Benchmarking Pre-trained Models for Image Classification

Ezra Kim

Agenda

- Problem Statement
- Hypothesis/Look Forward
- Exploratory Data Analysis
- Feature Engineering & Transformations
- Model Approach
- Results, Findings, & Learnings
- Future Work
- Important Links

Problem Statement

Transfer learning is exploding in popularity as companies like Huggingface, Steamship, etc continue to support and grow the machine learning community. With fine tuning, users can now utilize expensive models trained by big tech companies and use it to solve a multitude of problems.

In this presentation, we will examine out of box performance for image classification of four pre-trained models that has been fine tuned on the Indian food dataset.

The four models are Vision Transformers by Google, Res-Net by Microsoft, CovNext by Meta, and Mobile ViT by Apple.

Assumptions/Hypothesis

There were not too many assumptions being made in this experiment.

The main assumption is that Google's ViT would perform the best since it was trained on a larger dataset and the other three models would perform similarly

The pre-trained models chosen were all trained on the same ImageNet dataset with the difference coming in size of the dataset. The image transformation and fine-tuning would be the same so it's the pre-trained model itself that would be the differentiator.

Exploratory Data Analysis

It is difficult to do EDA on images but we have 20 classifications.

The dataset author did put in a caveat that this dataset does not have too much data per class so that will have a big effect on model performance since there was no data augmentation done.

All the images were different shapes so there was a need to standardize the images during image preprocessing

Data

- Dataset is from HuggingFace
- Indian Food Dataset
- Train Data Size: 4,795
- Validation Data Size: 533
- Test Data Size: 941

20 Classifications:

"burger", "butter_naan", "chai", "chapati", "chole_bhature", "dal_makhani", "dhokla", "fried_rice", "idli", "jalebi", "kaathi_rolls", "kadai_paneer", "kulfi", "masala_dosa", "momos", "paani_puri", "pakode", "pav_bhaji", "pizza", "samosa"

