Lean and Confident Learners

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

Name: ONG Ee Wen Lennard

Submission Date: 24 October 2023

Motivation & Research

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 achieving respectable performance.

Importance of Image 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.

Healthcare professionals traditionally rely on manual screening of blood films, which is time-consuming and subjective, leading to inconsistencies. Therefore, accurate identification of these cell types allows for effective diagnosis and prognosis of various medical conditions. [1] Using image classification of medical imaging for disease pathology enhances diagnostic efficiency and reduces the workload on healthcare professionals.

Challenges in Training Data for Pathology

Currently, the limited availability of large and diverse datasets in the medical field can pose challenges for training data for such 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] To overcome this challenge, patch augmentation and semantic segmentation have been explored as solutions. 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]

Recent Approaches to Image Models

Vision Transformers (ViTs) have emerged as a powerful tool in general image classification and also to medical pathology. At their core, ViTs break down images into smaller, fixed-size patches. Each patch is then transformed and processed through a neural network, which uses a special mechanism called self-attention. This allows ViTs to detect patterns in images with a level of detail that older methods might miss. This attention to detail is especially valuable when analyzing intricate patterns found in medical imaging. .

However, there is a tradeoff: ViTs often need a lot of data to work well. One common solution is to use models that have been pre-trained on other tasks to give ViTs a head start.

A recent innovation in this area is the Masked Autoencoder (MAE). The key principle behind MAEs is the masking of certain input features (in this case, parts of an image) and then challenging the network to predict or reconstruct the masked portions based only on the unmasked parts. This self-supervision technique forces the network to learn important features and representations of the data. The approach draws inspiration from the masking techniques used in NLP for models like BERT, but it's tailored for visual data. [4]

The key question now is: Can MAEs help reduce the amount of data we need to train ViTs effectively?

Approach

This project conducts an ablation study to explore the influence of transfer learning and the size of training data in the domain of medical pathology. In our approach,

1. **MAE Pre-Training:** We first train a Vision Transformer with Masked Autoencoder (ViT-MAE). Instead of using specialized medical images right away, we start with a broader set of medical images.
2. **ViT Classifier:** Once our ViT-MAE has learned from these general images, we introduce it to a more specific task: classifying white blood cells. By doing this, we're using what it has learned and applying it to a niche area.

By taking this route, we're exploring if the initial training with MAE-generated embeddings can give our model a better start when it faces the specific challenge of white blood cell classification.

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 utilizing distributed training to ensure efficient processing.

By default, the optimizer AdamW was employed for its ability to balance the speed of convergence and training stability. For all VIT and ViT-MAE models, we adopt the vanilla ViT-base architecture.

MAE Pretraining: Methodology & Results

Image Preprocessing & Augmentation

For our study, we utilized two primary datasets: pRCC and CAM16. The pRCC dataset images had a substantial resolution of 2000×2000 pixels, while the CAM16 images were more modest at 384×384 pixels. To ensure consistency in the input data format and to enable efficient processing, we segmented each pRCC image into quadrants measuring 500×500 pixels each, making their dimensions more comparable to the CAM16 images. This approach not only ensured uniformity in the starting image sizes but also effectively increased our initial training data by fourfold before any augmentation procedures.

For augmentation, we incorporated horizontal and vertical flipping into our augmentation techniques. This is given the non-directional nature of cellular images. Additionally, we applied scaling augmentation to our images, ranging from a factor of 0.2 up to 1. This ensured that our model was exposed to a broad range of details, from the fine granularities to the broader, more general patterns, which is crucial for the meticulous nature of medical image interpretation.

Given the importance of color in medical imaging, we consciously chose not to alter this aspect during augmentation. We hypothesized it was important for the model to familiarize itself with the authentic color distribution. However, subsequent observations from our experiments suggest that there might be room for reconsideration regarding this decision.

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 diverging, 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.

