Capstone 3 Proposal

Intro:

For my final capstone, I would like to follow up on my previous work which predicted cytokine production from images of human cells. The previous work was based on image embeddings from high-resolution images via extremely finely-tuned pre-trained models, but I did not generate these embeddings myself. Since I will likely focus on imaging professionally in the future, I would like to add some experience in this area to my portfolio.

Context:

Classification of microscopic images with machine learning has revolutionized biotechnology. However, the black-box nature of many machine learning approaches has left a gap between comfortable “knowns” of biologists and new “unknown” reasons for the classifications as generated by modeling. For example, some test drugs may cluster near control drugs and be listed as an exciting “hit” in a drug screen according to a clustering algorithm. But is the reason for the clustering really biological, or is it due to an artifact or a phenotype unrelated to the biology being studied?

Cytokines have been used as a known indicator of cellular activity for 50 years. By predicting cytokine production itself from images, rather than quantifying black-box similarity to control drugs, I aim to bridge the gap between the old-school and the new-school. My personal overall goal is to add confidence in the power and usefulness of ML approaches in drug discovery, improve throughput of wet-lab approaches and analysis methods, and thereby reduce the cost of drug discovery.

Spring Discovery is a biotechnology company that focuses on using and developing ML tools for drug discovery. They focus on imaging as their main data type. Being able to predict cytokines from images without actually needing to measure the cytokines would be a big boon to their suite of technologies.

Problem Statement:

The two questions addressed by this capstone are -

1) Can any cytokines be predicted from low resolution images of human cells via a CNN or a transferred ViT?

> More specifically, can even one of these 51 cytokines be predicted from these 384 64X64 tiffs?

2) Does employing data augmentation improve the performance of these models?

> The data augmentation strategies I will use will not be content generative - they will simply involve turning and flipping the pixel arrays I already have available.

Criteria for Success:

This work is simply proof-of-concept. A successful outcome is a clear answer to the question “can these low resolution images be used to predict any at all of these cytokines?”

For this analysis, I define predictable cytokines as those whose production can be predicted with at least 0.65 precision and 0.65 recall. (In this analysis, chance would result in scores of 0.5 since I will balance the classes.)

Constraints and Data Sources:

I am limited in my data selection because there are not many publicly available datasets that pair both cytokine data and images from the same exact wells. Generally, most people use one plate for imaging and then do another experiment entirely to collect cytokines to validate the drug hits. Luckily, there is one data set with cytokines pulled directly from the imaged wells, done by the Gates Foundation and Spring Discovery. This is publicly available since Gates is a philanthropic non-profit organization.

I know that embeddings of high-resolution images from thoroughly optimized pre-trained models (A combination of EfficientNetV2 and XLImageNet21) successfully predicted some cytokines in my part 1 analysis. Do I need such high resolution and so much effort transferring the models? This experiment will allow me to find out if I can use low-resolution images in a relatively small dataset and still successfully predict cytokines. This is important because Spring Discovery’s experiments are often relatively small, at least compared to how many images are used to pre-train neural networks.

Stakeholders:

Since this is a proof-of-concept exploration, the stakeholders are the internal scientists I will present this work to. There are no customers waiting on these results so that they can speedily deploy a cytokine prediction model to all of their future experiments. Instead, the stakeholders are simply curious if a thing like this is even possible. They are also curious about whether or not high-resolution is necessary, because if low-resolution images can be used then it could help them conveniently deploy this to future data (since the data is viewed and handled on a website where high-resolution images can really slow things down.)

Scope of solution space:

I will evaluate binary models for each cytokine individually. (In other words, I will predict whether each image is positive for a cytokine or negative for a cytokine.) I will try at least two CNNs and two ViTs, to get an idea of how adding and removing layers impacts the performance. I will not, however, do excessive fine-tuning, or employ detailed parameter sweeps for each cytokine's models. This work simply focuses on whether a CNN or ViT can be used at all, and whether a few simple augmentation approaches improves their performance. The stakeholders are not demanding deployable, robust models from this analysis. They only want to know if any of them has any potential.

Deliverables:

A report which gives a clear Yes/No on whether these low resolution images can predict the cytokine production of any cytokines.