Few Shot Image Classification

Introduction:

In the past decade, machine learning has experienced tremendous growth because of various factors like quality of data, increase in computational power and task specific model development. Another way to increase machine's learning efficiency is by training models with lots of data but doing so is not always practical because of obvious reasons. In many cases, data can be costly to annotate, and may not be readily available (like rare disease information).

In such a limited data situation Few Shot Learning has proved to be very beneficial in discovering patterns in data and making meaningful predictions. Few-shot learning refers to the process of feeding a network a small amount of training data instead of a large dataset as practiced normally. This technique is mostly utilized in the field of computer vision, where employing an object categorization model still gives appropriate results even without having several training samples.

Let us consider an example, we have many images of birds. But, due to birds being close to extinction we don't have many samples for those birds. The unbalanced distribution of images across classes can lead the model to skew towards classes with abundant images as it reduces the loss more and increases the accuracy a lot. But, now rare birds are *rarer* to be detected by the model. Hence, FSL urges the model to not learn to get lower loss or higher accuracy but actually learn to learn.

To summarize, we propose a **5-way-5-shot** classification model to facilitate FSL and perform quantitative analysis on the Hyper-parameters N & K, and evaluation metrics.

Dataset: Dataset used is **Caltech-UCSD Birds-200-2011**. It consists of 200 distincts bird classes and each class has around 60 images aggregating to a total of 11, 788 images in the dataset. Out of which 8251 images with 141 classes used for training the model, 1179 images with 21 classes used for validation and 2358 images with 40 classes used for testing.

Few Shot Image Classification

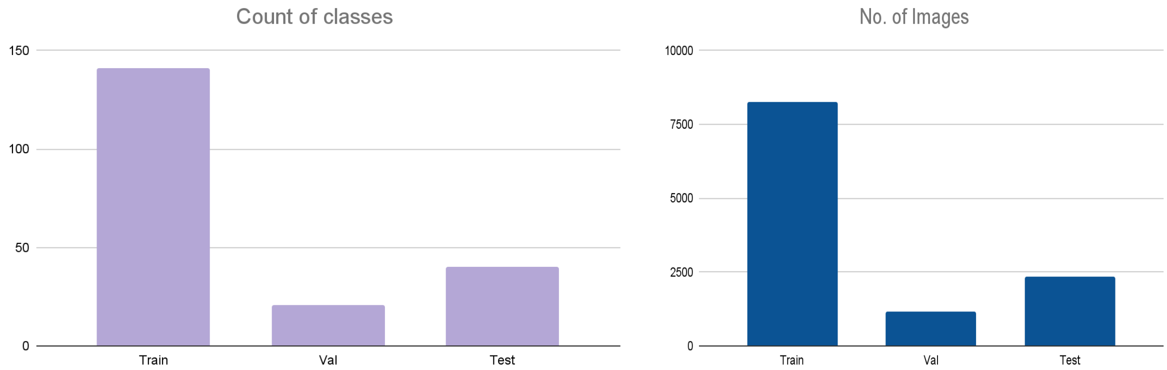

Figure: Number of Classes and Number of Images for training, Validation and Testing

Architecture and Method:

There are specific Meta-Learning algorithms used to solve Few-shot learning tasks. Some of them are Model-Agnostic Meta-Learning (MAML), Matching Networks, Prototypical Networks, Relation Network. In this project, we aim to achieve a model capable of few-shot classification using the above-mentioned hard to understand and implement complicated networks. Our model provides a three-stage approach to solve a few-shot problem. Let's understand how our model is working.

Initially, preprocess all images to (224,224,3) dimension, we randomly rotated and flipped just the training support set images. Then, we extracted 512-D features for all the images (both support and query images) using the ResNet18 model. We then pass this 512-D to an Auto-Encoder modeled with reconstruction loss(mean squared loss) to get high quality compact representations.

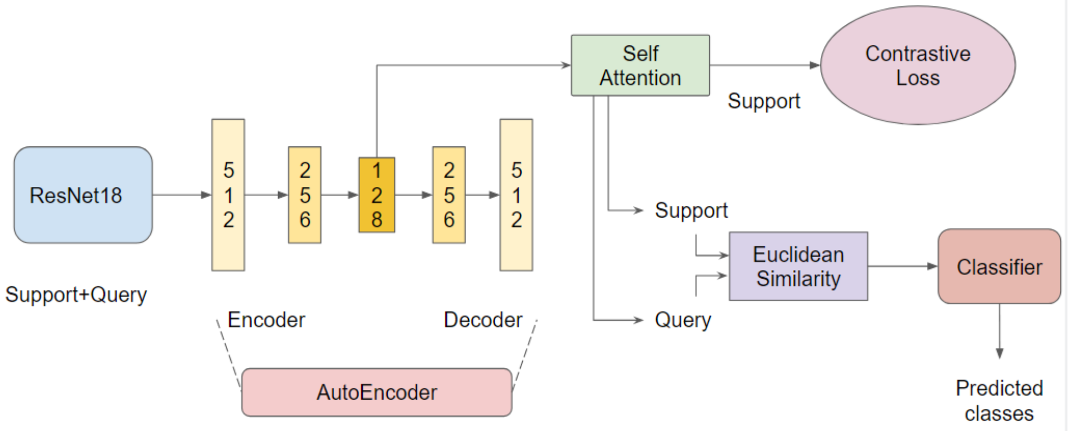

Figure: Model Architecture

Few Shot Image Classification

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$\mathbf{MSE}$	= mean squared error
n	= number of data points
Y_i	= observed values
$\hat{Y}_i$	= predicted values

Figure: Mean Square Error

The 128-D bottleneck of auto-encoder is used as the actual image representation. To further improve the representation we apply Self-Attention keeping the dimensions constant. The support set representations are passed to contrastive loss to bring the same class representation closer.