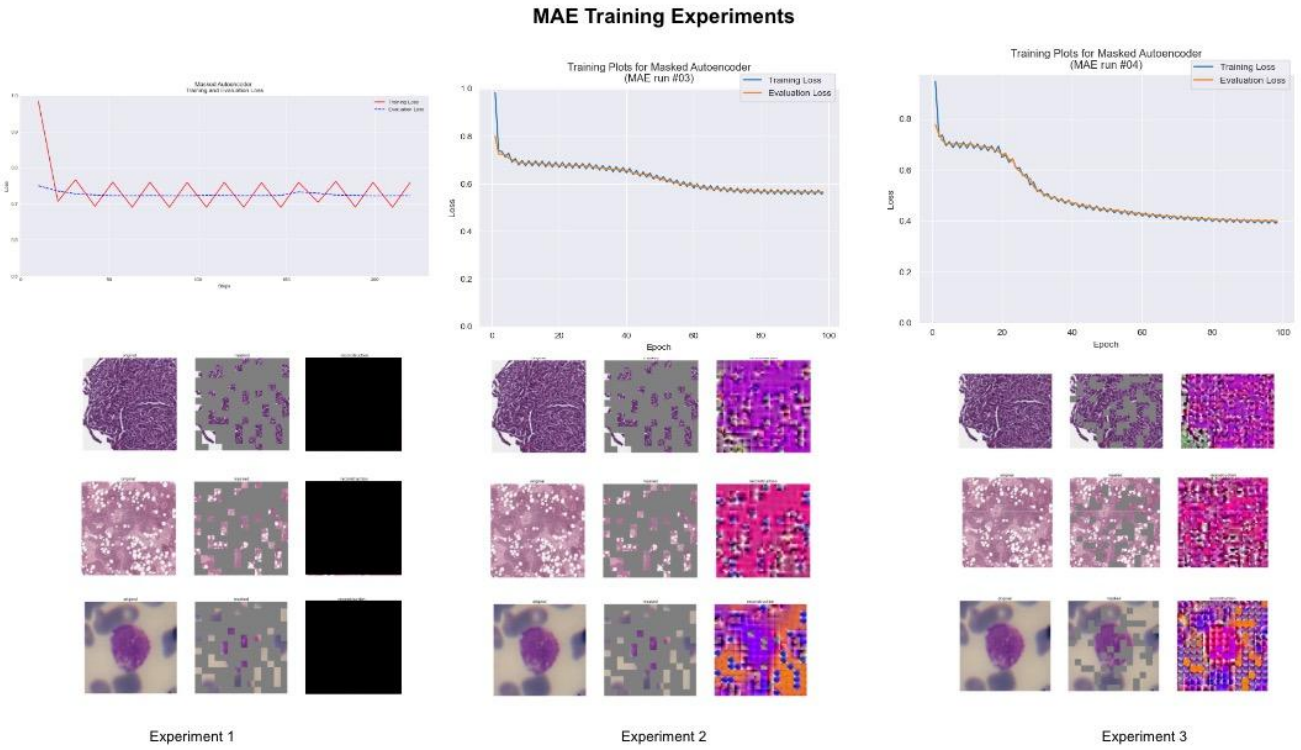
Though these changes facilitated better convergence with the loss consistently decreasing, a plateau was observed in the later phases. This suggested the presence of a local minimum, indicating the need for further optimization.

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 indicates effective hyperparameter tuning.

"Patch Spamming" Phenomenon: An interesting observation is the model's inclination to repeatedly use a specific patch in areas it finds ambiguous. While this behavior is more observable in the unmasked sections during experiment 2, it is predominantly present in the masked segments in experiment 3.

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

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.

Drawing from these observations, we can infer a "masked versus unmasked" tradeoff in the MAE architecture with this dataset. In simpler terms, while the model might excel in reconstructing masked areas, it might simultaneously face challenges in unmasked regions, and vice versa. This interplay offers avenues for further exploration to fine-tune and optimize masking percentage for medical imaging tasks.

