Final Project Instructions – Deep Learning Fall 2024
521153S-3005

Transfer Learning for Medical Image Classification

Motivation

In medical imaging, obtaining large, labeled datasets is often challenging due to privacy concerns, high annotation costs, and limited availability of expert knowledge. To effectively learn and boost performance on these smaller datasets we leverage transfer learning techniques which consist of models that are trained on huge amounts of data.

Goal

Improve the performance of diabetic retinopathy detection using transfer learning by fine-tuning models and understanding the classification results with visualizations and explainable AI.

Requirements

a) **Fine-tune a pretrained model using the DeepDRiD dataset**. (**5 points**)

1. Download the DeepDRiD dataset from the provided link along with the template code which would provide you with an explanation on diabetic retinopathy and the dataset itself. All the images for training, validation, and evaluation are 512 by 512 in size.

What should be output architecture? what should I predict? What hyperparameter should I use?

2. Fine-tune an ImageNet pretrained model (e.g., ResNet18, ResNet34, VGG, EfficientNet, DenseNet) on the DeepDRiD dataset. Evaluate and test it on the validation and test sets. Your goal here is to reach the highest Cohen Kappa score you can get.

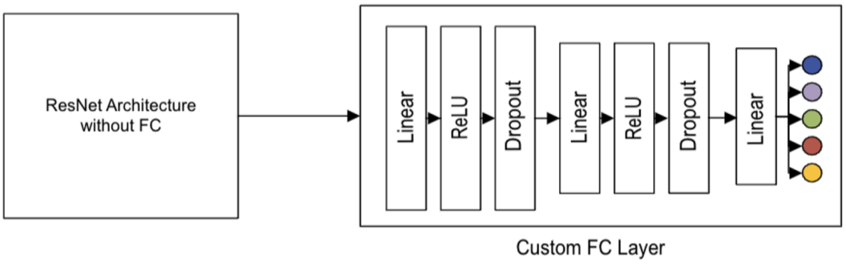
What are the augmentation techniques can be used here?

3. Play around different image augmentation techniques and check whether it boosts your evaluation metrics or decreases them.
4. Save the fine-tuned model.

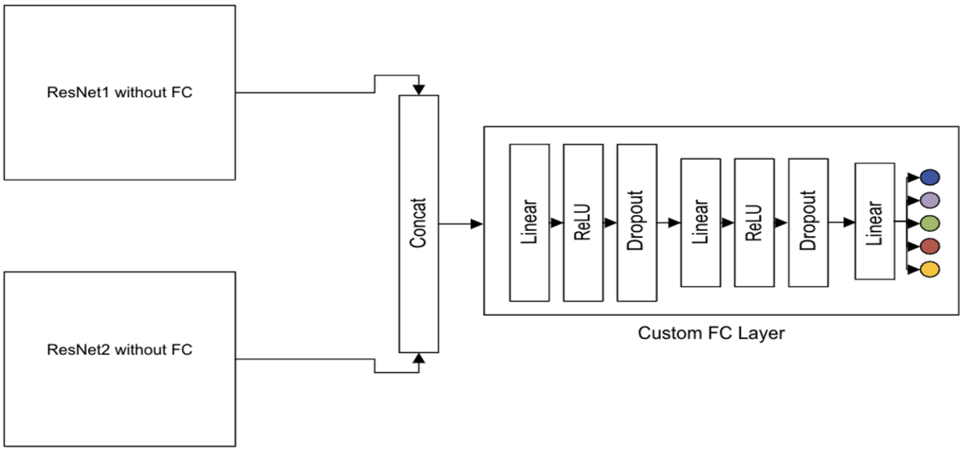
The goal of task(a) is to see how a model that is trained on a general dataset (pretrained) works for a specific task.

Reference Architecture

Here are two reference architectures described in the template code, i.e., **single image** and **dual image**. The single image architecture processes each image separately, while the dual image architecture takes two images of the same eye to fuse their features during training and evaluation. You can pick one or try both for tasks (a)-(e).



