Multimodal Representations of Biomedical Knowledge from Limited Training Whole Slide Images and Reports using Deep Learning

Prabal Ghosh$^{1[0009-0004-3449-5811]}$

Universite Cote d'Azur, Sophia Antipolis, France
prabal5ghosh@gmail.com

Abstract. Digital pathology has ushered in a new era of multimodal biomedical data acquisition, combining high-resolution whole slide images (WSIs) and detailed pathology reports. This paper presents a deep learning framework that fuses these two modalities to form a unified representation, thereby enhancing classification, retrieval, and interpretability in computational pathology. In this report, we summarize the paper's motivation, methodology, experimental evaluation, and contributions while providing a critical discussion relative to the state of the art. I also propose potential improvements to address identified weaknesses.

Keywords: Whole Slide Images (WSIs), Pathology Reports, Biomedical Imaging, Computational Pathology, PubMedBERT, Attention-based Deep Multiple Instance Learning (ADMIL), Histopathology.

1 Introduction

The advent of digital pathology has enabled the routine digitization of whole slide images (WSIs), capturing detailed cellular and tissue structures alongside comprehensive textual pathology reports that summarize clinical observations. Despite the complementary nature of these modalities, most existing deep learning methods in computational pathology focus exclusively on images, missing out on the semantic richness contained in text [1] [2]. Additionally, obtaining large-scale annotated datasets in the biomedical domain is challenging and expensive.

This paper addresses these challenges by proposing a multimodal deep learning framework that fuses WSIs with corresponding pathology reports. The objective is to develop a network that effectively combines visual and textual data to produce a robust joint representation, achieve superior classification performance compared to unimodal models, and support downstream tasks such as cross-modal retrieval and concept linking for visual ontology creation[3].

The contributions of this work include a dual-branch network architecture, a shared projection head for modality alignment, and a composite loss function that integrates supervised and self-supervised components. These methods have been demonstrated on both internal hospital datasets and publicly available datasets. This report summarizes the theoretical and experimental contributions of the paper and critically evaluates the proposed method in the context of current research.

2 Methodology

2.1 Data Processing

Image Data (WSIs) For image data, the method begins with processing whole slide images (WSIs), which are gigapixel images that need to be divided into manageable portions. These WSIs are split into smaller patches, typically measuring 224×224 pixels, at a magnification of 10×. To ensure that only regions containing tissue are processed, tissue segmentation tools such as HistoQC are applied, thereby excluding irrelevant background areas. In order to increase the diversity and robustness of the training data, various augmentations are performed on the image patches using the Albumentations library. These augmentations include applying random rotations at angles of 90°, 180°, and 270°, as well as horizontal and vertical flipping and adjustments in hue, saturation, and contrast.

Text Data (Pathology Reports) The processing of text data involves several steps to ensure that the pathology reports, which may originally be written in languages such as Italian or Dutch, are uniformly represented. Initially, these reports are translated into English using MarianMT models. Once translated, the text is tokenized using BERT's WordPiece tokenizer, with a maximum sequence length set to 512 tokens. To further enhance the variability of the text data, three augmentation techniques are employed. First, a back-to-back translation is performed where the reports are translated into one or more intermediate languages, such as French, German, or Spanish, and then translated back to English. Second, an insert or rephrase strategy is applied using the nlpaug library, where slight modifications, including insertions and paraphrasing, are introduced. Finally, the reports are also rephrased using GPT-3 (specifically via the text-davinci-003 backend), ensuring that the semantic content is maintained while the textual expression is varied.

2.2 Network Architecture

Image Branch In the image branch, the framework employs a convolutional neural network (CNN) that uses a ResNet34 backbone to extract deep features from each image patch. Whole slide images (WSIs) are initially divided into smaller patches, and each patch is processed by the CNN to capture important visual features. After obtaining patch-level features, the model applies an attention-based deep multiple instance learning (ADMIL) framework. This mechanism aggregates the features from all patches by assigning higher weights to those patches that are diagnostically significant. The result is a single, fixed-size, 128-dimensional embedding that represents the entire slide, effectively summarizing the complex visual information present in the WSI.

