Quick and Confident Learners

An Ablation Study of Masked Autoencoders for Transfer Learning in Medical Imaging

Name: ONG Ee Wen Lennard

Matric: A0034832L

Date: 24 October 2023

Motivation **2**
- Importance of Scalable Classification in Disease Pathology 2
- Domain specific challenges to training data 2

Research **2**
- General Challenges in ViTs: Large Datasets 2
- Domain-Specific Challenges: Pertubations & In-Class Variations 2

Approach **2**
- Intended Outcomes 3
- Datasets 3
- General Setup 3

Methodology & Results (Pretraining Masked Autoencoders) **3**
- Image Preparation 3
- Image Augmentation 3
- Experiment 1: Initial Setup 3
- Experiment 2: Adjusting Learning Rate 4
- Experiment 3: Refining Hyperparameters 4
- Results + Conclusions: Masked vs Unmasked Tradeoffs 4

Methodology & Results (Vision Transfomer Classifiers) **5**
- Architecture + Image Augmentation 5
- Experiments 5
- Results + Conclusions: Training 5
- Results + Conclusions: Accuracy, Precision, Recall, F1 6
- Results + Conclusions: “Confidence” Scores 6

Summary Conclusions **7**
- Further Opportunities: Domain-Specific MAEs 7

References **8**

Motivation

In this paper, we explore the value of pre-training domain-specific models for downstream medical imaging tasks. The goal is to understand if a pre-trained model can meaningfully reduce the need for training data while maintaining SOTA performance.

Importance of Scalable Classification in Disease Pathology

Take for example the task of white blood cell (WBC) classification. There are 5 types of blood cells: basophil, eosinophil, lymphocyte, monocyte and neutrophil. Determining the correct type and number of white blood cells is crucial for diagnosing various diseases.

Variations in their counts or alterations in their characteristics can be indicative of different diseases or health conditions. For example, an increase in lymphocytes may indicate acute and chronic leukemia, hypersensitivity reaction, or viral infection, while a decrease might be a sign of AIDS, influenza, sepsis, or aplastic anemia. Similarly, alterations in shapes, sizes, and structures of these cells can also indicate disease. Therefore, accurate identification of these cell types allows for effective diagnosis and prognosis of various medical conditions. [1]

Healthcare professionals traditionally rely on manual screening of blood films, which is time-consuming and subjective, leading to inconsistencies.

Domain specific challenges to training data

Currently, the limited availability of large and diverse datasets in the medical field can pose challenges for training data for medical tasks.

Annotated medical datasets are often expensive and time-consuming to create, leading to smaller datasets. Furthermore, labeling medical data requires expert knowledge and can be subjective, leading to potential inconsistencies and uncertainties in the training data. The dynamic nature of medical data, including evolving disease patterns and treatment protocols, requires continuous updates and retraining of models to ensure their relevance and accuracy. [1]

Researchers also highlight significant variation in imaging patterns among different samples and the complex perturbations in images, such as color distribution, brightness, and contrast, caused by different staining qualities. [8]

Research

A lot of attention has been given to the application of Vision Transfomers to the task of medical imaging. A Vision Transformer (ViT) is a type of neural network that processes images by breaking them into fixed-size patches, linearly embedding each patch, and then feeding them into a transformer architecture. This allows it to handle visual data using self-attention mechanisms, making it effective for various image-related tasks.

While it is able to achieve highly promising results, instability during training has been highlighted as a significant issue that can degrade the accuracy of ViT models. Researchers highlight that adjusting factors such as batch size and learning rate can improve stability during training. [5]

General Challenges in ViTs: Large Datasets

A significant challenge to ViTs is the amount of data required. A significant volume of data is needed for the model to perform optimally.

Generating pre-trained models have shown promising results. A novel approach to self-supervised learning in computer vision using Masked Autoencoders (MAEs) has been studied and shown non-trivial impact on training ViTs.

Domain-Specific Challenges: Pertubations & In-Class Variations

To overcome this challenge, patch augmentation and semantic segmentation have been explored as solutions.

