TRANSFORMER-BASED MULTIMODAL FUSION FOR SURVIVAL PREDICTION BY INTEGRATING WHOLE SLIDE IMAGES, CLINICAL, AND GENOMIC DATA

Yihang Chen[⋆]*, Weiqin Zhao*[†]*, Lequan Yu*[†]

[⋆]Renmin University of China, Beijing, China
[†]The University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Survival prediction using whole slide images (WSIs) is a complex and difficult task, as handling gigapixel WSI directly is computationally impossible. In the past few years, people have worked out multiple instance learning (MIL) strategies to deal with WSIs, *i.e.*, splitting WSI into many patches (instances) and aggregating features across patches. Moreover, to better predict the survival outcome of patients, different modalities have been explored, among which gene features are used the most frequently. In this paper, we explore a graph-based strategy to handle WSIs and investigate a transformer-based strategy to combine different modalities for survival prediction. Moreover, clinical data was also adopted and different encoding manners of clinical information were explored. Experiments on two public datasets from The Cancer Genome Atlas (TCGA) demonstrate the effectiveness of the proposed graph-transformer framework for survival prediction.

Index Terms— Survival Prediction, Whole Slide Image, Multi-modality, Transformer, Graph Neural Network

1. INTRODUCTION

Although computer vision has been in rapid development in general with novel and revolutionary methods proposed one after another, how to deal with gigapixel whole slide image (WSI) remains challenging and complex. The difficulty comes from: (1) WSIs are too large so processing them directly is computationally infeasible and (2) WSIs usually only have slide-level labels (*e.g.*, a binary label or a cancer type label for one WSI) as the detailed pixel-level annotations are expensive and hard to obtain, which means tasks on WSIs are weakly supervised tasks in nature.

The past few years have witnessed large improvements in processing WSIs. People usually adopted the multiple instance learning (MIL) frameworks to deal with gigapixel WSI, which can be described as two stages: (1) splitting WSIs into bags of patches (instances) and extracting features on those patches, (2) aggregating features across instances to acquire global features for slide-level prediction [1, 2]. Although the coarse framework has handling WSIs is established, representation learning is still being explored. At the bag (WSI) level, conventionally popular methods are based on sets, which simply don't take the order and dependency within a bag (*i.e*, across instances) into consideration, and hence it induces permutation-invariant feature aggregation on instances [3, 4]. However, since set-based methods assume neither dependency nor ordering across instances, these methods lack the ability to utilize the interactions between instances (*e.g.*, interactions across cells in different patches), and the whole topology structure of WSI is ignored although this problem could be alleviated to some extend when combining set-based methods with attention mechanism.

Recently, graph-based methods have emerged, by which people not only focus on feature extraction and aggregation of patches but also explore the topology structure of the WSI as a whole [5]. To better make use of multi-scale and heterogeneous information of WSIs, some researchers propose network architecture to exploit hierarchical features as well as spatial structures for different resolutions [6, 7]. From the perspective of graph-based methods, we regard a WSI as a graph and patches as nodes of the graph, which has proven to surpass the state-of-the-art performance using set-based methods on some datasets. At the instance (patch) level, how to encode information about different patches has been a classic question. People have worked out lots of methods to encode information on patches, ranging from simple CNN to more complex networks [1, 8]. Apart from that, people consider using multi-modality methods to boost performance, among which many works have been done to fuse gene features into pathology (*i.e.*, WSI features) since tumors are often correlated with genes [9]. However, people seldom consider using other information like clinical information of patients. And how to effectively combine different modalities is still worth exploring.

In this paper, we focus on multi-modal WSI survival prediction and present a novel graph-transformer architecture to effectively learn the slide-level representation of WSI and integrate multi-modal information. To utilize the spatial information of different patches in WSI, we formulate WSI as a graph data structure and employ a graph convolutional layer to conduct local feature aggregation.

We then adopt a transformer architecture to effectively aggregate the WSI node features from a global perspective. More importantly, we encode the clinical data and genomic data information as extra feature tokens and adopt the transformer architecture to learn the relationship between them with WSI features, so that we can effectively integrate multimodal data for more accurate prediction. Experiments on two public TCGA datasets demonstrate the effectiveness of the proposed graph-transformer framework for survival prediction.