https://huggingface.co/datasets/rajistics/indian_food_images

Feature Engineering & Transformation

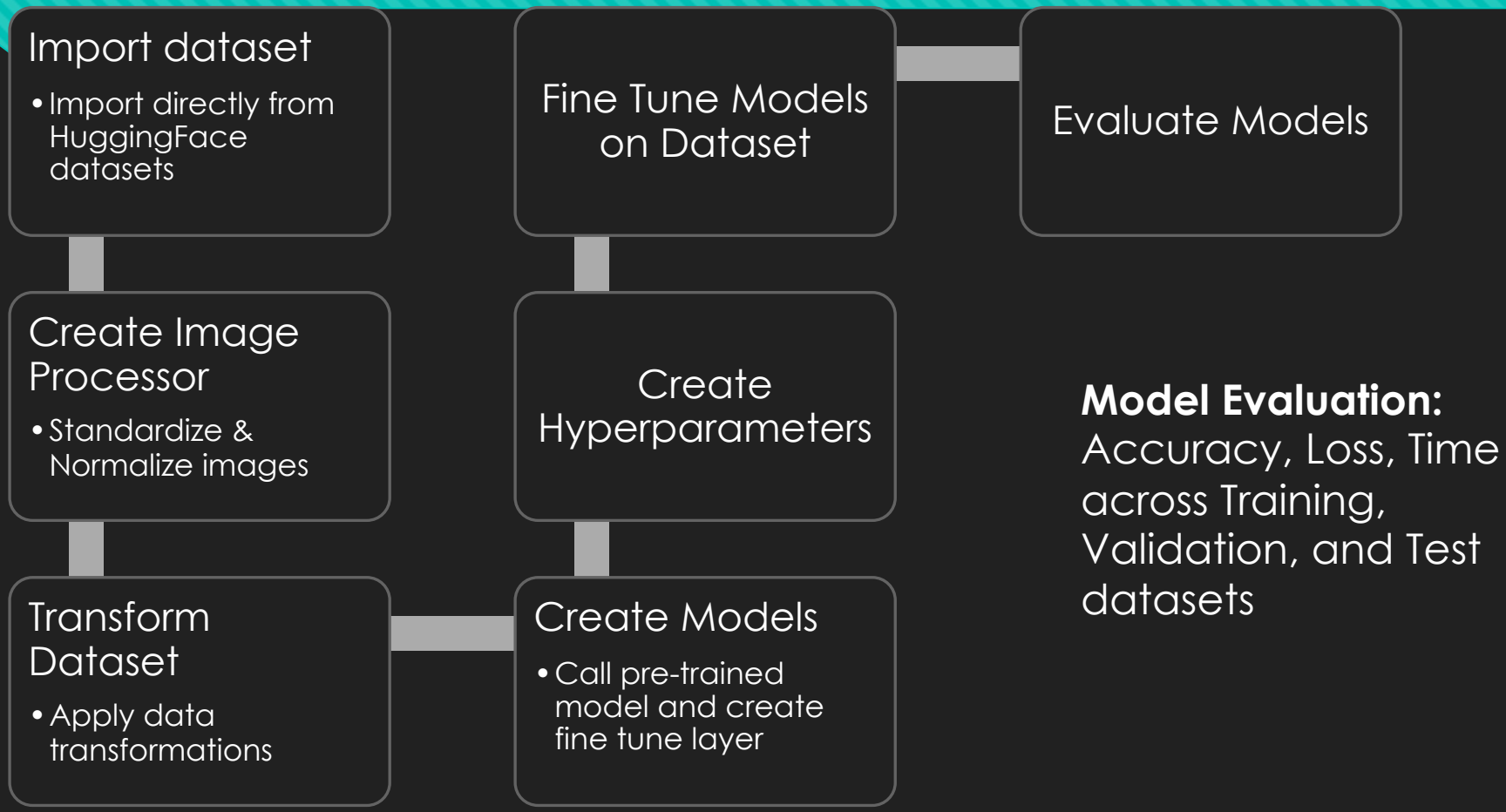
Hyperparameter Name	Hyperparameter Value
Training Epoch	3
Gradient Accumulation Steps	4
Learning Rate	5e-5
Logging Steps	10
Batch Size	32

Since the purpose of this exercise is to test the pretrained models, the hyperparameters and image transformations were kept to the minimum .

Pretrained Model	HuggingFace Model Name	Pre-Trained DataSet
Vision Transformer (base-sized model)	google/vit-base-patch16-224-in21k	ImageNet-21k (14 million images, 21,843 classes) at resolution 224x224.
ResNet-50 v1.5	microsoft/resnet-50	ImageNet-1k at resolution 224x224
ConvNeXT (large-sized model)	facebook/convnext-large-224	ImageNet-1k at resolution 224x224
MobileViT (small-sized model)	apple/mobilevit-small	ImageNet-1k at resolution 256x256

Image Transformations (PyTorch)
Center Crop
Compose
Normalize
Random Horizontal Flip
Random Resized Crop
Resize

Approach: Test models under similar conditions



Model Evaluation: Accuracy, Loss, Time across Training, Validation, and Test datasets

Results and Learnings

	Train			Validation			Test		
	Accuracy	Loss	Runtime	Accuracy	Loss	Runtime	Accuracy	Loss	Runtime
Google | ViT	0.919	2.13	00:09:51	0.919	1.58	00:00:17	0.874	1.62	00:00:32
Microsoft | Resnet	0.135	2.95	00:08:16	0.135	2.93	00:00:16	0.158	2.93	00:00:31
Meta | ConvNeXT	0.790	2.59	00:11:50	0.790	2.22	00:00:18	0.739	2.26	00:00:27
Apple | Mobile ViT	0.433	2.89	00:08:23	0.433	2.78	00:00:16	0.39	2.78	00:00:30

