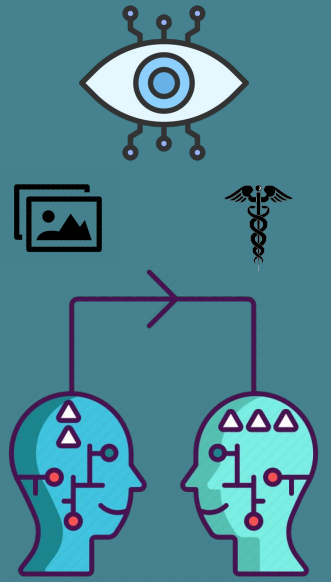
Università
degli Studi
di Padova

Self-supervised Transfer Learning for Medical Image Classification



Lorenzo Baietti
ID: 2130676

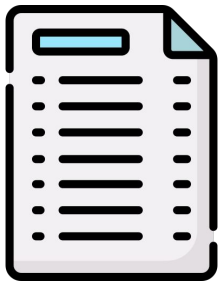
Francesco Carlesso
ID: 2125806

MSc Data Science

Vision and Cognitive Systems

A.Y. 2024/2025

Index



Introduction

Introduction

Motivation:

- Medical imaging datasets require expert-labeled annotations, which can be quite costly.

Approach:

- Self-supervised Learning (SSL), which extracts representations from unlabeled data by leveraging pretext tasks.

Objective:

- Demonstrate the effectiveness of SSL using a pre-trained BEiT model (Bidirectional Encoder representations from Image Transformers) with transfer learning in a domain-shifted task.

Related Work

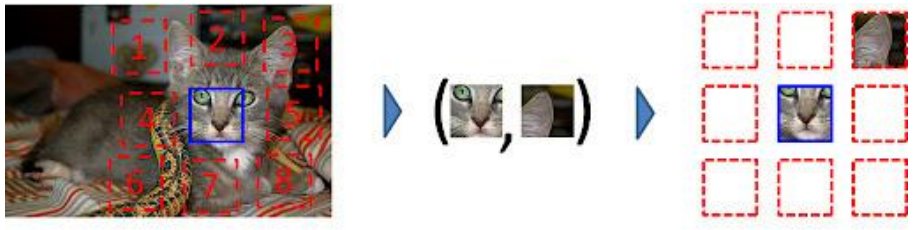
Predictive Learning

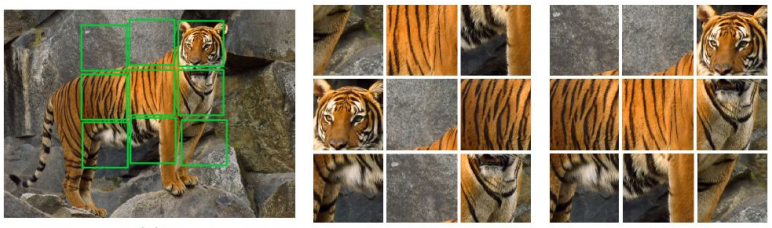
Predictive Learning is one of the foundational approaches in SSL, where models are trained to predict missing or occluded information in input data.

Seminal Papers:

- ”Unsupervised Learning by Context Prediction” (Doersch et al.)
- “Unsupervised learning of visual representations by solving jigsaw puzzles” (Nor. and Fav.)

Focus on capturing semantic information and spatial hierarchies, which are essential for downstream visual understanding tasks.