Text Branch The text branch processes pathology reports using a BERT-based encoder, specifically PubMedBERT, which is pretrained on biomedical literature to capture domain-specific language nuances. The report text is first tokenized, and then the encoder processes the sequence to generate contextualized representations. From the output, the final hidden state corresponding to the special [CLS] token is extracted as a summary representation of the entire report. However, since this embedding is originally 768-dimensional, a projection layer is applied to reduce its dimensionality to 128. This projection ensures that the text representation has the same dimensionality as the image embedding.

Shared Projection and Fusion Both image and text embeddings are fed into a single, shared projection head, aligning them into a common latent space. This unified representation is essential for effective multimodal learning and supports downstream tasks such as cross-modal retrieval and concept linking by combining complementary information from both modalities.

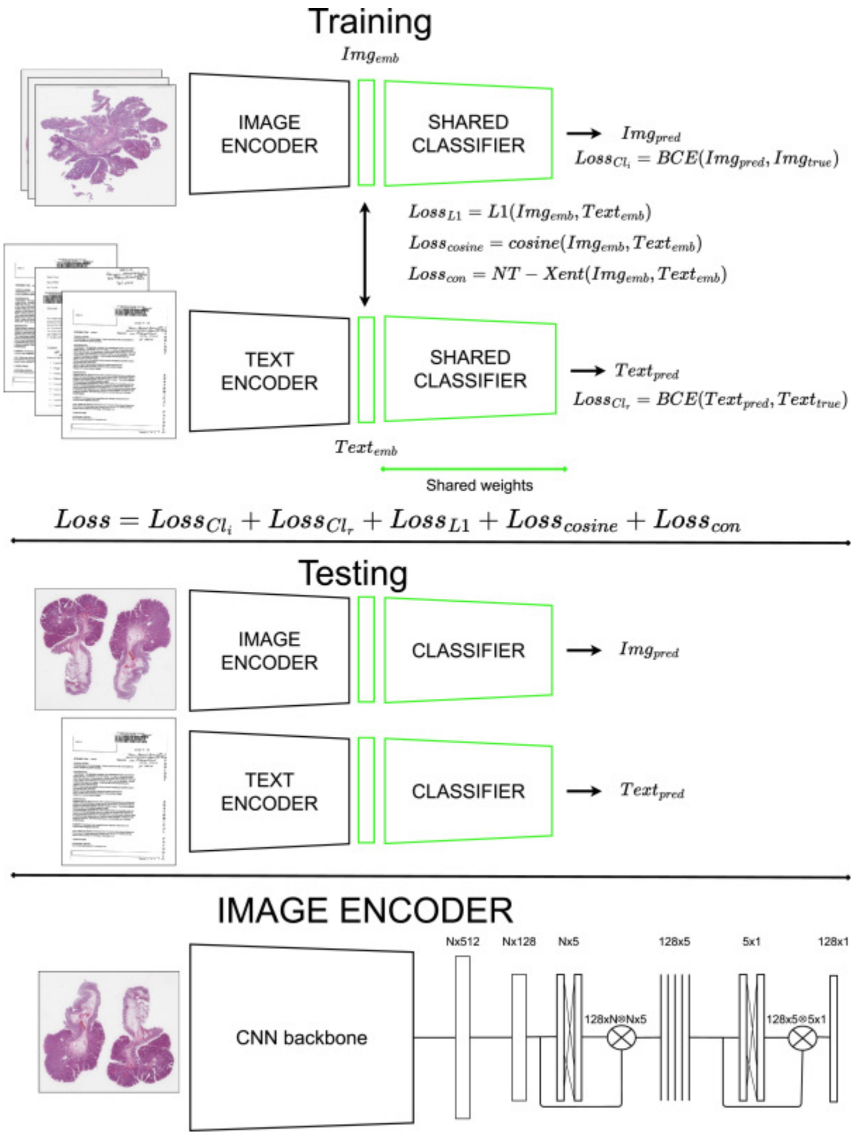
Training
Img_{emb}
IMAGE ENCODER
SHARED CLASSIFIER
Img_{pred}
$Loss_{Cl_i} = BCE(Img_{pred}, Img_{true})$
$Loss_{L1} = L1(Img_{emb}, Text_{emb})$
$Loss_{cosine} = cosine(Img_{emb}, Text_{emb})$
$Loss_{con} = NT - Xent(Img_{emb}, Text_{emb})$
TEXT ENCODER
SHARED CLASSIFIER
$Text_{pred}$
$Loss_{Cl_r} = BCE(Text_{pred}, Text_{true})$
$Text_{emb}$
Shared weights
$Loss = Loss_{Cl_i} + Loss_{Cl_r} + Loss_{L1} + Loss_{cosine} + Loss_{con}$
Testing
IMAGE ENCODER
CLASSIFIER
Img_{pred}
TEXT ENCODER
CLASSIFIER
$Text_{pred}$
IMAGE ENCODER
CNN backbone
Nx512
Nx128
Nx5
128x5
5x1
128x1
128xN⊗Nx5
128x5⊗5x1