2. METHODS

2.1. Problem Formulation

We use Multiple Instance Learning (MIL) as our framework to handle WSIs for weakly supervised learning, which could be formulated as:

$$F(X_i) = h(g\{(f(x_{ij})\}) \quad x_{ij} \in X_i$$

where $X_i = \{x_{i1}, ..., x_{iN}\} \in R^{N \times d_1}$ represents the bag of instance features of the i-th sample, and x_{ij} stands for the j-th patch in the i-th sample. We define function $f : R^{d_1} \rightarrow R^{d_2}$ as the encoding function to map raw features into embeddings. The function $g : R^{N \times d_2} \rightarrow R^{d_3}$ is an aggregating function to aggregate features across patches to get our overall features. And $h : R^{d_3} \rightarrow R^{\#class}$ is a task-specific function, it could be a softmax function to output the probabilities of all classes for classification tasks.

For the survival prediction task, instead of estimating the survival time of patients, we want to get an ordinal risk value acquired through the survival function:

$$S(T \geq t, X_i) = \prod_{i=1}^{t} (1 - H(T = t | T \geq t, X_i))$$

To simplify the task, we divide survival time into intervals (represented as $i \in \{0, 1, 2, 3\}$ in the formula) according to quartiles so that we can use accumulative multiplication of H to obtain S instead of integration.

2.2. Graph-Based WSI Feature Aggregation

To better utilize the spatial information of WSIs, rather than treat different patches within a bag as independent from each other using set-based methods, we use graph-transformer architecture [5]. This model regards WSIs as graphs, splits WSIs into patches, and treats them as nodes of graphs. Based on the spatial structure of patches, nodes are connected as a graph, then a WSI is converted to a graph $G = (V, A)$, where $V = \{v_1, v_2, ..., v_N\}$ is the set of nodes denoting patches and containing node features $\vec{v_i} \in R^{d_2}$, and A is the adjacency matrix. $A_{ij} = 1$ if node v_i and v_j are adjacent to each other. To extract features from raw image data of patches as node features, we use pre-trained KimiaNet as our feature extractor [10]. Since every patch has 8 neighbors at most, the sum of each row or column of A is at least 1 and at most 8.

Then we pass the graph through a graph convolutional (GC) layer to get denser and more representative features [11]. The basic process during which the GC layer implements the message passing and aggregation could be formulated as:

$$H^{(l+1)} = ReLU(\widetilde{D}^{-\frac{1}{2}} \widetilde{A} \widetilde{D}^{-rac12} H^{(l)} W^{(l)})$$

$\widetilde{A} = A + I$ is the adjacency matrix with self-connections added. $\widetilde{D}_{ii} = \sum_j \widetilde{A}_{ij}$ and $W^{(l)}$ is a trainable matrix. $H^{(0)}$ is initialized as the feature matrix $X = [\vec{v_1}, \vec{v_2}, ..., \vec{v_N}]^T \in R^{N \times d_2}$. After a GC layer, we use mincut pooling to get our final graph tokens, the loss of which is unsupervised and differentiable and is derived from mincut optimization objective: partitioning node set V in K disjoint subsets by removing the minimum volume of edges [12].

2.3. Clinical and Genomic Data Encoding

Apart from pathological features, people often use gene features as another modality to boost performance [9, 13]. In this paper, we use gene data and clinical information of patients besides pathological features. Since clinical and gene data are both tabular data in nature, we adopt TabNet to encode them as dense embeddings. TabNet uses sequential attention to choose which features to reason from at each decision step with an encoder-decoder structure and it is proven empirically to surpass state-of-the-art tree-based methods when handling some tabular data [14]. To extract latent information contained in tabular data, both the encoder and decoder of TabNet could be used for iterations (3 iterations in our experiment for both). We then use the reconstructed features output by the decoder as our encoded embeddings. For the clinical information, besides encoding using TabNet, we also try encoding them directly (denoted as emb in section 3): for categorical attributes, we use categorical embeddings while we transfer numeric attributes to categorical ones according to their distributions.