OR



b) **Two stage training with additional dataset(s)**. (**5 points**)

1. Choose a diabetic retinopathy dataset from either Kaggle DR Resized or APTOS 2019 Blindness Detection (links are provided below).
2. Fine-tune an ImageNet pretrained model (e.g., ResNet18, ResNet34, VGG, EfficientNet, DenseNet) on the selected dataset by unfreezing all pretrained layers. If you have any difficulties, you can also use pretrained weights of Kaggle DR Resized (pretrained_DR_resize) to skip this step: https://www.kaggle.com/competitions/521153S-3005-final-project/data
3. Next up, fine-tune this trained model on the DeepDRiD dataset (keep all the layers unfrozen) and see how it impacts your Cohen Kappa score.
4. Save the fine-tuned model.

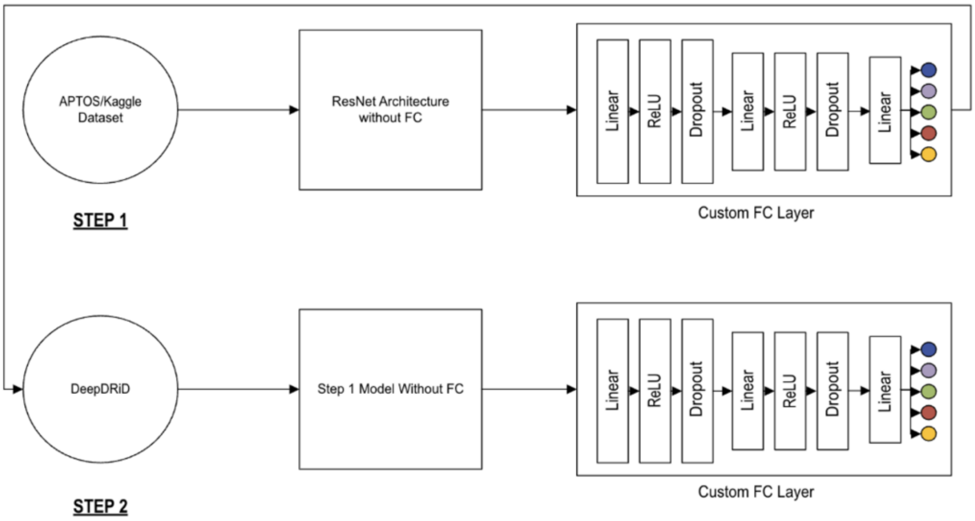
just use the pretrained weight. for step 2

What kind of difficulties I may have?

Initialize the pretrained model with custom weight.
Add custom layers at the end.
Trained this model on DeepDrid dataset.

The goal of task (b) is to compare the performance of a deep model that is trained and fine-tuned on a task-specific dataset with that of a model that is first trained on a general dataset and then fine-tuned on the same task-specific dataset in task (a).

Reference Architecture



Datasets:

Kaggle DR Resized: https://www.kaggle.com/datasets/tanlikesmath/diabetic-retinopathy-resized
APTOS – 2019: https://www.kaggle.com/datasets/mariaherrerot/aptos2019

Hints:

How to do this oversampling?

- When experimenting with different approaches for tasks (a) and (b), such as balancing the dataset, consider oversampling at the patient level instead of individual images. This ensures that all four images from a single patient are kept together as a unit.

When should we calculate Kappa score? In each epoch or at the end of all epoch?

- Keep in mind that the evaluation metric, "Cohen's Kappa," may show variance after each epoch. Therefore, training the model for too many epochs does not necessarily guarantee a higher Kappa score.

c) Incorporate attention mechanisms in the model. (**10 points**)

What is attention mechanism? How do I add it with the previous architecutre?

1. Implement attention mechanisms (e.g., self-attention, channel attention, or spatial attention) in your DeepDRiD model architecture.
2. Evaluate the impact of attention mechanisms on model performance.

The goal of task(c) is to apply and see how attention is affecting the model and its performance.

d) Compare the performance of different models and strategies. (**20 points**)

1. Use at least three transfer models that you've trained using task(b) and perform ensemble learning. Try out the following ensemble techniques (Stacking, Boosting, Weighted Average, Max Voting, Bagging) and analyze whether the performance increases or not.

