distinguish between circular, white pills. Furthermore, many pills have identical reverse sides making it impossible to predict a pill class with a single-sided pill query, making this a problem with the dataset rather than the training of the model. This prompted us to consider the models proposed in Iteration 2.

**Iteration 2.** As mentioned in Section 4, two iterations of this model were created. The first achieved double-sided pill queries by concatenating the two query images and using the resultant rectangle image as input (Iteration 2.1), the second involved using two TransFG modules, and concatenating output tokens in order to make predictions (Iteration 2.2). Iteration 2.1. The concatenated image model achieves similar performance to the single-image query model (Iteration 1). We attribute this poor performance to how vision transformers split the input image leading to sub-images that contained parts of both the front and back image. We believe this may have engineered non-discriminative features for the model as pill classes that scored highly during prediction did not necessarily look similar to the query pill class. Iteration 2.2. Studying the performance of the second iteration on the validation holdout set, we found that there existed a very small euclidean distance between learned embeddings of similar-looking pill classes which prompted us to consider metric learning techniques to better discriminate similar-looking pill classes.

**Iteration 3.** The performance of the our Dual Transformer model were impressive in the context of current attention-based approaches; however, did not perform as well as the state-of-the-art benchmark model. We illustrate a representative, qualitative analysis of the model in Figure 5. In the first row, we see that our attention-based model is responsive to larger text, even in low-lite consumer images. In the second row, we understand that our attention-based model is sensitive to changes in color and lighting that are seen in the consumer images. The brightening of the green pill causes the model to preference lighter green pills instead of giving preference to the text on the pill. Finally, in the bottom row (pictured right in Figure 5) we understand that still, both the attention-based and baseline models struggle to detect smaller text on the pills.

We hypothesize that the inability of our model to detect smaller text is due to the size of the window that the vision transformer splits the larger image into. Further study should be conducted into decreasing the size of this window, increasing the granularity of features for pill classes with smaller text. This hypothesis is reinforced by Wang et. al's [32] recent paper from NeurIPS 2021 that the classification of different images in contingent on the size of the window utilized by vision transformers. We also recommend more drastic image augmentation of color, brightness, and hue is implemented in order for the model to preference text instead of getting confused by pill color.

We also observe that in over 92% of cases where the Dual Transformer model provides a correct prediction on the holdout dataset, the correct confidence score is $> 0.2$ than any other class. This could provide a reliable start to automating pill detection. Further research should be conducted to study the accuracy of this model if it were to give a prediction when it is highly confident and request human intervention when it is not. We attribute this property to the triplet loss which separates the representation of distinct image classes.

## 7 References

## References


[1] T. A. Anto nio Cunha and P. Trigueiros. Helpmepills: A mobile pill recognition tool for elderly persons. *Procedia Technology*, 16:1523–1532, 2014.

[2] A. Antoniou, H. Edwards, and A. J. Storkey. How to train your MAML. *ICLR*, 2019, 2018.

[3] J. T. Ash and R. P. Adams. On the difficulty of warm-starting neural network training. *CoRR*, abs/1910.08475, 2019.

[4] X. Bai, M. Yang, P. Lyu, and Y. Xu. Integrating scene text and visual appearance for fine-grained image classification with convolutional neural networks. *CoRR*, abs/1704.04613, 2017.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers, 2020.

[6] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019.

[7] L. Delgado. Fast and accurate medication identification. *npj Digit. Med.*, 2, 10, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] T. G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[12] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. *CoRR*, abs/1511.06062, 2015.

[13] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman. Video action transformer network, 2019.

[14] R. Z. Gregory Koch and R. Salakhutdinov. Siamese Neural Networks for One-shot Image Recognition. 2015.

[15] S. Hashem. Optimal linear combinations of neural networks. *Neural Networks*, 10(4):599–614, 1997.

[16] J. He, J. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, C. Wang, and A. L. Yuille. Transfg: A transformer architecture for fine-grained recognition. *CoRR*, abs/2103.07976, 2021.

[17] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. Transreid: Transformer-based object re-identification, 2021.

[18] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[19] S. Ling1, A. as Pastor, J. Li, Z. Che, J. Wang, J. Kim, and P. L. Callet. Few-shot pill recognition. *CVPR*, 2020, 2020.

[20] T. Luo, A. Li, T. Xiang, W. Huang, and L. Wang. Few-shot learning with global class representations. *ICCV*, 2019, 2019.

[21] K. Musgrave, S. Belongie, and S.-N. Lim. Pytorch metric learning, 2020.

[22] F. Perronnin, Y. Liu, and J.-M. Renders. A Family of Contextual Measures of Similarity between Distributions with Application to Image Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2358–2365, Miami Beach, Florida, United States, June 2009. IEEE.

[23] T. Ridnik, E. B. Baruch, A. Noy, and L. Zelnik-Manor. Imagenet-21k pretraining for the masses. *CoRR*, abs/2104.10972, 2021.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[25] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo. Transtrack: Multiple object tracking with transformer, 2021.

[26] Y. Sun, K. Fu, Z. Wang, C. Zhang, and J. Ye. Road network metric learning for estimated time of arrival. *ICPR*, 2021, 2020.

[27] N. Usuyama, N. L. Delgado, A. K. Hall, and J. Lundin. epillid dataset: A low-shot fine-grained benchmark for pill identification. *CoRR*, abs/2005.14288, 2020.

[28] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

[31] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. *CoRR*, abs/1404.4661, 2014.

[32] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang. Not all images are worth 16x16 words: Dynamic vision transformers with adaptive sequence length. *CoRR*, abs/2105.15075, 2021.

[33] Y. Wang and Q. Yao. Few-shot learning: A survey. *CoRR*, abs/1904.05046, 2019.

[34] X. Wei, J. Wu, and Q. Cui. Deep learning for fine-grained image analysis: A survey. *CoRR*, abs/1907.03069, 2019.

[35] Y. Wong, H. Ng, K. Leung, K. Chan, S. Chan, and C. Loy. Development of fine-grained pill identification algorithm using deep convolutional network. *Journal of biomedical informatics*, 74:130-136. doi: 10.1016/j.jbi.2017.09.005, 2017.

[36] K. C. Xiao Zeng and M. Zhang. Mobiledeeppill: A small-footprint mobile deep learning system for recognizing unconstrained pill images. *ACM International Conference on Mobile Systems, Applications, and Services*, 2017.

[37] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo. Segmenting transparent object in the wild with transformer, 2021.

[38] Z. Yaniv. The national library of medicine pill image recognition challenge: An initial report. *IEEE Appl Imag Pattern Recognit Workshop*, 10.1109/AIPR.2016.8010584, 2016.

[39] A. K. J. Young-Beom Lee, Unsang Park and S.-W. Lee. Pill-id: Matching and retrieval of drug pill images. *Pattern Recognition Letters*, 33(7):904–910, 2012.

[40] H. Zheng, J. Fu, Z. Zha, and J. Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. *CoRR*, abs/1903.06150, 2019.

[41] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition, 2019.