2.4. Transformer-based Multimodal Fusion

When combining different modalities with pathological features, we simply treat gene embeddings or clinical embeddings extracted above as tokens and concatenate them with previous graph tokens as well as a CLS token. Figure 1 displays how we combine pathological features with gene data using TabNet. For gene data or clinical data extracted by TabNet, it will yield one reconstructed embedding to be treated as one token while for clinical data encoded directly, each attribute will become one embedding (*e.g.*, 9 attributes yield 9 tokens).

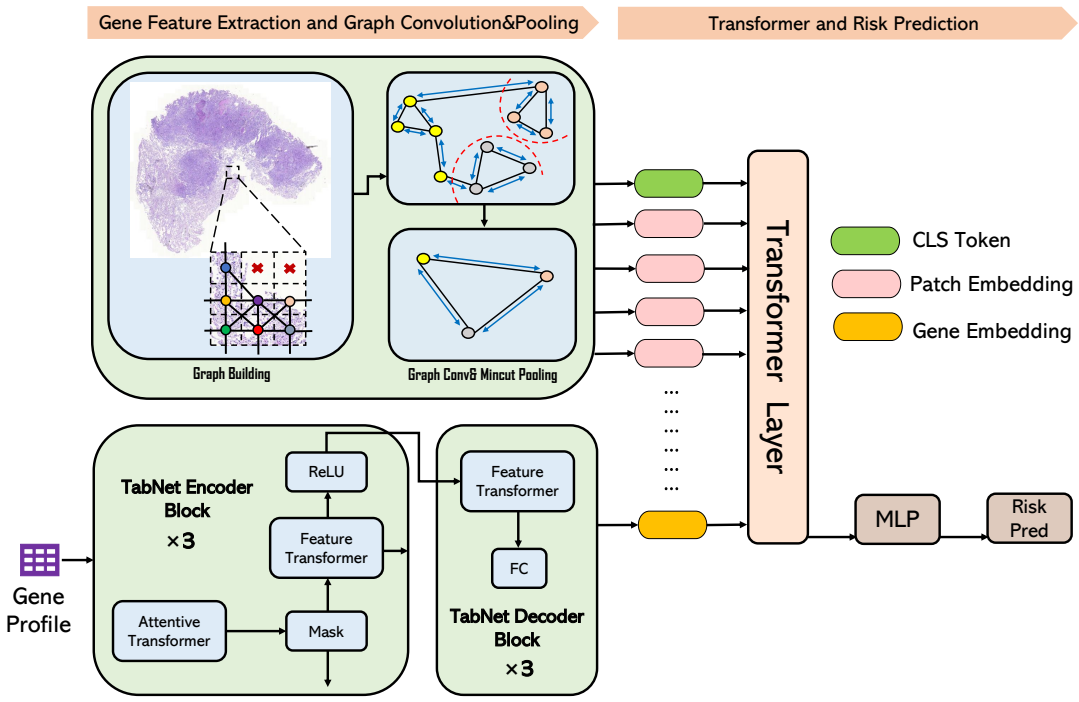

Fig. 1. Illustration of our transformer-based framework for the survival of the prediction by combining WSI with gene profile with TabNet. Pathological features are extracted by GraphTransformer and genomic features are encoded by TabNet and they are both treated as tokens to be combined and then fed into a Transformer layer.

Lastly, we pass all the tokens through a transformer layer [15, 16], where feature nodes are regarded as tokens in a sequence and the adjacency matrix is treated as positional information and self attention is used to capture importance across patches. Given that $x = \{e_1, ..., e_N\} \in R^{N \times D}$ is the sequence of feature nodes (tokens) extracted from the graph, qkv could be computed using the standard self-attention (SA) method (N is the number of tokens).