Methods have been explored for improving the quality of annotated samples in pathology image analysis. By shuffling image patches within the same batch and using a curriculum learning-inspired strategy, the method preserves the relationships between instances and achieves state-of-the-art performance[7]

Similar techniques such as the shuffle instances-based Vision Transformer (SI-ViT) approach have been introduced to reduce image perturbations and enhance modeling among instances. This approach focuses on cells rather than various perturbations, ensuring accurate classification. [8]

Finally, techniques such as DINO (Emerging Properties in Self-Supervised Vision Transformers) have been used to provide explicit information about the semantic segmentation of images. Self-supervised methods can achieve high accuracy when combined with Vision Transformers. [6]

Approach

This project conducts an ablation study to explore the influence of pretraining and the size of training data in the domain of medical pathology.

In our approach, we start by pretraining a ViT-MAE model using generalized domain images. Following this, we employ transfer learning techniques to construct a white blood cell classifier, leveraging a domain-specific dataset.

Intended Outcomes

Through this ablation study, we aim to shed light on the following lines of inquiry:

1. **Pretraining Influence**: To what extent does pretraining on generalized domain images influence the model's proficiency in the specific task of white blood cell classification?
2. **Dataset Size Impact**: How does varying the size of the training dataset impact the classifier's performance metrics?
3. **Performance Plateaus**: Are there diminishing returns in performance improvements with increasing dataset size, especially after introducing pretraining?

We aim to determine this by assesing the model's performance, post-ablation, using the following established metrics: Accuracy, precision, recall, F1 and “confidence”, which will be defined as the p-score of the predicted class.

Datasets

For the initial pre-training phase, we utilize domain-appropriate datasets to provide our model with a broad understanding of medical imaging:

pRCC Dataset: This dataset is centered around Papillary Renal Cell Carcinoma (pRCC). It contains high-resolution images from renal tumor pathology slides, capturing the nuanced characteristics of pRCC. While not exclusively focused on hematologic cells, we hypothesize the diverse range of images within this dataset is instrumental in refining our model's ability to recognize biological patterns, thus enhancing its overall generalization capabilities.

Camelyon16 Dataset: This dataset is derived from a challenge that aimed to detect metastatic breast cancer in lymph node sections using whole-slide images. Even though Camelyon16 doesn't revolve around hematologic cell categorization, we hypothesize that it introduces the model to the intricate details found in pathology slides, optimizing its subsequent capability to classify blood cells effectively.

For the specific task of WBC classification, we employ the **Rabin-maas White Blood Cell Dataset**. This dataset features microscopic imagery of various white blood cell types, namely basophils, eosinophils, lymphocytes, monocytes, and neutrophils. Each image carries a definitive label of the cell type it represents, forming the foundation of our white blood cell classification endeavor.

General Setup

The training sessions were executed on 5x A6000 GPU processors. The model was put through three distinct experiments, with each one encompassing a total of 100 epochs.

The optimizer AdamW was employed for its ability to balance the speed of convergence and training stability.

Methodology & Results (Pretraining Masked Autoencoders)

For training the MAE, we adopted for the ViT-base architecture.

Image Preparation

The two datasets employed for our experiments were the pRCC and CAM16. While the images from the pRCC dataset came in a considerably larger resolution of 2000×2000 pixels, those from the CAM16 dataset were of 384×384 pixel dimensions. To ensure a harmonized input data format and to facilitate more streamlined processing, we opted to crop the pRCC images down to a more manageable size of 500×500 pixels, bringing them closer in dimension to the CAM16 images. This decision not only optimized our computational resources but also yielded a more extensive pool of initial training data, essential for the initial stages of model training.

Image Augmentation

Cellular images, by nature, lack a specific directionality. To this, we introduced horizontal and vertical flips as a part of our augmentation strategy. We also subjected our images to a scaling augmentation to between a factor of 0.2 to 1. This enabled our model to capture a spectrum of details, from intricate local nuances to overarching global patterns, essential for precise medical image analysis.

Given the importance of colour, we decided not to provide any augmentations for this. We felt it important for the model to learn the distribution of the actual colour space. This is later suggested otherwise in the experiment observations.

Experiment 1: Initial Setup

For the first experiment, we adopted the baseline settings from the MAE paper. the parameters listed below were utilized:

- base_learning_rate: 1.5e-4
- mask_ratio: 0.75
- weight_decay: 0.05
- per_device_train_batch_size: 48
- gradient_accumulation_steps: 8
- lr_scheduler: cos

This setup, however, encountered difficulties. The training loss diverged wildly, implying the learning rate was too aggressive for this dataset.