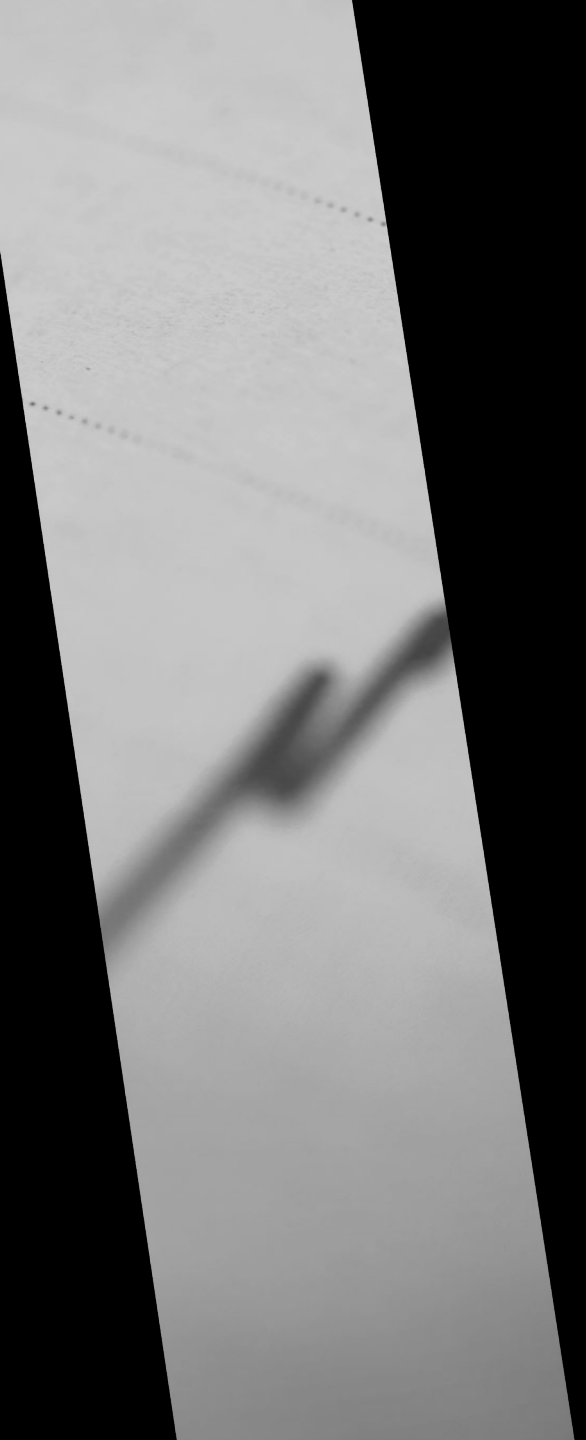
We can see that Google's ViT is the superior model, most likely due to the amount of data this pre-trained model has been trained on.

It is interesting to see that Meta's model was trained on the same data set as the Mobile ViT, and ResNet but performed comparably to Google's ViT but the ResNet model was so much worse which suggests that this model underfit and the model capacity is much lower compared to the other pre-trained models.

Future Work

- Data Augmentation
  - Per the description on HuggingFace, it is recommended to augment the data due each class not having too many images
  - With many of the pre-trained models, the models are not robust enough to properly train with the sparse classes
- Hyperparameter Tuning
  - Implementing hyperparameter tuning should increase performance. My hypothesis here is that with proper hyperparameter tuning, we should see the CovNext model perform as well as the ViT model
- Test more pre-trained models
  - There are many pre-trained models to choose from on HuggingFace. For this project I chose some of the more popular ones but there could be diamonds in the rough to use
- Test models built from scratch
  - I'd like to test models built from scratch to see if there are ways to outperform the pre-trained models

Thank You



Appendix

Important Links:

- Github: https://github.com/bigtreesfallhard/ml-final
- HuggingFace Data: https://huggingface.co/datasets/rajistics/indian_food_images
- Google ViT Model: https://huggingface.co/google/vit-base-patch16-224-in21k
- Microsoft ResNet: https://huggingface.co/microsoft/resnet-50
- Meta CovNeXT: https://huggingface.co/facebook/convnext-large-224
- Apple MobileViT: https://huggingface.co/apple/mobilevit-small