The American Journal of Pathology, Vol. 194, No. 5, May 2024

ELSEVIER

The American Journal of
PATHOLOGY

ajp.amjpathol.org

MACHINE LEARNING, COMPUTATIONAL PATHOLOGY, AND BIOPHYSICAL IMAGING

Model-Agnostic Binary Patch Grouping for Bone Marrow Whole Slide Image Representation

Youqing Mu,[*][†] Hamid R. Tizhoosh,[‡] Taher Dehkharghanian,[†][§] Saghir Alfasly,[‡] and Clinton J.V. Campbell[†]

From the Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario, Canada; the Department of Pathology and Molecular Medicine,[†] McMaster University, Hamilton, Ontario, Canada; the Rhazes Lab,[‡] Artificial Intelligence and Informatics, Mayo Clinic, Rochester, Minnesota; and the Department of Nephrology,[§] University Health Network, Toronto, Ontario, Canada*

Accepted for publication January 10, 2024.

Address correspondence to Clinton J.V. Campbell, M.D., Ph.D., Juravinski Hospital, 711 Concession St., Hamilton, ON L8V1C3, Canada.
E-mail: campbecj@mcmaster.ca.

Histopathology is the reference standard for pathology diagnosis, and has evolved with the digitization of glass slides [ie, whole slide images (WSIs)]. While trained histopathologists are able to diagnose diseases by examining WSIs visually, this process is time consuming and prone to variability. To address these issues, artificial intelligence models are being developed to generate slide-level representations of WSIs, summarizing the entire slide as a single vector. This enables various computational pathology applications, including interslide search, multimodal training, and slide-level classification. Achieving expressive and robust slide-level representations hinges on patch feature extraction and aggregation steps. This study proposed an additional binary patch grouping (BPG) step, a plugin that can be integrated into various slide-level representation pipelines, to enhance the quality of slide-level representation in bone marrow histopathology. BPG excludes patches with less clinical relevance through minimal interaction with the pathologist; a one-time human intervention for the entire process. This study further investigated domain-general versus domain-specific feature extraction models based on convolution and attention and examined two different feature aggregation methods, with and without BPG, showing BPG's generalizability. The results showed that using BPG boosts the performance of WSI retrieval (mean average precision at 10) by 4% and improves WSI classification (weighted-F1) by 5% compared to not using BPG. Additionally, domain-general large models and parameterized pooling produced the best-quality slide-level representations. *(Am J Pathol 2024, 194: 721–734; https://doi.org/10.1016/j.ajpath.2024.01.012)*

Histopathology is the extraction of microscopic information from preserved sections of human tissue by a pathologist, providing information about the biological state of a patient. It is considered the reference standard method for supporting a diagnosis in pathology.[1] Digitizing tissue glass slides into whole slide images (WSIs), known as digital pathology, has revolutionized computational pathology techniques, enabling the efficient processing and storage of large amounts of histopathology data.[2]

Trained histopathologists can make an accurate tissue assessment by visually inspecting WSIs on high-resolution settings. However, the large spatial dimensionality of WSIs makes analyzing these images a time-consuming and inefficient process. In response to this challenge, scientists in this domain have embraced automated analysis of WSIs as a means to quantify pertinent anatomic structures. Deep learning models (ie, deep neural networks) have emerged as a promising solution, providing state-of-the-art performance in a wide variety of tasks, including tumor grading, tumor segmentation, and disease prediction.[3–6] However, despite the potency of deep learning, several challenges persist. The lack of labeled WSI data, heterogeneity in images, and complexity of features all pose technical difficulties that need to be overcome to achieve reliable results.[7] Thus, pipelines are designed to address these issues and to make deep learning models more practical for the histology-specific domain.[8]

Previous studies involving WSI processing and analysis focused primarily on patch-level classification,

Supported by New Frontiers and Research Fund Exploration grant NFRFE-2019-00754.

https://doi.org/10.1016/j.ajpath.2024.01.012

original glass slides, the remaining 44 WSIs with fewer than 64 patches were disregarded.

WSI-level labels were generated by simplifying the predictions of a fine-tuned Bidirectional Encoder Representations from Transformers language model on WSI synopses, as previously reported by Mu et al.[33] The Bidirectional Encoder Representations from Transformers language model predictions took the form of a multilabel task. In this work, these predictions were simplified to broad diagnostic categories using the following rules.