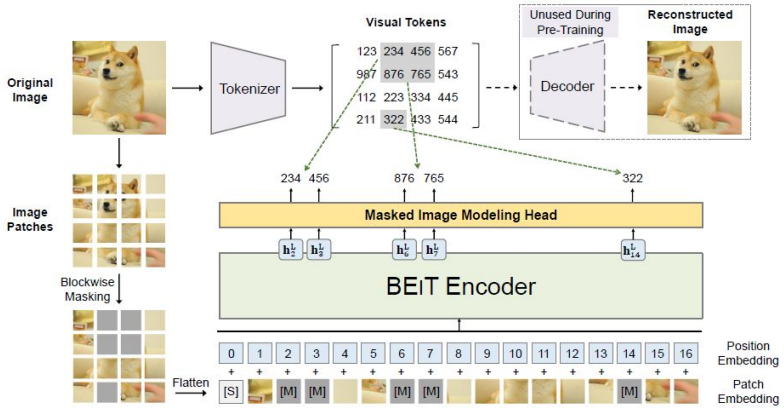
Predictive Learning

Predictive Learning is one of the foundational approaches in SSL, where models are trained to predict missing or occluded information in input data.

State-of-the-art:

- “Beit: Bert pre-training of image transformers” (Bao et al.)

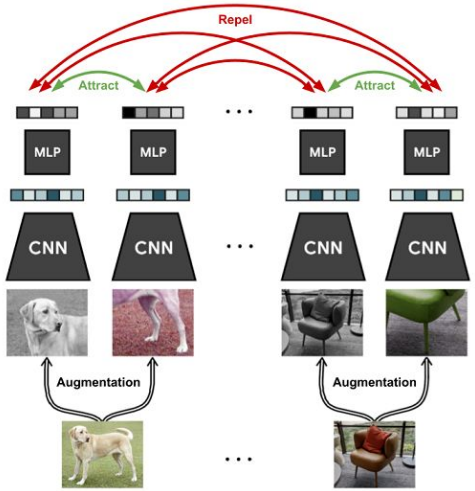
The model is trained to predict visual tokens using a tokenizer to discretize image data, and it learns to reconstruct masked image patches through a discrete variational autoencoder.



Predictive Learning vs. Contrastive Learning

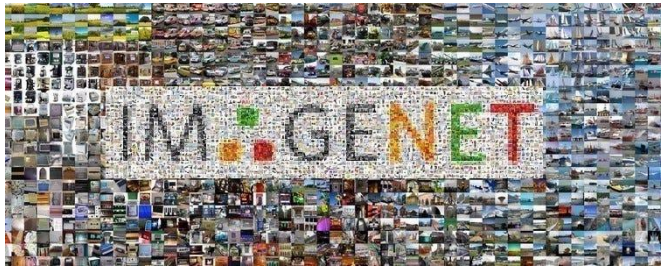
Predictive Learning focuses on contextual information, making it particularly suitable for dense prediction tasks, but not only.

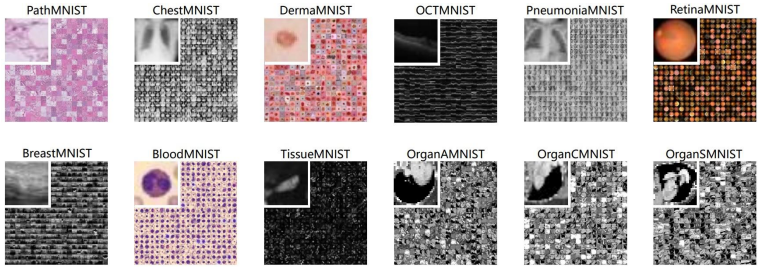
Contrastive Learning is another dominant approach in SSL (e.g. SimCLR), it focuses on instance-level discrimination.



Self-supervised Transfer Learning

Recent advances in self-supervised transfer learning have demonstrated the efficacy of the relative models in domain adaptation, showing that self-supervised models pre-trained on large datasets, such as ImageNet, can generalize well to domain-shifted tasks with minimal fine-tuning.





Self-supervised Transfer Learning

- Explored by MoCo (Momentum Contrast) and DINO (self-DIstillation with NO labels) models.
- “Big self-supervised models advance medical image classification” (Azizi et al.) applied contrastive pre-training, achieving good results on several benchmarks.

Our work aims at further testing the outlined discoveries, using a slightly different method to confirm the upsides of taking advantage of self-supervised pre-training.