Based on the similarity between tokens, we can get attention weights $\{A_{ij}\}$. Multihead Self-Attention (MSA) is also utilized to combine information extracted by m number of heads. The whole basic framework of a transformer layer could be formulated as:

$$
\begin{aligned}
&q = xW_q, k = xW_k, v = xW_v \quad && W_q, W_k, W_v \in R^{D \times D_h} \\
&A = softmax(qk^T/\sqrt{d_h}), SA(x) = Av \quad && A \in R^{N \times N} \\
&MSA(x) = [SA_1(x)U_1; ...; SA_m(x)U_k] \quad && U_i \in R^{D_h \times D} \\
&t_0 = [x_{class}; h_1; ...; h_N], h_i = e_i E && E \in R^{D \times D'} \\
&t_l{}' = MSA((LN(t_{l-1})) + t_{l-1} && l = 1...L \\
&t_l = MLP(LN(t_l{}')) + t_l{}' && l = 1...L
\end{aligned}
$$

where LN denotes Layer Normalization, MLP denotes a multi-layer perceptron and L is the number of MSA blocks. In the end, we get our output:

$$H = Sigmoid(LN(t_L{}^{(0)}))$$

which serves as the hazards of patients. The whole framework is trained in an end-to-end manner.

3. EXPERIMENTS

3.1. Datasets

For this work, we use Lung Adenocarcinoma (LUAD) and Stomach Adenocarcinoma (STAD) datasets from The Cancer Genome Atlas (TCGA), a public data consortium including matched WSIs, and clinical and gene data. Clinical information contains a patient's age, race, gender, etc. And for gene data, we selected the most strongly expressed 250 genes. Before analyzing these datasets, we filter out little WSIs, which contain less than 100 nodes, to ensure data quality. LUAD contains 317 patient samples with 822 WSIs (approx 35 GB) and an average bag size of 1546 patches (nodes). STAD contains 372 patient samples with 397 WSIs (approx 39 GB) and an average bag size of 3002 patches. For each dataset, we divide patients into four intervals according to the quartiles of their survival time.

Table 1. Experiment Results (c-index ± a standard error of a mean)

Modalities and Models	LUAD	STAD
WSI(ABMIL)	0.5377 ± 0.0153	0.5208 ± 0.0229
WSI(GraphTransformer)	0.5576 ± 0.0556	0.5468 ± 0.0422
WSI(GraphTransformer) + Clinical Info (TabNet)	0.5367 ± 0.0272	0.5566 ± 0.0506
WSI(GraphTransformer) + Clinical Info (Emb)	**0.5962 ± 0.0194**	0.5527 ± 0.0418
WSI(GraphTransformer) + Gene (TabNet)	0.5605 ± 0.0630	0.5432 ± 0.0222
WSI(GraphTransformer) + Gene (TabNet) + Clinical Info(Emb)	0.5804 ± 0.0358	**0.5581 ± 0.0427**

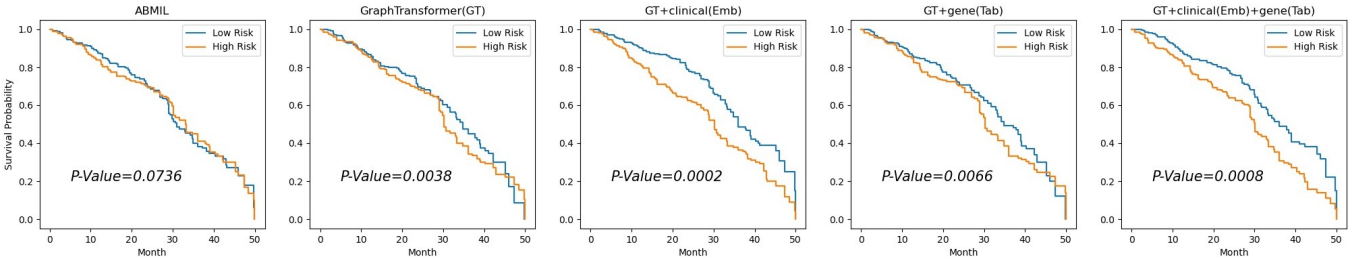

Fig. 2. Kaplan-Meier Analysis based on risk stratification on LUAD dataset. The survival curves are plotted using risk scores given by models w.r.t. ground-truth survival time (in months). The Logrank test is performed to compare the two survival distributions statistically.

