Self-Supervised Vision Transformers for Breast Histopathology Image Embeddings in Invasive Ductal Carcinoma Detection

Faiaz Rahman
CS 482: Applied Machine Learning
Department of Computer Science
Yale University
faiaz.rahman@yale.edu

Abstract

Breast histopathology, which studies diseases in breast tissue such as breast cancer, is both time-consuming and challenging since malignant spots can often be small and difficult to detect amongst the large regions of benign tissue. Self-supervised Vision Transformers have been shown to learn features which contain explicit information about images' semantic segmentation and thus have promising potential in working with breast histopathology data. We apply self-supervised Vision Transformers for generating image embeddings of breast histopathology data and evaluate their performance on the task of detecting invasive ductal carcinoma (IDC). The **DINO** algorithm was developed by Facebook AI for self-supervised training of the Vision Transformer, which is a form of self-**di**stillation with **no** labels. After DINO's self-supervised pretraining, we fine-tune our Vision Transformers on breast histopathology data, and then apply the Vision Transformers to generate image embeddings in the IDC detection architecture. We compare the performance of the DINO-trained Vision Transformers with a baseline model built on ResNet. Additionally, we explore the effects of patch size in the DINO algorithm along with the model size of pretrained DINO models on performance.[1]

1 Introduction

Invasive ductal carcinoma (IDC) is the most common phenotypic subtype of breast cancer, comprising nearly 80% of all breast cancers. Isolating invasive tumor tissue from non-invasive or healthy tissues allows for medical analysis, and precise delineation of IDC is crucial for grading tumor aggressiveness and, in turn, predicting patient outcome (Cruz-Roa et al., 2014). This task, however, is both time-consuming and challenging since malignant spots can often be small and difficult to detect amongst the large regions of benign tissue.

Deep learning is ideally suited for image analysis challenges in breast histopathology, and digital pathology in general, due to its ability to learn representations of the image data without any handcrafted feature engineering (Janowczyk and Madabhushi, 2016). Previous work exploring deep learning approaches for detecting breast cancer has primarily focused on convolutional neural networks (Shen et al., 2019; Cruz-Roa et al., 2014; Tsochatzidis et al., 2019), with some work employing autoencoders (Xu et al., 2014; Feng et al., 2017). These simpler deep learning approaches leave much room for future work with breast histopathology data incorporating more complex vision models. We aim to help advance work on applying more complex deep learning models to breast histopathology by applying DINO-trained Vision Transformers to the task of detecting invasive ductal carcinoma.

2 Related Work

2.1 Vision Transformers

The Transformer (Vaswani et al., 2017) is a self-attention–based architecture which has become the predominant model in natural language processing, given its ability to be pretrained on huge, unlabeled text corpora and then be fine-tuned for specific tasks (Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2020).

In computer vision, convolutional architectures have been dominant, but recent work (inspired by the success of Transformers in NLP) has explored self-attention–based architectures. Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT), which applies the Transformer directly to images by dividing an image into patches

[1] For reproducibility and extensibility in future work, we make our code available at https://github.com/faiazrahman/Self-Supervised-Breast-Histopathology-Transformers.

and providing the sequence of their resulting embeddings to the Transformer as input. Given that Dosovitskiy et al. (2020) further found that ViT had excellent performance on image recognition benchmarks (e.g., ImageNet, CIFAR-100, VTAB), the Vision Transformer is a promising architecture in the future of computer vision and image processing.

2.2 Self-Supervised Learning

Self-supervised learning learns from unlabeled training data through both *generative* and *contrastive* approaches which learn the underlying representations of the data (Jaiswal et al., 2020). Previous work in self-supervised learning explored contrastive, discriminative approaches which group similar samples closer, e.g., through contrastive loss (Chen et al., 2020; He et al., 2019). Grill et al. (2020) explore a different direction for learning unsupervised features by using metric-learning, inspiring the student and teacher networks in the DINO algorithm (Caron et al., 2021).