Datasets

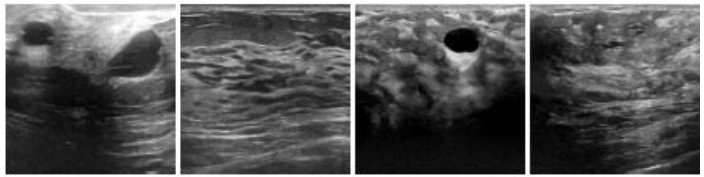
MedMNIST

MNIST-like collection of biomedical datasets for diverse classification tasks, covering various modalities, scales, and resolutions.

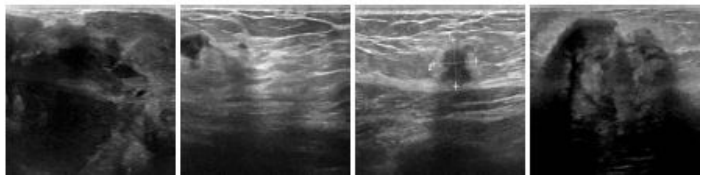
BreastMNIST:

Dataset of 780 breast ultrasound images for tumor diagnosis, categorized into 3 classes: normal, benign, and malignant. Simplified to a binary classification.

Normal or Benign:



Malignant:



MedMNIST

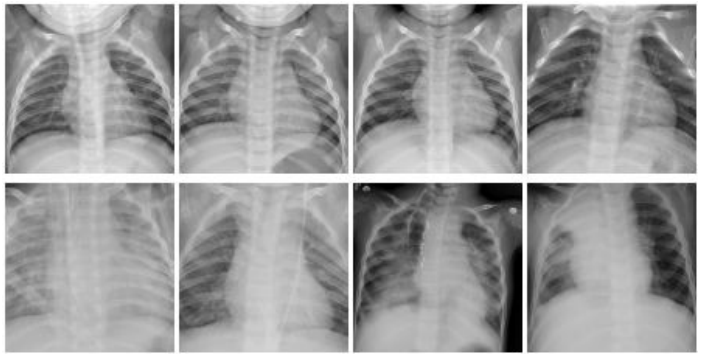
MNIST-like collection of biomedical datasets for diverse classification tasks, covering various modalities, scales, and resolutions.

PneumoniaMNIST:

Dataset of 5,856 pediatric chest X-Ray images for pneumonia diagnosis, where the task is binary classification of pneumonia against normal conditions.

Normal:

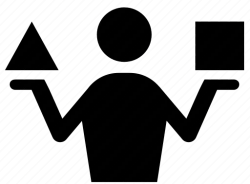
Pneumonia:



Method

Method

1) Freeze most layers of the pre-trained BEiT
2) Fine-tune (train) in a supervised way on the medical task for each input condition
3) Test computing the evaluation metrics
4) Compare BEiT methods with supervised methods benchmarks

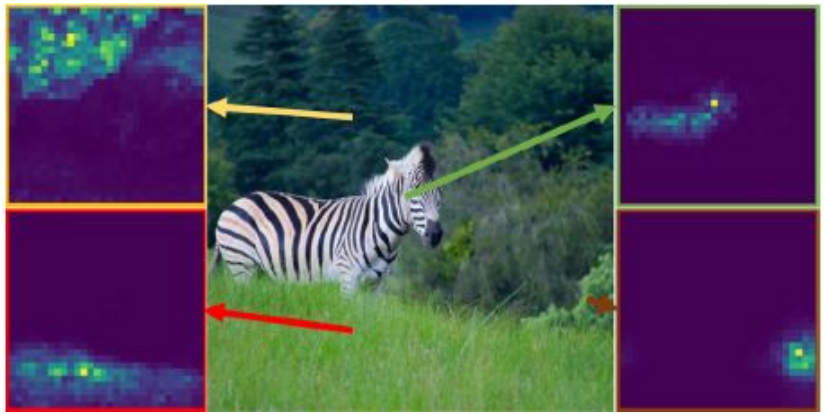


Self-supervised Learning with BEiT

BEiT-large:

- ➢ 24 transformer layers, hidden size of 1024, and 16 attention heads, total of 304M parameters
- ➢ Pre-trained on 14 million 224x224 images from ImageNet

Vision Transformer (ViT) backbone employing self-attention maps, which are not biased by any label, so highly transferable.