Experiment 2: Adjusting Learning Rate

To address loss divergence, the following adjustments were introduced:

- The learning rate was reduced to 5e-5.
- The random seed was varied to provide a fresh perspective on the loss landscape.
- Weight decay reduced to 0.01 to lessen regularization and support learning.
- Gradient accumulation steps halved to 4 to update the weights more frequently.

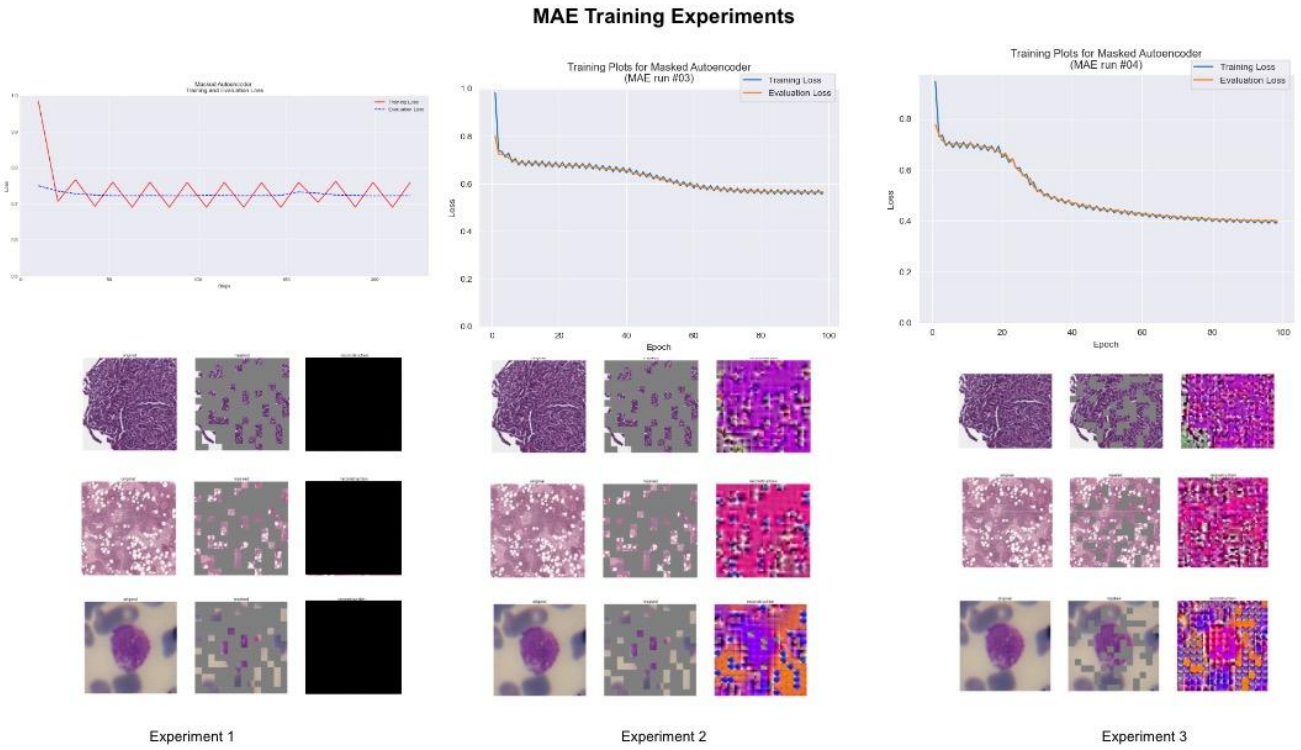
Though these changes facilitated better convergence with the loss consistently decreasing, a plateau was observed in the later phases. This suggested the presence of a local minimum, indicating the need for further optimization.