- If normal is in a multilabel prediction, this prediction will be simplified as normal.
- If any label is acute leukemia related, this prediction will be simplified as acute leukemia (eg, acute myeloid leukemia; hypercellular becomes acute leukemia).
- If any label is myelodysplastic syndrome related, this prediction will be simplified as myelodysplastic syndrome (eg, hypercellular; myelodysplastic syndrome becomes myelodysplastic syndrome).
- If any label is plasma cell neoplasm related, this prediction will be simplified as plasma cell neoplasm (eg, hypercellular; plasma cell neoplasm becomes plasma cell neoplasm).
- If any label is lymphoproliferative disorder related, this prediction will be simplified as lymphoproliferative disorder (eg, hypercellular; lymphoproliferative disorder becomes lymphoproliferative disorder).
- If any label is myeloproliferative neoplasm related, this prediction will be simplified as lymphoproliferative disorder (eg, myeloproliferative neoplasm; fibrosis becomes myeloproliferative neoplasm).

The 40 WSIs that did not meet the previously mentioned criteria were discarded because their label groups had fewer than 10 WSIs, making it inadequate to generate new classes. As a result, a primary data set was generated consisting of 633 WSI and label pairings for experiments on a slide-level representation pipeline (Supplemental Table S1). For these experiments, five-repeat Monte Carlo cross-validation[34] was performed. For each cross-validated fold, 633 WSIs were partitioned randomly into a training set (50% of cases; ie, 316 WSIs) and a test set (50% of cases; ie, 317 WSIs) and maintained the proportion of classes consistently in each set.

Feature Extractors

The performance of the pipeline arguably is affected by the choice of domain-general (DG) versus domain-specific (DS) pretrained models as feature extractors. The main differences in the pretrained models are in their training data set and model complexity (ie, size). Typically, DG data sets are large and diverse data sets because it is easier to collect general data than specific data such as medical images. For instance, ImageNet, a general data set, comprises more than 1.2 million labeled images belonging to 1000 classes,[35] whereas The Cancer Genome Atlas, a pathology-specific data set, comprises approximately 32,000 hematoxylin and eosin−stained images from more than 11,000 specimens belonging to 32 classes (subtypes).[36] Previous studies have shown that DS models trained on domains closely related to the transfer-learning domain outperform DG models.[37,38] However, large models trained on these relatively small DS data sets, compared with smaller models, have a higher probability of overfitting and being sensitive to input changes,[37] even though empiric evidence suggests that larger models may outperform small models. However, as pretrained methods shift from supervised to self-supervised learning (SSL),[39,40] this paradigm may no longer hold. Unlike supervised learning, SSL is designed to learn representations, which are less affected by domain differences.[41] Thus, a DG large model trained with SSL potentially can outperform DS models in transfer learning because it produces richer features for the downstream model to learn and is less prone to overfitting thanks to its larger and more diverse training data sets.

The study used pretrained models from the two main deep network architectures, namely convolution and transformer, including DenseNet-121[22] and KimiaNet[38] to represent convolution in DG and DS models, respectively. In addition, the ViT-16/256 model from HIPT_4K (ViT-16/256)[42] and the ViT-S/16 model from DINO[32] were used to represent a transformer in DG and DS models, respectively. To distinguish the vision transformer models (ViT) between DINO's from HIPT_4K's, DINO's ViT was named DINO in this study. Patches (512 × 512 pixels) at ×40 magnification were resized to 256 × 256 pixels to align with the ViT-16/256 model's size and ×20 magnification input condition and DINO's input size condition.

DenseNet-121 and KimiaNet are convolutional neural network−based networks with the same architecture, whereas HIPT (ViT-16/256) and DINO are built on the transformer architecture.[43] DenseNet-121 and KimiaNet have 7,978,856 parameters (relatively small), whereas HIPT (ViT-16/256) and DINO have 21,665,664 parameters (ie, 270% larger). Both KimiaNet and ViT-16/256 were pretrained on histopathology data (The Cancer Genome Atlas, DS), whereas DenseNet-121 and DINO were pretrained only on ImageNet (DG) without specific domain adaptation to histopathology images. By using these four different feature extractors, this study compared the performance of convolutional neural network and transformer models and the impact of DS versus DG feature extractors on the final results (ie, WSI representation quality). This experimental setting was designed to enable verification of the robustness and generalizability of this method, regardless of the feature extractor chosen.

BPG

BPG training was prepared by first randomly sampling n patches (ie, n feature vectors) per slide from all 717 scanned WSIs, except for one randomly chosen WSI that served as a