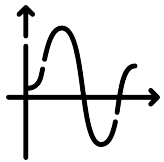
Self-supervised Learning with BEiT

Masked Image Modeling (MIM) with 16x16 image patches, so 196 per image, to predict the corresponding visual tokens:

$$\hat{x}_m = \arg\max_{x \in \mathcal{V}} P(x|\mathbf{x}_{\setminus m})$$

Reconstruction-based loss:

$$\mathcal{L}_{\text{MIM}} = -\sum_{m \in \mathcal{M}} \log P(\hat{x}_m|\mathbf{x}_{\setminus m})$$

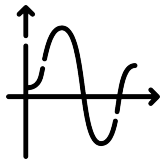


Fine-tuning

This process bridges the gap between the self-supervised pre-training objective and the supervised downstream classification objective, enabling the model to specialize in the new domain while leveraging the general purpose features learned. Transitioning from the reconstruction-based objective of MIM to the cross-entropy loss for classification, aligning the learned representations to the class labels of the medical datasets.

Binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

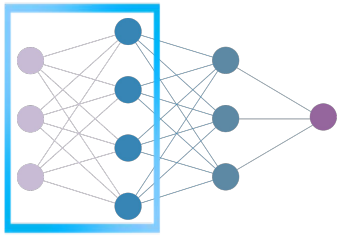


Fine-tuning

The BEiT model is fine-tuned on the datasets by adjusting the classifier head, the pooler, and the top 5 transformer layers, while keeping all of the others frozen.

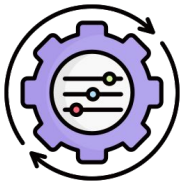
➢ Approximately 63M trainable parameters, around 20% of the architecture.

Pre-trained features act as a strong foundation, ensuring that the model efficiently learns the decision boundaries in the new domain without overfitting.



Hyperparameters:

➔ Epochs = 5
➔ Learning rate = 3e−4
➔ Batch size and Weight decay are adapted on each method



Evaluation Metrics

Area Under the Curve (AUC):

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds. It provides a single scalar value summarizing the curve, which ranges from 0.5 (random guessing) to 1.0 (perfect classification).

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \qquad \text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

Accuracy:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

Benchmarks

The benchmarks given mainly represent supervised learning approaches:

- **ResNet-18:** widely used CNN which mitigates the vanishing gradient problem.
- **ResNet-50:** deeper variant of ResNet which is able to learn more complex features.
- **auto-sklearn:** statistical ML framework leveraging Bayesian optimization and meta-learning.
- **AutoKeras:** deep learning AutoML framework leveraging neural-architecture-search.
- **Google AutoML Vision:** black-box AutoML tool leveraging proprietary transfer learning.

Experiments

Experiments

Class Imbalance & Undersampling

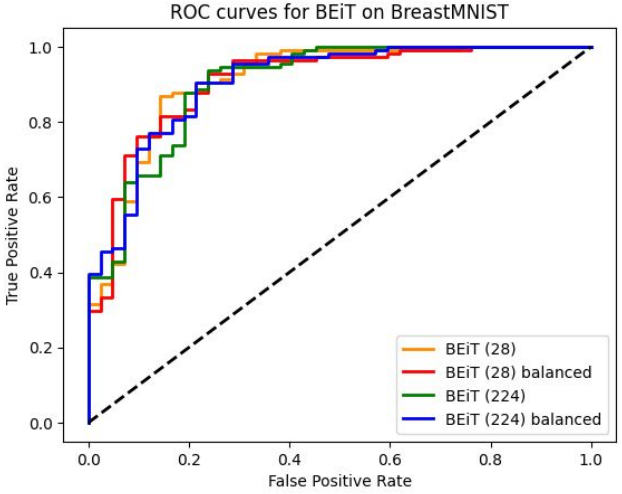
- ➔ Applied undersampling to balance classes.
- ➔ BreastMNIST: Adjusted to 60/40 to avoid excessive sample size reduction.
- ➔ PneumoniaMNIST: Achieved 50/50 given the larger dataset.