Experiment 3: Refining Hyperparameters

The objective of the third experiment was to simplify the challenge, aiding the model's learning capacity. To this end, a series of hyperparameters were fine-tuned to make learning more efficient:

- The mask ratio was brought down to 0.25, revealing more pixels to the model, potentially simplifying the task of learning
- The learning rate experienced a minor boost to 8e-5, aiming for quicker learning.
- Normalization was incorporated to render consistent input data, which could potentially enhance training stability.
- The cosine scheduler was discarded. Given our shorter epoch range in contrast to the 800 epochs used in the MAE paper, it was surmised that a cosine scheduler might be premature.
- Image scaling parameters underwent adjustments to concentrate on a specific size range, thereby shrinking the problem space.

Results + Conclusions: Masked vs Unmasked Tradeoffs



Observing the charts and reconstructed images, we can derive the following insights:

Progress in Training Metrics: The training plots for both experiments 2 and 3 exhibit a declining trend in training and evaluation losses, showcasing the model's consistent learning capability. Notably, in experiment 3, the decline appears more uniform. This suggests the applicability of MAEs to the medical domain for pretraining.

"Patch Spamming" Phenomenon: The model displays a tendency to replicate a specific patch in unfamiliar areas. While in experiment 2, this pattern appears predominantly in unmasked regions, experiment 3 reveals it more prominently within the masked areas.

Experiment 3 = Enhanced Reconstruction in Unmasked Regions: The findings from experiment 3 suggest a the model learning the latent characteristics of the training data. This is particularly evident as the unmasked portions in the reconstructed images of experiment 3 align more closely with their original counterparts.For example, in experiment 3, there's a better ability to replicate unmasked white regions, especially evident in CAM 16 representations.

Experiment 2 = Superior Inference in Masked Regions: During experiment 2, the model showcases an aptitude for deducing and filling in missing segments. This is evident in the WBC reconstructions, portraying an emergent understanding of the cellular structures.

Out of Distribution Color Challenge: The beige colorspace unique to White Blood Cells (WBC) emerges as a pronounced challenge. This is accentuated in experiment 3, suggesting that while a heightened masking area imposes stricter constraints, it can also guide the model more effectively. These observations resonate with findings from the MAE paper.

Methodology & Results (Vision Transfomer Classifiers)