2. Try out different image preprocessing techniques such as, Ben Graham, Circle Cropping, CLAHE, adding gaussian blur, sharpening up the images etc.

The goal of task(d) is to perform ensemble learning by training various models and combining their predictions and analyzing whether it boosts the performance. Along with that, applying multiple preprocessing techniques to see if that has any effect on the model.

Note: If any of the terms feel unfamiliar, kindly refer to the information below

e) Creating Visualizations and Explainable AI. (**5 points**)

1. Implement visualizations (e.g., scatter plots and line graphs) for your models' losses and accuracies on training and validation datasets to analyze the convergence and overall performance of the model.
2. Use Explainable AI techniques such as GradCAM to analyze what features in the image are contributing the most and the least in the model's decision-making process. Please also include a few visualization results in the report.

The goal of task(e) is to visualize the performance of the model by creating graphs and understanding the model's decision making through explainable AI.

References for Task(e)

- Ben Graham’s Preprocessing: https://medium.com/@astronomer.abdurrehman/enhancing-image-quality-for-machine-learning-ben-grahams-preprocessing-e795ad982abe
- Circular Cropping: https://www.geeksforgeeks.org/cropping-an-image-in-a-circular-way-using-python/
- CLAHE: https://www.geeksforgeeks.org/clahe-histogram-eqalization-opencv/
- Different type of blurs: https://docs.opencv.org/4.x/d4/d13/tutorial_py_filtering.html
- Ensemble Learning and types: https://corporatefinanceinstitute.com/resources/data-science/ensemble-methods/
- More ensemble types: https://www.geeksforgeeks.org/ensemble-methods-in-python/

Making your online submission:

You are required to include the results of the DeepDRiD test set for task (a)-(d) in the report.

To make your submission and get the performance on DeepDRiD test set, you can follow the template code and submit it to Kaggle for online evaluation. Your goal here is to predict labels (DR Levels) for those images and then submit those outputs as a csv file in the following form:

ID	TARGET
347_l1.jpg	0
347_l2.jpg	0
347_r1.jpg	0
347_r2.jpg	0

Key Terminologies

- **Fine-tuning:** It is the process of taking a pre-trained model and further training it on a specific dataset to boost performance. This may include unfreezing one or more layers of the pretrained model.
- **Augmentations:** These are the transformations applied to the training data to increase the diversity of samples and improve model generalization. Common augmentation techniques include rotations, flips, color adjustments, and cropping.
- **Kaggle:** A community platform for data science and machine learning competitions and datasets.
- **Cohen Kappa Score:** A statistical measure of inter-rater reliability used to assess the agreement between two raters (or models) when assigning categorical labels to data points.

  To calculate Cohen's Kappa, we have the following formula:

$$K = \frac{(p_o - p_e)}{(1 - p_e)}$$

  Where p_o denotes the observation agreement and p_e refers to the expected agreement between the raters.

  Initially, a confusion matrix is created where the raters (A and B) categorize the data, the rows of the matrix represent the categories predicted by rater A and columns by rater B.

A	B		
		X	Y
	X	10	2
	Y	5	20

In the above example both raters A and B agreed that 10 data points belong to the X category and 20 datapoints belong with Y category. With this we can now find the value for P_o which is the observed agreement between A and B (10 + 20)

$$P_o = \frac{10 + 20}{37} = 81\%$$

Hence the observed agreement is that in 81% of the cases both raters judge the same while 19% of the times they do not.

Now to find P_e which is the probability of getting a random match or agreement we first sum the rows and columns in our confusion matrix:

A	B			
		X	Y	
	X	10	2	12
	Y	5	20	25
		15	22	

$$P_e = \frac{12}{37} * \frac{15}{37} + \frac{25}{37} * \frac{22}{37} = 0.131 + 0.401 = 0.532$$

At first, we calculate the probability of both raters arriving at the decision for X by chance and then adding it with arriving at decision Y by chance. The two raters coincidentally agree 0.532 (53.2%) times.