3 Approach

Our approach involves fine-tuning DINO-based Vision Transformers to generate image embeddings of breast histopathology image scan data, from which we apply a classification head (several fully-connected feedforward layers) for the IDC detection task. We describe the DINO algorithm, the Vision Transformer architecture, and our overall model architecture in the following.

3.1 DINO Algorithm: Self-Supervised Learning with Knowledge Distillation

Caron et al. (2021) developed DINO as a self-supervised method incorporating knowledge distillation for training Vision Transformers (ViTs). *Knowledge distillation* is a learning paradigm where a student network g_{θ_s} parameterized by θ_s is trained to match the output of a given teacher network g_{θ_t} parameterized by θ_t. Given an input image x, both networks output probability distributions P_s and P_t over K dimensions. The output of the network g is normalized with a softmax function to produce the probability P, as follows (where τ_s and τ_t are temperature parameters that control the sharpness of the output distributions).

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s},(x)^{(i)}/\tau_s}{\sum_{k=1}^{K} \exp(g_{\theta_s}(x)^{(k)}/\tau_s)} \qquad (1)$$

$$P_t(x)^{(i)} = \frac{\exp(g_{\theta_t},(x)^{(i)}/\tau_t}{\sum_{k=1}^{K} \exp(g_{\theta_t}(x)^{(k)}/\tau_t)} \qquad (2)$$

Thus, the student network g_{θ_s} learns to match its distributions to that of g_{θ_t} by minimizing the cross-entropy loss.

$$\min_{\theta_s} H(P_t(x), P_s(x)) \qquad (3)$$

Caron et al. (2021) adapt this knowledge distillation problem to *self-supervised learning* by constructing different distorted views (also referred to as crops) of an image by applying a set of transforms to the original image, producing both *global views* and *local views* of smaller resolution. The loss is thus adapted as follows.

$$\min_{\theta_s} \sum_{x\in\{x_1^g, x_2^g\}} \sum_{x'\in V, x'\neq x} H(P_t(x), P_s(x')) \qquad (4)$$

Thus, the model can be pretrained on unlabeled image data by generating these distorted transformations of the original images.

Both networks share the same architecture g with separate sets of parameters θ_s, θ_t. Since this is self-supervised learning, there is no a priori for the teacher network, and thus it is built from past iterations of the student network (e.g., freezing the teacher network over an epoch or a specified number of training steps).

3.2 DINO Vision Transformer Architecture

The architecture g consists of a backbone f (in our experiments, we use only ViT, although Caron et al. (2021) also use ResNet as a backbone) and a projection head $h : g = h \circ f$. ViT (Dosovitskiy et al., 2020) takes as input a grid of non-overlapping contiguous $N \times N$ image patches, passes them through a linear layer to generate embeddings, adds a `[CLS]` token, and feeds them into a standard Transformer network, i.e., a sequence of self-attention and feedforward layers parallelized with skip connections (Vaswani et al., 2017).

3.3 Invasive Ductal Carcinoma Detection Model Architecture

Our DINO-based IDC detection model architecture consists of a DINO Vision Transformer, which is used to generate an embedding of the input breast histopathology image. This image embedding is

Model	Patch Size (DINO)	Accuracy
ResNet (baseline)	—	84.05%
dino-small	16	89.32%
	8	**91.41%**
dino-base	16	90.21%
	8	**92.14%**

Table 1: Results from experiments, reporting the accuracy of each model on the invasive ductal carcinoma (IDC) detection task.

Model	Patch Size	**Performance Improvement** over ResNet baseline model	
		Absolute Accuracy	**Relative Accuracy**
dino-small	16	+5.27%	+6.27%
	8	**+7.36%**	**+8.76%**
dino-base	16	+6.16%	+7.32%
	8	**+8.09%**	**+9.63%**

Table 2: Performance improvement of the DINO-based IDC detection models over the ResNet baseline.

then passed through two fully-connected feedforward layers, mapping from the DINO ViT embedding to a hidden layer, and then to the final output layer with 2 nodes.