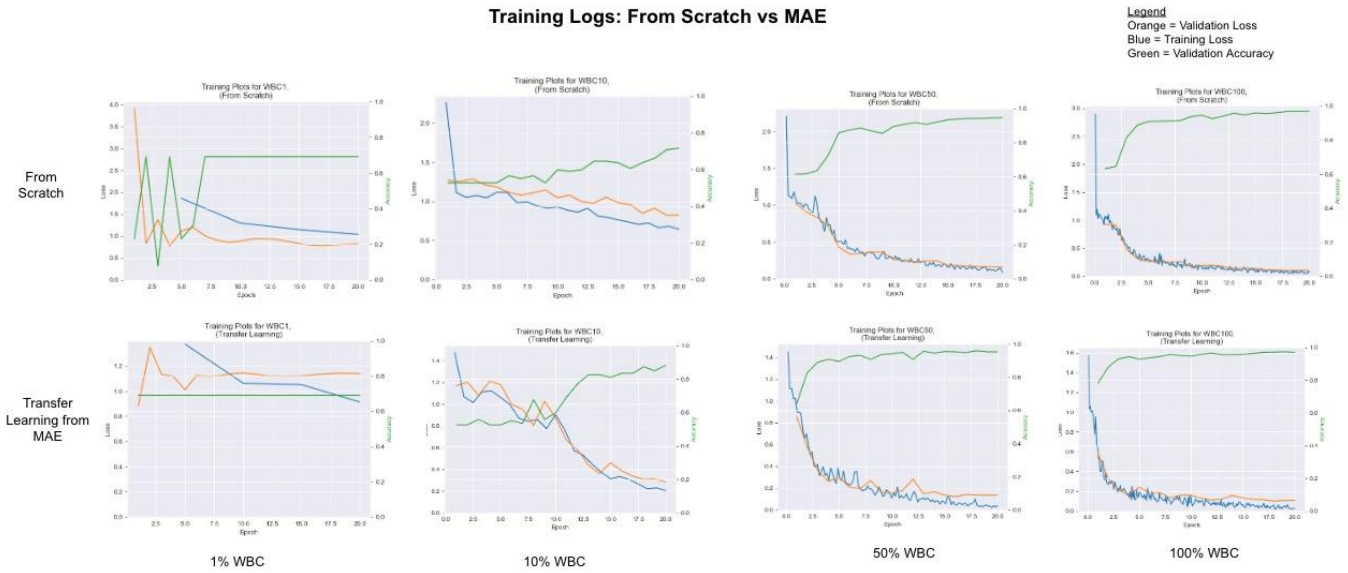
Architecture + Image Augmentation

Image augmentation was carried out in alignment with the protocols established in the MAE pipeline. For the transfer-learning approach, we adapted the encoded segment of the MAE model and incorporated a classifier head. This configuration was subsequently fine-tuned in its entirety using the WBC dataset. This strategy diverges from Linear-Probing, wherein only the classifier head undergoes retraining.

Experiments

We engaged in base hyperparameter tuning to guarantee loss convergence. It's worth mentioning that while there exists a potential avenue for further optimization via advanced hyperparameter adjustments, we chose not to delve into this to maintain the integrity of the ablation study. Having achieved satisfactory foundational performance, we built 8 models to a training regimen spanning 20 epochs for the ablation study.

Results + Conclusions: Training



From the training plots, we note the following observations about the training process:

Efficient Performance Gains for Smaller Datasets: Upon examination of the 10% WBC dataset, it's clear that transfer learning provided a significant performance advantage compared to training from scratch. This finding underscores the potential of a well-trained MAE model in facilitating transfer learning, especially when only a limited dataset is available.

Training Kick-Start from Transfer Learning: The sharp decline in loss during the initial epochs in transfer learning models across all datasets underlines the robustness of the MAE pre-trained model and its ability to kick-start the learning process.

Saturation Threshold: When observing the 100% WBC dataset, beyond the preliminary training phase, minimal distinction was noted between models trained from scratch and those benefiting from transfer learning. This observation might indicate a saturation threshold. Specifically, with a sufficiently large dataset, the benefits derived from a pre-trained model may become negligible.

Quicker Convergence in "From Scratch": The learning curves, especially the validation accuracy, for models trained from scratch tend to stabilize earlier than their transfer learning counterparts in datasets of lesser size. This suggests quicker convergence, but at a potentially suboptimal performance level.

Results + Conclusions: Accuracy, Precision, Recall, F1

	Accuracy			Precision			Recall			F1		
	From Scratch	w/ MAE	MAE Gains	From Scratch	w/ MAE	MAE Gains	From Scratch	w/ MAE	MAE Gains	From Scratch	w/ MAE	MAE Gains
1% WBC	0.61227	**0.61285**	0.06%	0.12252	0.12257	0.00%	0.19981	**0.2**	0.02%	0.1519	**0.15199**	0.01%
10% WBC	0.73148	**0.90104**	**16.96%**	0.36244	**0.78875**	**42.63%**	0.37126	**0.78074**	**40.95%**	0.36438	**0.78292**	**41.85%**
50% WBC	0.94965	**0.95486**	0.52%	0.89466	**0.90775**	1.31%	0.89335	**0.91559**	2.22%	0.89361	**0.9102**	1.66%
100% WBC	**0.97164**	0.97049	-0.11%	**0.95332**	0.93961	-1.37%	0.93709	**0.93979**	0.27%	**0.94481**	0.93882	-0.60%

biggest gains from MAE

The above table compares the accuracy, precision, recall and F1 score of the ablation study. From the above, the following observations:

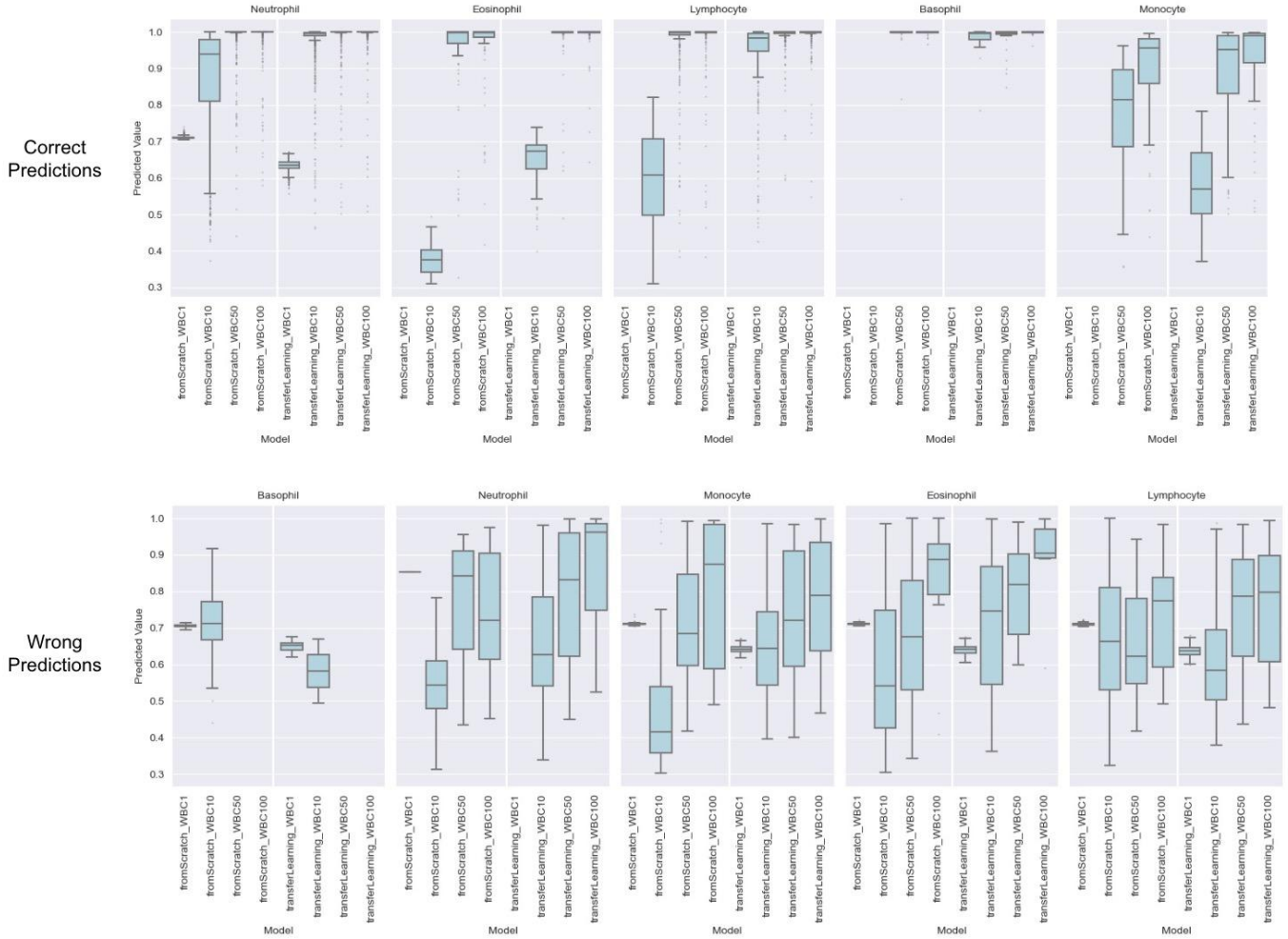
Major Gains with Small Data: At low data volume (1% WBC and 10% WBC), the use of MAE transfer learning consistently outperforms the from-scratch models, especially in precision, with gains reaching as high as 42.63%.

Performance Plateaus: As the dataset size increases (50% and 100% WBC), the gains from using MAE transfer learning become minimal, and in some cases, the from-scratch models slightly outperform the transfer learning models, such as in the accuracy metric for 100% WBC.

Recall Consistency: For recall, the gains from using MAE transfer learning remain positive across all data volumes, but the benefits are most significant at 10% WBC with a 40.95% improvement.

Mixed Results for F1 Score: The F1 score shows a mixed trend. For 1% WBC and 10% WBC, transfer learning offers better or equal performance, but as the dataset size grows, the gains reduce and even become negative (e.g., -0.60% for 100% WBC).

Results + Conclusions: "Confidence" Scores



The above table compares the prediction score from the ablation study. From the above, the following observations:

Transfer Learning = More Confidence: For most cell types, models trained with transfer learning appear to have a higher median confidence in their predictions, both correct and incorrect, compared to models trained from scratch. This is especially so frp the 10%

dataset. For wrong predictions, transfer learning models, in general, have a higher "confidence" score, especially for labels like 'Monocyte' and 'Lymphocyte'. This could be a point of concern, as having high confidence in a wrong prediction can be misleading.