Fig. 1: Overview of the multimodal architecture.

2.3 Loss Functions and Training Strategy

The network is trained using a composite loss function that combines:

The training of the network relies on a composite loss function that combines both supervised and self-supervised components. For the supervised part, a binary cross-entropy loss is employed, which is designed for multilabel classification. This loss function is used to predict the presence or absence of each pathology class—specifically Adenocarcinoma, High-Grade Dysplasia, Low-Grade Dysplasia, Hyperplastic Polyp, and Normal Glands. By penalizing incorrect predictions across these multiple classes, the supervised loss ensures that the model learns to accurately classify the pathology of each slide.

In addition to the supervised loss, the training process incorporates several self-supervised losses that help align the representations from the image and text modalities. One of these is the NT-Xent loss, also known as a contrastive loss. This loss encourages the embeddings derived from paired image-text samples (i.e., those that belong together) to be very similar, while simultaneously ensuring that embeddings from non-matching pairs remain distinct. A temperature parameter, set to 0.07, is used within this loss function to control the scaling of similarity scores, which in turn affects how strongly the loss penalizes dissimilarities.

Furthermore, an L1-loss is applied to minimize the absolute difference between the corresponding image and text embeddings. This loss function helps to fine-tune the alignment between the two modalities by reducing any numerical discrepancies between their representations. Complementing this, a cosine similarity loss is used to ensure that the embeddings not only have similar magnitudes but also point in the same direction in the latent space. By minimizing the angular difference between paired embeddings, this loss reinforces the idea that the image and text representations are semantically aligned.

The total loss, a weighted sum of these components, is optimized using gradient-based methods (e.g., Adam) over a 10-fold cross-validation framework.

The overall training objective is obtained by combining these losses into a weighted sum. This composite loss function drives the model to learn a robust, joint representation that integrates both the fine-grained visual details from the images and the high-level semantic information from the text. Despite the relatively limited number of training samples, the inclusion of these self-supervised components, alongside the supervised classification loss, enables the model to generalize well and perform effectively across various tasks.

3 Experimental Evaluation

3.1 Datasets and Labels

The evaluation of the proposed multimodal model is conducted using two main types of datasets: internal pathology workflow data and publicly available datasets. The internal dataset consists of whole slide images (WSIs) paired with pathology reports, collected from two hospitals, namely the Catania cohort and Radboudumc. Each WSI is linked to a corresponding pathology report, which provides textual descriptions of the observed tissue characteristics. The dataset is annotated using five pathology classes: Adenocarcinoma, High-Grade Dysplasia (HGD), Low-Grade Dysplasia (LGD), Hyperplastic Polyp, and Normal Glands. Importantly, these classes are not mutually exclusive, meaning that a single image can contain multiple pathological features, leading to a multilabel classification task. The dataset comprises approximately 6,176 samples in the training set and 868 samples in the testing set.

In addition to the internal dataset, two publicly available datasets are used to assess the generalizability of the model. The first dataset, UNITOPatho, contains histopathology images annotated with four classes relevant to colon pathology. The second dataset, IMP-CRC, is composed of colorectal images categorized into three different classes: High-Grade Lesions, Low-Grade Lesions, and Non-Neoplastic Lesions. These publicly available datasets provide an additional evaluation benchmark by introducing new data distributions, staining variations, and potentially different annotation guidelines, helping to assess how well the multimodal model can generalize beyond the training domain.