For our baseline, our ResNet-based IDC detection model uses ResNet modified such that the last layer is overwritten with a fully-connected feedforward layer mapping to the image feature dimension. This image embedding is then passed through the same architecture as was the DINO image embedding to produce the IDC prediction.

4 Data

Cruz-Roa et al. (2014) developed a dataset of breast cancer whole mount slide images of women diagnosed with IDC at the Hospital of the University of Pennsylvania and the Cancer Institute of New Jersey. Specifically, the data originally contained whole mount slide images of 162 women which were digitized at 40x magnification (0.25 μ/pixel resolution), from which 277,524 patches of size 50 × 50 were extracted, where 198,738 samples were non-IDC and 78,786 were IDC. The dataset is available on Kaggle[2] and via their CL API.

For our DINO-based models, we resize the images to 224 × 224, since that is what the DINO ViT models expect as input. For our ResNet baseline model, we apply a normalization transform as per the `torchvision` documentation.[3]

5 Experiments

We run experiments to (1) compare the performance of DINO-based Vision Transformer models against a ResNet baseline, (2) explore the effect of patch size on DINO ViT's performance, and (3) explore the effect of initial model size on DINO ViT's performance.

5.1 Models

We continue fine-tuning from Facebook AI's initial DINO ViT checkpoints and use the resulting fine-tuned models in our IDC classification architecture. For our baseline model, we use ResNet-152 via `torchvision.models.resnet152` in our IDC classification architecture.

5.2 Experiment Settings

We train on 2–8 NVIDIA GeForce RTX 3090 GPUs with Driver version 470.103.01 and CUDA version 11.4. We use PyTorch (version 1.11.0 with CuDNN version 8.2.0) to implement our models, including using PyTorch Lightning for our model training and evaluation. We run our training and evaluation on multiple GPUs in data parallel (i.e., splitting each batch across all GPUs specified for the experiment), with batch size of 64 (which, after also trying 32 and 128, we found maximized GPU utilization without exceeding CUDA memory, with each GPU processing an even number of items per batch and the root node aggregating the

[2] https://www.kaggle.com/paultimothymooney/breast-histopathology-images

[3] https://pytorch.org/docs/stable/torchvision/models.html

Model	Patch Size	Performance Improvement over respective models with patch size 16	
		Absolute Accuracy	Relative Accuracy
dino-small	8	+2.09%	+2.34%
dino-base	8	+1.93%	+2.11%

Table 3: Exploring the effect of patch size in the DINO Vision Transformer on performance in the IDC detection task. The smaller patch size of 8 performed better than 16 for both DINO model sizes.

results). We use a learning rate of `1e-4` and train for 5 epochs. We use Adam as our optimizer and cross-entropy as our loss function. We save model checkpoints every 100 train steps using PyTorch Lightning callbacks.

We ran hyperparameter tuning and found that learning rates of `1e-2`, `1e-3`, and `1e-5`, along with using SGD as the optimizer, did not perform as well, with the loss either not decreasing or overall performance simply being lower than Adam with learning rate `1e-4`.

6 Results and Analysis

6.1 IDC Detection Performance and Improvements

The results of our experiments are recorded in Table 1. We find that all DINO-based Vision Transformer models outperform the ResNet baseline, which achieved an accuracy of 84.05%. Specifically, the best-performing DINO models were `dino-base` with patch size of 8, which achieved an accuracy of 92.14%, and `dino-small` with patch size of 8, which achieved an accuracy of 91.41%. The performance improvements in absolute accuracy and relative accuracy for all DINO-based models are recorded in Table 2.

We hypothesize that our DINO-based Vision Transformer models were able to outperform the ResNet baseline due to the ability of the DINO algorithm's self-supervised learning to learn explicit information about the semantic segmentation of an image, which Caron et al. (2021) find does not emerge as clearly with supervised pretraining of other computer vision models. Such image segmentation is particularly useful on breast histopathology data in the IDC detection task, since the model aims to find segments of malignant tissue.