Preprocessing for BEiT Fine-tuning

- ➔ BEiT model was pre-trained on RGB images (224×224 resolution).
- ➔ Fine-tuning requires inputs to follow the same format.
- ➔ Hugging Face preprocessor:
  - ◆ Automatically resizes images to 224×224.
  - ◆ Normalizes across RGB channels.

BreastMNIST

Method	AUC	Accuracy
⋆ BEiT (28)	0.910	**0.891**
† BEiT (28) balanced	0.910	0.872
† BEiT (224)	0.904	0.885
† BEiT (224) balanced	0.908	0.878
ResNet-18 (28)	0.901	0.863
ResNet-18 (224)	0.891	0.833
ResNet-50 (28)	0.857	0.812
ResNet-50 (224)	0.866	0.842
auto-sklearn	0.836	0.803
AutoKeras	0.871	0.831
Google AutoML Vision	**0.919**	0.861

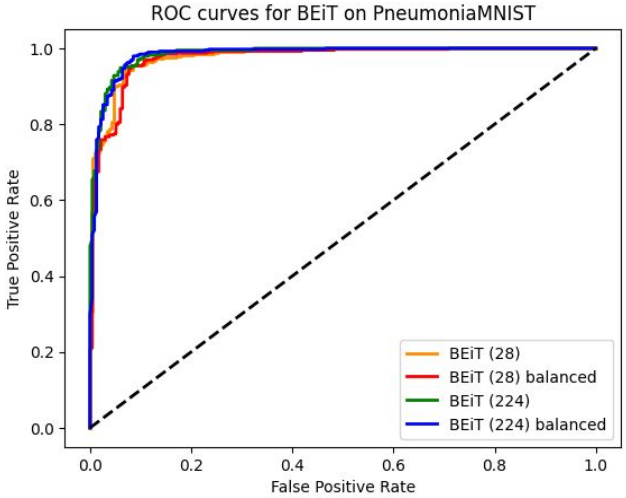


➔ **Balanced training** does not improve performance (likely due to the reduction of samples).

➔ **Resizing** benefits performance by smoothing images and reducing noise.

PneumoniaMNIST

Method	AUC	Accuracy
† BEiT (28)	0.979	0.881
† BEiT (28) balanced	0.974	0.926
† BEiT (224)	0.985	0.912
⋆ BEiT (224) balanced	0.983	0.918
ResNet-18 (28)	0.944	0.854
ResNet-18 (224)	0.956	0.864
ResNet-50 (28)	0.948	0.854
ResNet-50 (224)	0.962	0.884
auto-sklearn	0.942	0.855
AutoKeras	0.947	0.878
Google AutoML Vision	**0.991**	**0.946**



➔ **ROC plots** show only slight variations, and excellent performance overall.

➔ **Balancing** helps accuracy while negligibly reducing AUC.

Conclusion

Conclusion

Findings:

- ★ Strong generalization capabilities from ImageNet pre-training
- ★ Competitive performance against supervised methods
- ★ Self-supervised transfer learning is highly viable
- ★ Predictive learning with ViTs is effective in medical imaging

Future Work:

- ❖ Experiment with hybrid SSL approaches (predictive + contrastive learning)
- ❖ Test on larger and more diverse datasets
- ❖ Improve interpretability to build trust for clinical applications

Thank You!