3.2 Downstream Tasks and Performance

The performance of the multimodal model is assessed across three key tasks: classification, multimodal retrieval, and concept linking for visual ontology creation. The first and primary task is the classification of WSIs into the predefined pathology classes. Results indicate that the multimodal model significantly outperforms unimodal models that rely solely on images. The weighted macro F1-score, a key metric used to evaluate classification performance while addressing class imbalances, is consistently higher for the multimodal model across all experiments. An important finding from this evaluation is that the multimodal model maintains strong classification accuracy even when trained with only half of the available data. This suggests that incorporating textual pathology reports into the training process enables the model to extract richer representations and improve performance despite data limitations, making it a particularly effective approach in scenarios where acquiring large labeled datasets is challenging.

Another key advantage of the multimodal architecture is its ability to support cross-modal retrieval. The shared latent space, which aligns image and text embeddings, allows for the retrieval of a relevant pathology report given an input WSI, or conversely, the retrieval of the most relevant WSIs given an input text query. This retrieval process is based on cosine similarity, which measures the alignment between embeddings in the shared representation space. The performance of the retrieval system is quantified using precision@k and mean average precision (mAP). Precision@k evaluates how many of the top-k retrieved results are relevant, while mAP provides a more holistic evaluation of retrieval effectiveness by measuring how well relevant items are ranked across the dataset. The results demonstrate that the multimodal model is capable of retrieving relevant reports and images with high precision, confirming the robustness of the learned representation and its potential application in medical information retrieval systems.

The third and final downstream task involves concept linking and the construction of visual ontologies. The model is capable of associating image patches with high-level textual concepts derived from the ExaMode ontology, an existing structured knowledge base of medical terms and conditions. This linking process allows the model to create a visual ontology in which distinct pathological structures within WSIs are mapped to corresponding diagnostic terms found in pathology reports. This feature is particularly promising for enhancing diagnostic interpretation, as it provides an interpretable and human-understandable mapping between textual descriptions and visual evidence. Moreover, this approach has significant potential for medical education, where students and trainees can benefit from an enriched learning experience by interacting with a dataset that directly links textual knowledge with visual examples.

Overall, the experimental results confirm the effectiveness of the proposed multimodal model in addressing classification, retrieval, and knowledge representation tasks. The ability to combine low-level image features with high-level semantic content from pathology reports not only improves classification accuracy but also enhances the interpretability and usability of the learned representations. This demonstrates the potential of multimodal deep learning to revolutionize the field of computational pathology, making diagnostic workflows more efficient and accessible.

4 Discussion and Critical Analysis

4.1 Methodological Strengths and Weaknesses

The proposed framework presents several significant strengths, making it a valuable contribution to the field of computational pathology. One of its most notable advantages is its data efficiency. Despite using only 6,000 paired whole slide images (WSIs) and pathology reports, the model achieves high classification performance, effectively addressing the issue of data scarcity that is prevalent in biomedical imaging. In many machine learning applications within the medical domain, obtaining large-scale annotated datasets is challenging due to the need for expert labeling. By leveraging both image and text modalities, this framework demonstrates that multimodal learning can enhance predictive performance even when the amount of available training data is limited.

Another strength of the model is its ability to effectively integrate multimodal information. Histopathology images provide detailed visual features related to tissue morphology, whereas pathology reports contain high-level semantic information describing diagnostic findings. The proposed approach successfully combines these two complementary sources of information, generating a robust joint representation that enhances downstream tasks such as classification, cross-modal retrieval, and concept linking. Unlike unimodal approaches that rely solely on either images or text, this model benefits from the interplay between the two modalities, leading to a richer and more informative feature space.

The framework also incorporates an integrated loss function that ensures proper alignment between image and text embeddings. The use of multiple self-supervised loss components, including the NT-Xent loss for contrastive learning, L1-loss for numerical similarity, and cosine similarity loss for directional alignment, plays a crucial role in creating a well-structured shared latent space. These self-supervised objectives, when combined with the supervised classification loss, enable the model to learn meaningful associations between visual features and textual descriptions, improving its ability to generalize to unseen data. This well-designed optimization strategy contributes to the model's effectiveness across multiple tasks beyond classification, including cross-modal retrieval, where the model successfully retrieves corresponding reports based on WSIs and vice versa. Additionally, the model's capability to establish links between visual features and high-level textual concepts facilitates the creation of visual ontologies that can aid diagnostic interpretation and medical education.