From Scratch = Variability in Confidence: Models trained from scratch seem to exhibit a larger variability in their prediction scores, as evidenced by the longer whiskers in the box plots for several cell types, particularly for correct predictions. This might indicate that from scratch models are less consistent in their confidence levels.

Robustness of Transfer Learning: Regardless of the data size, transfer learning models consistently show a narrower interquartile range in their prediction scores compared to from scratch models. This suggests that transfer learning models are more robust and less influenced by outlier data or specific subsets of the data.

Summary Conclusions

Pretraining Influence: The efficacy of transfer learning is clearly demonstrated in enhancing the training process. By using a pre-trained MAE model, the model's proficiency in white blood cell classification is substantially improved. Notably, even when relying on limited datasets, the training process accelerates, yielding better results in a shorter time frame. This is particularly evident in the pronounced improvement seen with the 10% dataset size, containing 842 samples across all classes.

Dataset Size Impact: Varying the size of the training dataset had a noticeable influence on the classifier's performance metrics. Transfer learning consistently led to faster convergence during training as compared to models started from scratch, a trend observed across multiple dataset sizes. The impact of transfer learning was especially prominent in scenarios with limited data, highlighting its utility in such cases.

Performance Plateaus: When examining the full 100% WBC dataset, the advantages of transfer learning began to wane after the initial epochs. This indicates a diminishing return in performance improvements with the increase in dataset size after leveraging pretraining. Additionally, as the dataset size grew, validation accuracy seemed to level off. This suggests that there might be a saturation point beyond which merely augmenting the training data does not yield considerable improvements in accuracy.

Further Opportunities: Domain-Specific MAEs

The task training of pre-training a domain-specific MAE has shown promising signs of progress and should be investigated furhter.

Observing the transition from MAE experiment 2 to MAE experiment 3, there's a notable improvement in the model's performance. While there's still room for enhancement, especially in the masked areas, the model's ability to boost a downstream process is noteworthy.

The contrast between experiments 2 and 3 reveals a nuanced trade-off: As the extent of pixel masking increases, the model becomes adept at processing masked regions. However, this appears to come at the cost of accuracy in the unmasked areas, as indicated by the “patch spamming” behavior. This observation emphasizes the significance of the pixel masking rate, identifying it as a crucial hyperparameter warranting tuning for the dataset at hand.

Moving forward, It would be interesting research avenue to implement the methods used to manage pertubations & in-class variations to this MAE pre-training phase. [7][8]

References

1. Kouzehkanan, Z.M., Saghari, S., Tavakoli, S., Rostami, P., Abaszadeh, M., Mirzadeh, F., Satlsar, E.S., Gheidishahran, M., Gorgi, F., Mohammadi, S. and Hosseini, R., 2022. A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. Scientific reports, 12(1), p.1123.

2. Gao, Z., Hong, B., Zhang, X., Li, Y., Jia, C., Wu, J., Wang, C., Meng, D. and Li, C., 2021. Instance-based vision transformer for subtyping of papillary renal cell carcinoma in histopathological image. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27– October 1, 2021, Proceedings, Part VIII 24 (pp. 299-308). Springer International Publishing.

3. Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R. and Manson, Q.F., 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience, 7(6), p.giy065.

4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R., 2022. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

5. Chen, X., Xie, S. and He, K., "An Empirical Study of Training Self-Supervised Vision Transformers. arXiv preprint arXiv:2104.02057

6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A., 2021. Emerging properties in self- supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9650- 9660).

7. Zhang, T., Yan, Z., Li, C., Ying, N., Lei, Y., Feng, Y., Zhao, Y. and Zhang, G., 2023. CellMix: A General Instance Relationship based Method for Data Augmentation Towards Pathology Image Analysis. arXiv preprint arXiv:2301.11513.

8. Zhang, T., Feng, Y., Feng, Y., Zhao, Y., Lei, Y., Ying, N., Yan, Z., He, Y. and Zhang, G., 2022. Shuffle Instances-based Vision Transformer for Pancreatic Cancer ROSE Image Classification. arXiv preprint arXiv:2208.06833.