Now we find the kappa score:

$$K = \frac{(0.81 - 0.532)}{(1 - 0.532)} = \frac{0.278}{0.468} = 0.594$$

Interpreting Kappa:

- < 0: Less than chance agreement
- 0.01–0.20: Slight agreement
- 0.21–0.40: Fair agreement
- 0.41–0.60: Moderate agreement
- 0.61–0.80: Substantial agreement
- 0.81–1.00: Almost perfect agreement

The use of Cohen Kappa in medical analysis allows researchers to measure how well the machine learning models are classifying data compared to the human experts. By considering the possibility of agreement by chance.

- **Explainable AI:** It refers to techniques which makes the AI decision-making process more interpretable and explainable to humans.
- **GradCAM:** Gradient-weighted Class Activation Mapping (Grad-CAM) is a technique used to visualize which parts of an image contribute the most to a model's prediction.
- **Attention:** Attention mechanisms that allow the model to focus on specific regions of an input image that are most relevant to the task at hand.
- **Ensemble Learning:** This involves combining multiple models to improve predictive performance. Ensemble methods can help reduce overfitting and increase model robustness.
- **Image Pre-processing:** This refers to the steps taken to prepare images for input into a machine learning model, such as resizing, normalization, and augmentation.

This two part is
important to learn
before hand

References

What is Fine-tuning: https://www.techtarget.com/searchenterpriseai/definition/fine-tuning

pytorch Augmentations: https://anushsom.medium.com/image-augmentation-for-creating-datasets-using-pytorch-for-dummies-by-a-dummy-a7c2b08c5bcb

Kaggle: https://www.kaggle.com/

Cohen’s Kappa Explanation: https://datatab.net/tutorial/cohens-kappa

Explainable AI (XAI): https://builtin.com/artificial-intelligence/explainable-ai

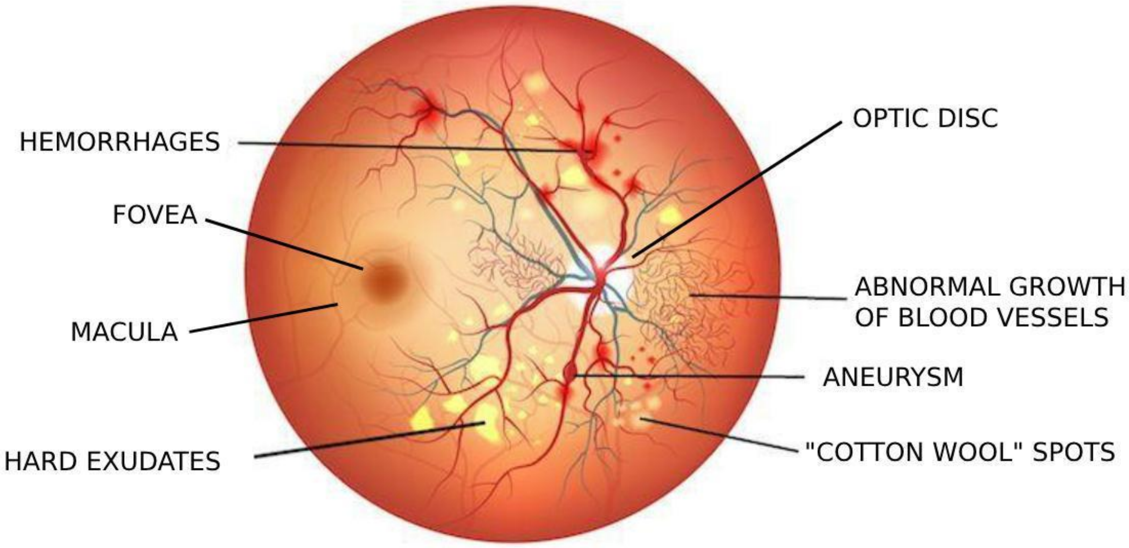
GradCAM Article: https://datascientest.com/en/what-is-the-grad-cam-method

Ensemble Learning and techniques: https://www.geeksforgeeks.org/a-comprehensive-guide-to-ensemble-learning/

Diabetic Retinopathy Grading

Diabetic Retinopathy

Diabetic Retinopathy is a complication of diabetes that affects the human eyes, specifically the retina. The cause of which is high blood sugar levels that over an extended period damage blood vessel. These damaged blood vessels can leak blood or fluid, leading to swelling of the retinal tissue and blurry vision. In severe cases it can also lead to permanent blindness.