Despite these strengths, there are several limitations to the proposed approach. One of the primary concerns is the model's reliance on a complex set of hyperparameters. The composite loss function includes multiple weighting factors, such as those controlling the contributions of the NT-Xent loss, L1-loss, and cosine similarity loss, as well as a temperature parameter used in contrastive learning. These hyperparameters require extensive tuning to achieve optimal performance, making the model's training process more time-consuming and less easily reproducible. Since the optimal choice of these parameters may vary depending on the dataset, additional experimentation would be needed to ensure consistent performance across different medical imaging tasks.

Another limitation is the scalability of the approach. While the model has demonstrated strong results on the datasets used in this study, which include around 6,000 training samples from pathology workflows and additional publicly available datasets, it remains unclear how well the method would perform on significantly larger and more diverse datasets. Histopathology images can vary widely in terms of staining techniques, scanner specifications, and institutional differences in diagnostic protocols. If the model were to be applied to larger datasets containing more variations in tissue morphology, color distributions, and diagnostic styles, it is uncertain whether the current architecture would maintain its high performance. The extent to which the learned representations generalize to new clinical settings or different disease types is an important aspect that requires further investigation.

The design of the shared projection head, although effective for aligning image and text embeddings, may also represent a potential weakness. The model uses a single shared projection head for both modalities without modality-specific normalization or additional mechanisms to capture more complex relationships between images and text. While this design choice simplifies the training process, it may not fully exploit the rich multimodal interactions present in the data. More sophisticated fusion techniques, such as separate projection heads followed by an attention-based fusion mechanism, could potentially enhance the model's ability to capture intricate dependencies between textual descriptions and image features. Alternative methods, such as modality-specific embeddings combined with late fusion strategies, might allow for a more fine-grained representation that could improve overall performance in classification and retrieval tasks.

Additionally, while the model applies multiple text augmentation techniques to improve robustness, fully capturing the complexities and nuances of clinical language remains a challenge. Pathology reports are highly structured yet variable, as different pathologists may use different terminology, abbreviations, and phrasing conventions. The augmentation techniques used in this study, such as back-to-back translations, insert/rephrase strategies, and ChatGPT-based rewording, help introduce linguistic variability but may not fully address the issue of domain-specific jargon or ambiguous phrasing. Further refinement of these augmentation strategies, possibly through specialized medical language models trained on a larger corpus of clinical texts, could improve the robustness of the text branch.

4.2 Potential Improvements

- **Adaptive Loss Weighting:** Implement dynamic adjustment of loss weights (e.g., via uncertainty-based weighting) to simplify hyperparameter tuning.
- **Dual-Projection Architecture:** Experiment with separate projection heads for each modality followed by a fusion layer to capture more complex interactions.
- **Scaling and Domain Adaptation:** Extend the framework to larger and more diverse datasets and incorporate domain adaptation techniques to improve generalization.
- **Enhanced Text Augmentation:** Utilize state-of-the-art NLP models for paraphrasing and domain-specific augmentation to better capture clinical language variability.
- **Automated Hyperparameter Optimization:** Incorporate methods such as Bayesian optimization to streamline the tuning process.

5 Implementation

I tried to do some basic implementation of this multimodal architecture using pytorch. Its available in my github. Source code: https://github.com/prabal5ghosh/Multimodal-Representation-of-Whole-Slide-Images-and-Reports-implementation-using-pytorch

References

1. Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.
2. Niccolò Marini, Stefano Marchesin, Lluis Borras Ferris, Simon Püttmann, Marek Wodzinski, Riccardo Fratti, Damian Podareanu, Alessandro Caputo, Svetla Boytcheva, Simona Vatrano, et al. Automatic labels are as effective as manual labels in biomedical images classification with deep learning. *arXiv preprint arXiv:2406.14351*, 2024.
3. Niccolò Marini, Stefano Marchesin, Marek Wodzinski, Alessandro Caputo, Damian Podareanu, Bryan Cardenas Guevara, Svetla Boytcheva, Simona Vatrano, Filippo Fraggetta, Francesco Ciompi, et al. Multimodal representations of biomedical knowledge from limited training whole slide images and reports using deep learning. *Medical Image Analysis*, 97:103303, 2024.