ViT Classifiers: Methodology & Results

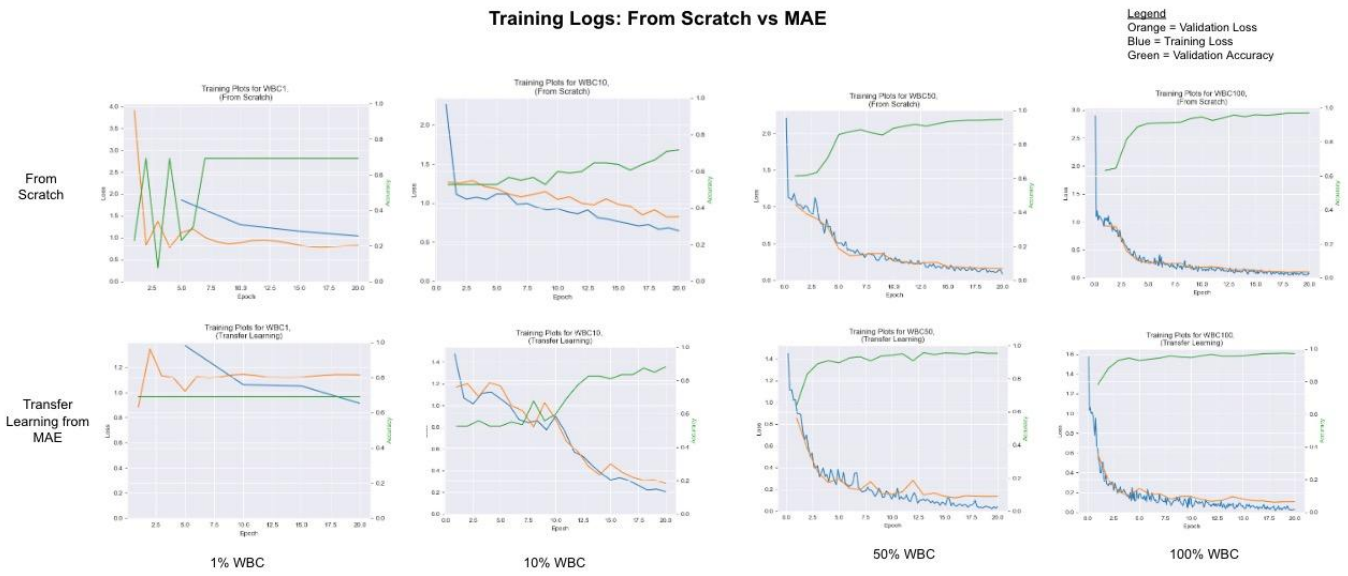
Architecture + Image Augmentation

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:

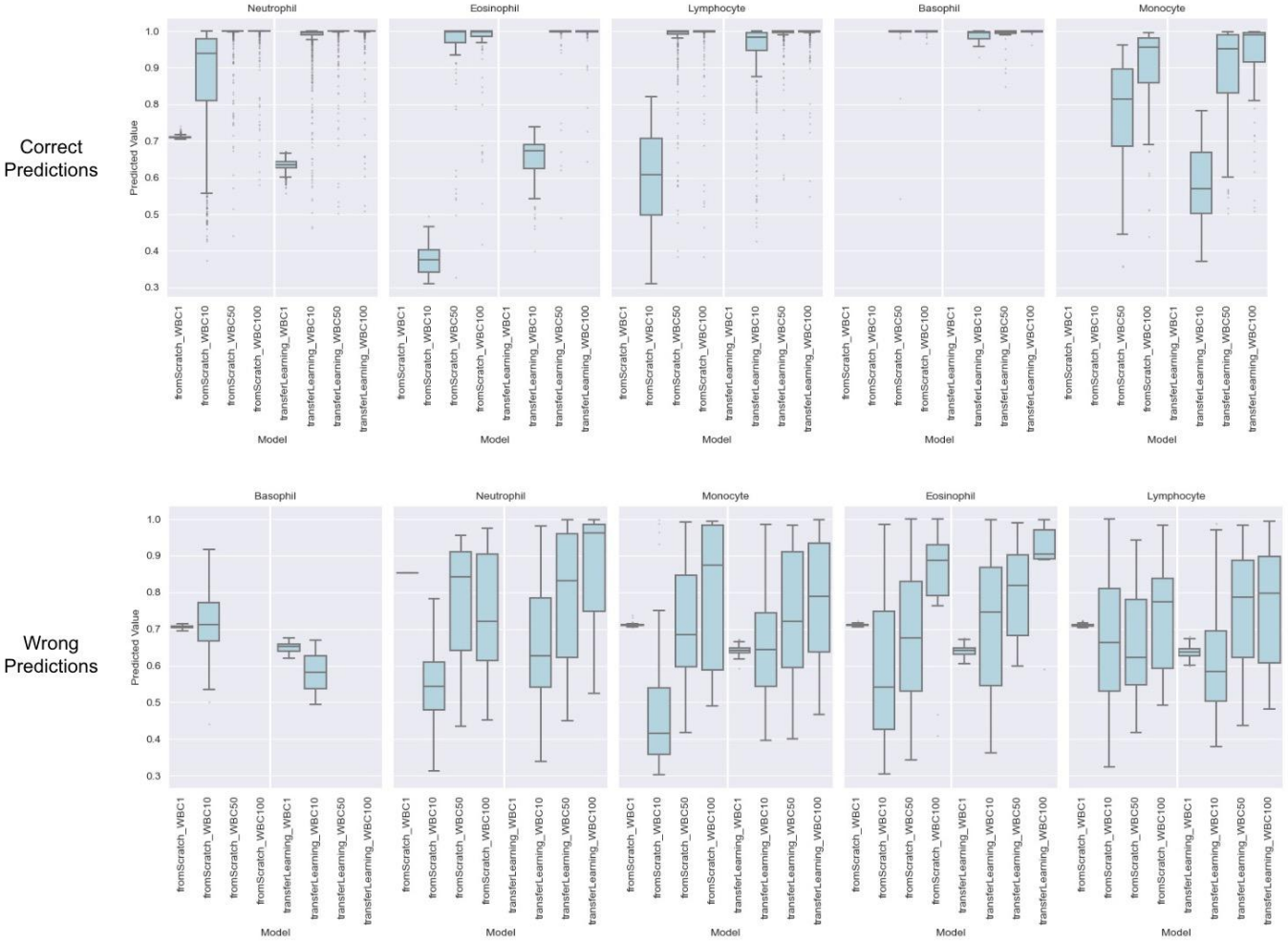
Major Gains with Small Data: At low data volume (1% WBC and 10% WBC), the use of MAE transfer learning consistently outperforms the from-scratch models, especially in precision, with gains reaching as high as 42.63%.