6.2 Exploring the Effect of Patch Size

Additionally, we can explore the effect of patch size in our DINO-based Vision Transformer models in performance on the IDC detection task. As explained in 3.2, the Vision Transformer model takes an image, produces a grid of $N \times N$ image patches, and then passes them through a linear layer to generate embeddings, which are in turn fed into a standard Transformer network. This patch size, in turn, affects the scale at which the Vision Transformer "sees" the image, along with the overall length of the generated embeddings sequence.

Our results indicate that for the IDC detection task, the *smaller* patch size of 8 (versus the larger patch size of 16) performs better. Specifically, the `dino-small` model with patch size 8 outperformed the respective model with patch size 16 (which had accuracy of 89.32%) by 2.09% absolute accuracy and 2.34% relative accuracy. The `dino-base` model with patch size 8 outperformed the respective model with patch size 16 (which had accuracy of 90.21%) by 1.93% absolute accuracy and 2.11% relative accuracy. Thus, the `dino-small` model had slightly greater improvements for its larger-patch version than the `dino-base` model did for its larger-patch version. These improvements are recorded in Table 3.

We hypothesize that the models with smaller patch size performed better since, as Cruz-Roa et al. (2014) described, malignant spots can often be small and difficult to detect amongst the large regions of benign tissue. A smaller patch size increases the number of overall patches which are saturated with malignant tissue, and even though there would be more patches overall, those malignant patches (which are "confidently" determined by the model to be malignant) help improve the overall IDC detection performance.

6.3 Exploring the Effect of Model Size

From our results in Table 1, we can see that, as likely expected, the larger `dino-base` model outperformed the `dino-small` model. However, the model size of the initial DINO Vision Transformer did not signficantly affect overall performance: dino-base had a 0.89% greater absolute accuracy than dino-small for patch size 16, and 0.73% for patch size 8. It is also important to note that the base-sized model had 86.8M parameters, while the small-sized model had 20.0M parameters; thus, the tradeoff in computational cost (e.g., compute resources, training time, trained model size) may not be worth the slight increase in performance. We hypothesize that the additional parameters in the base-sized model likely did not learn much additional information for this specific task of IDC detection, but may be more useful for other downstream tasks using DINO-based Vision Transformers.

7 Conclusion

Efficiently and effectively processing breast histopathology data allows for improved detection of breast cancer and invasive tissue, which in turn can improve medical analysis, the grading of tumor aggressiveness, and overall predictions of patient outcome. Our experiments illustrated the powerful potential of Vision Transformers (ViTs) trained through the DINO self-supervised learning algorithm in better detecting invasive ductal carcinoma (IDC), outperforming a ResNet baseline. Our experiments also show the effect of patch size in the DINO-based Vision Transformers on overall performance, where models with the smaller patch size of 8 slightly outperformed those with patch size of 16.

Further work can be done experimenting with various backbones in the DINO ViT architecture (as described in 3.2), various additional patch sizes, and pretraining on other kinds of breast histopathology data. Additionally, future work can apply Vision Transformers to other digital pathology tasks to see if the promising results shown in this paper can further extend to other areas of medicine, leveraging the computational power of cutting-edge deep learning architectures towards improving medical diagnoses and healthcare.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. 2014. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *SPIE Proceedings*. SPIE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Yangqin Feng, Lei Zhang, and Zhang Yi. 2017. Breast cancer cell nuclei classification in histopathology images using deep neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 13(2):179–191.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *CoRR*, abs/2011.00362.

Andrew Janowczyk and Anant Madabhushi. 2016. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1):29.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. 2019. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9(1).

Lazaros Tsochatzidis, Lena Costaridou, and Ioannis Pratikakis. 2019. Deep learning for breast cancer diagnosis from mammograms—a comparative study. *Journal of Imaging*, 5(3).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jun Xu, Lei Xiang, Renlong Hang, and Jianzhong Wu. 2014. Stacked sparse autoencoder (ssae) based framework for nuclei patch classification on breast cancer histopathology. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 999–1002. IEEE.