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)} \ ,$$

Figure: Contrastive Loss

This step is quite useful as this urges the model to push similar images close to each other in hidden space minimizing inter-class variance. To classify the query set, the euclidean similarity between query image representations and aggregated(mean over same class) support image representations. Cross-Entropy loss is applied over this euclidean similarity score.

$$L = -\frac{1}{m}\sum_{i=1}^{m} y_i \cdot \log(\hat{y}_i)$$

Figure: Cross Entropy Loss

The euclidean similarity score is passed through softmax to classify the query to a class with minimum distance for that query image. Model is trained for 20 epochs over episodes (support set, support label, query set, query label) with learning rate 3e-4 and we divide the learning rate in every 2 epochs by 2. Our final loss is:

$$\mathcal{L}_{final} = \mathcal{L}_{MSE} + \mathcal{L}_{contrastive} + \mathcal{L}_{cross-entropy}$$

Few Shot Image Classification

Results: We have taken the famous hypothesis into consideration that accuracy or similarity metric should increase with decreasing **N** and increasing **K.** To prove this we have considered two cases. First is by keeping **K=5** constant, the model is trained at increasing **N= 3, 5, 7** and accuracy is noted in each of the cases. Next we kept **N=5** at constant and the model is trained at increasing **K=3, 5, 7.** (Due to computation limitation we have below present results over 10 epochs)

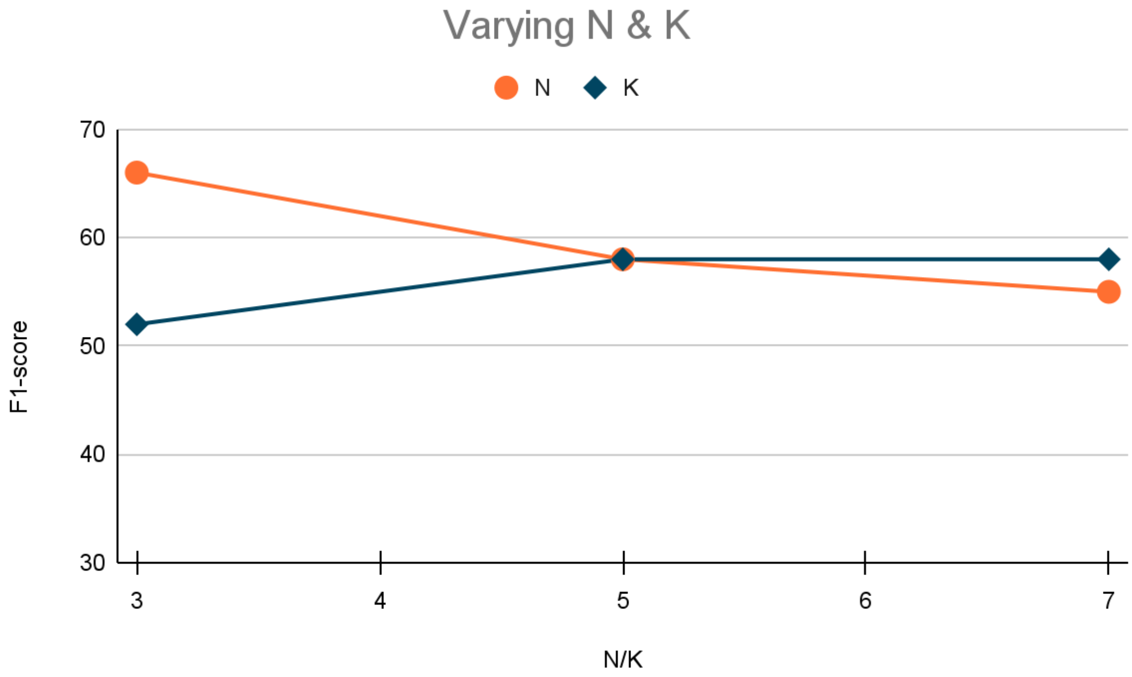

Figure: Result(F1-Score vs N/K)

From the graph it can be inferred that the hypothesis mentioned above is true indeed. This can also be used as evidence to prove the correctness of our model as the hypothesis holds true.

Our best model accuracy is 62% and F-score: 61% aggregated over all the episodes. We suspect that the accuracy can be further increased using the ResNet152 model (we tried this but as this is much deeper due to computation limits, we were not able to train successfully). A 2-stage fine tuning approach could also be beneficial.

Few Shot Image Classification

Conclusion & Consensus: Our model accurately came out as the Few-Shot classification model for variable N-way-K-shot. Hypothesis is proven to be true and it can be concluded that a simple model for few-shot classification can be realized. Our model performs adequately given the nature of classes during set are completely disjoint as that present while training.