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



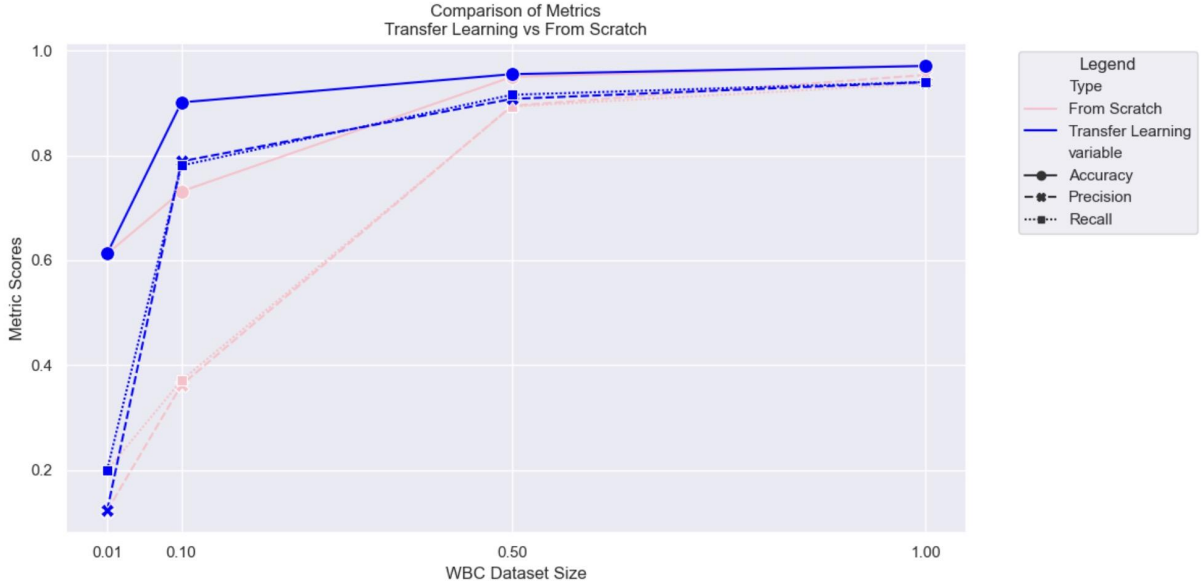
The above table compares the prediction score from the ablation study. From the above, the following observations:

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.

Finally, 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



Overall, we can infer that MAE transfer learning delivers substantial advantages, particularly when training on limited data. But as the dataset size increases, the benefits taper off, while still yielding slight gains in recall. It is especially promising that there are no negative effects to using a pre-trained MAE for transfer learning.

Exploring further opportunities, th task of pre-training a domain-specific MAE should be investigated further.

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.

It will be interesting to further understand if a diversity of morphological characteristics will add value to MAE pretraining. For example, will x-ray imaging help or hinder a downstream cellular pathology task?

Finally, would implementing the methods used to manage pertubations & in-class variations add value 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.