Fundus Image

Key terms to understand Diabetic Retinopathy (DR)

The following terms are essential for understanding and diagnosing Diabetic Retinopathy (DR):

- **Fundus Images:** Pictures of the back part of the eye, which includes the retina.
- **Hemorrhages:** Areas in the retina where blood vessels have leaked or burst, creating small blood spots.
- **Fovea:** The darker pit in the retina responsible for sharp vision, containing a high concentration of cone cells.
- **Macula:** The small circular region around the fovea responsible for central vision (clear and detailed vision).
- **Soft Exudates / Cotton Wool Spots:** Fluffy white patches on the retina caused by tiny areas of damage where the blood supply to the retina has been blocked, like small strokes. They show that the retinal tissue is severely damaged.
- **Hard Exudates:** Yellowish deposits of fats and proteins that leak from blood vessels.
- **Optic Disc**: The brighter region where the optic nerve connects to the retina. It is where visual information is transmitted from the eye to the brain.
- **Aneurysm:** Small, balloon-like bulges in weakened blood vessels that can leak fluid or blood into the retina.

Classification of Diabetic Retinopathy

Diabetic Retinopathy can be classified into two main categories:

1) Non-Proliferative Diabetic Retinopathy (NPDR)

NPDR is the early DR stage with minimal symptoms like microaneurysms (tiny bulges) and dot hemorrhages (small red spots), but good vision is often maintained during this stage. This type can be further classified into:

- **Mild NPDR:** Its symptoms are the presence of microaneurysms. These are small bulges which can be detected through retinal imaging, they are scattered throughout the retina often in areas with high capillary density. Patients often do not notice any changes in their vision during this stage.
- **Moderate NPDR:** Increased number of microaneurysms and dot hemorrhages (small red dots) appearing throughout the retina, appearance of hard exudates (yellowish spots) often near the areas of leakage, and mild macula edema which refers to swelling in the macula that may cause slight vision changes.

2) Proliferative Diabetic Retinopathy (PDR)

PDR is the advanced stage of DR with the growth of abnormal new blood vessels on the retinal surface which significantly increases the risk of vision loss by bleeding and scarring. This growth of new vessels is known as neovascularization, and it often occurs at the edge of the retina (peripherals) and near the optic disc. These new blood vessels can form scar tissue which can cause problems with the macula or lead to a detached retina.

DeepDRiD Dataset

DeepDRiD (Diabetic Retinopathy Image Dataset) is a comprehensive dataset designed to aid in the diagnosis and grading of diabetic retinopathy. It covers a wide range of DR stages, from mild NPDR to Severe PDR.

The dataset includes a total of 2,000 images from patients with diverse backgrounds, taken under different camera angles and lighting conditions. Each patient has four images in total, two of their left eyes and two for the right eye.

Following is the list of relevant features provided in the dataset for the final project, refer to the DeepDRiD Challenge in the reference section if you wish to see all the dataset features:

I need this five example images.

Column	Level	Description
patient_id		Patients with the serial number.
image_id		Image sequence number.
patient_DR_Level	0	No apparent retinopathy
	1	Mild – NPDR
	2	Moderate – NPDR
	3	Severe – NPDR
	4	PDR
	5	Both eye fundus images is low and cannot be diagnosed and graded.

References

- What is Diabetic Retinopathy: https://www.aao.org/eye-health/diseases/what-is-diabetic-retinopathy
- Fundus Image Reference: https://www.themedicaleyecenter.com/diabetic-eye-disease-management-manchester/
- General Information on DR: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy
- More information on DR, prevention, and its types: https://www.mayoclinic.org/diseases-conditions/diabetic-retinopathy/symptoms-causes/syc-20371611
- DeepDRiD Challenge: https://www.sciencedirect.com/science/article/pii/S2666389922001040