3.2. Evaluation Metrics

On each dataset, we choose five-fold cross-validation to evaluate our models and concordance index (c-index)

$$c = \frac{\#concordant\,pairs}{\#concordant\,pairs + \#discordant\,pairs}$$

is used as our standard to measure the performance of risk prediction.

3.3. Experiment Results

The experimental results are shown in Table 1. Firstly, we compare our graph-based method with the state-of-the-art set-based method Attention-based Deep Multiple Instance Learning (ABMIL) [3]. An increase of 3.7% and 5.0% is achieved respectively on LUAD and STAD datasets. We also compare different modalities based on our graph-based models. Our experiments show that when other modalities are added, performance could improve most of the time, and a maximum increase of 6.9% and 2.1% is achieved respectively. Moreover, we evaluate the model's ability in risk stratification, which is widely used for survival analysis models [7, 9]. We stratify patients into two groups according to the median of the predicted risk scores. From Figure 2, we can see that graph-based methods could separate patients into low and high risk groups better than the set-based method ABMIL. Judging from the p-values of Logrank test, graph-based methods are much better than ABMIL, and adding different modalities could bring higher performance with smaller p-values too.

4. CONCLUSION

In this paper, we compare the graph-based method graph-transformer with the traditional state-of-the-art set-based method ABMIL for WSI analysis. We also propose a new transformer-based framework to effectively combine different modality data for survival prediction. Extensive experiments carried out prove that graph transformer has better performance than ABMIL. Empirically, with other modalities added with pathological features, modals could improve their performance in general. However, it's not always promising to combine other modalities and even doing so would hurt the performance, which indicates that how to extract features from genes and clinical information and how to combine different modalities remain open and challenging. Future work would focus on how to extract features of different modalities better, as well as investigating fusion methods of different modalities.

5. ACKNOWLEDGEMENTS AND ETHICAL COMPLIANCE

The work described in this paper is supported in part by the Natural Science Foundation of China under grant 62201483 and in part by HKU Seed Fund for Basic Research (Project No. 202009185079 and 202111159073). Ethical approval was not required as confirmed by the license attached with the open access data.

6. REFERENCES

[1] Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," 2015.

[2] Yu Zhao, Fan Yang, Yuqi Fang, Hailing Liu, Niyun Zhou, Jun Zhang, Jiarui Sun, Sen Yang, Bjoern Menze, Xinjuan Fan, et al., "Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4837–4846.

[3] Maximilian Ilse, Jakub Tomczak, and Max Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, PMLR.

[4] Shivam Kalra, Mohammed Adnan, Graham Taylor, and Hamid R Tizhoosh, "Learning permutation invariant representations using memory networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 677–693.

[5] Yi Zheng, Rushin H. Gindra, Emily J. Green, Eric J. Burks, Margrit Betke, Jennifer E. Beane, and Vijaya B. Kolachalama, "A graph-transformer for whole slide image classification," 2022.

[6] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang, "H^2-mil: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 933–941, Jun. 2022.

[7] Pei Liu, Bo Fu, Feng Ye, Rui Yang, Bin Xu, and Luping Ji, "Dsca: A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis," 2022.

[8] Zhen Chen, Jun Zhang, Shuanlong Che, Junzhou Huang, Xiao Han, and Yixuan Yuan, "Diagnose like a pathologist: Weakly-supervised pathologist-tree network for slide-level immunohistochemical scoring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 47–54.

[9] Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew F.K. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood, "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 4015–4025.

[10] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, Maral Rasoolijaberi, Milad Sikaroudi, Mohd Adnan, Sultaan Shah, Charles Choi, Savvas Damaskinos, Clinton JV Campbell, Phedias Diamandis, Liron Pantanowitz, Hany Kashani, Ali Ghodsi, and H. R. Tizhoosh, "Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides," 2021.

[11] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," 2016.

[12] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi, "Mincut pooling in graph neural networks," 2020.

[13] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood, "Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, 2020.

[14] Sercan Ömer Arik and Tomas Pfister, "Tabnet: Attentive interpretable tabular learning," *CoRR*, vol. abs/1908.07442, 2019.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.