Multimodal Co-Attention based MIL on Gigapixel Whole Slide Images

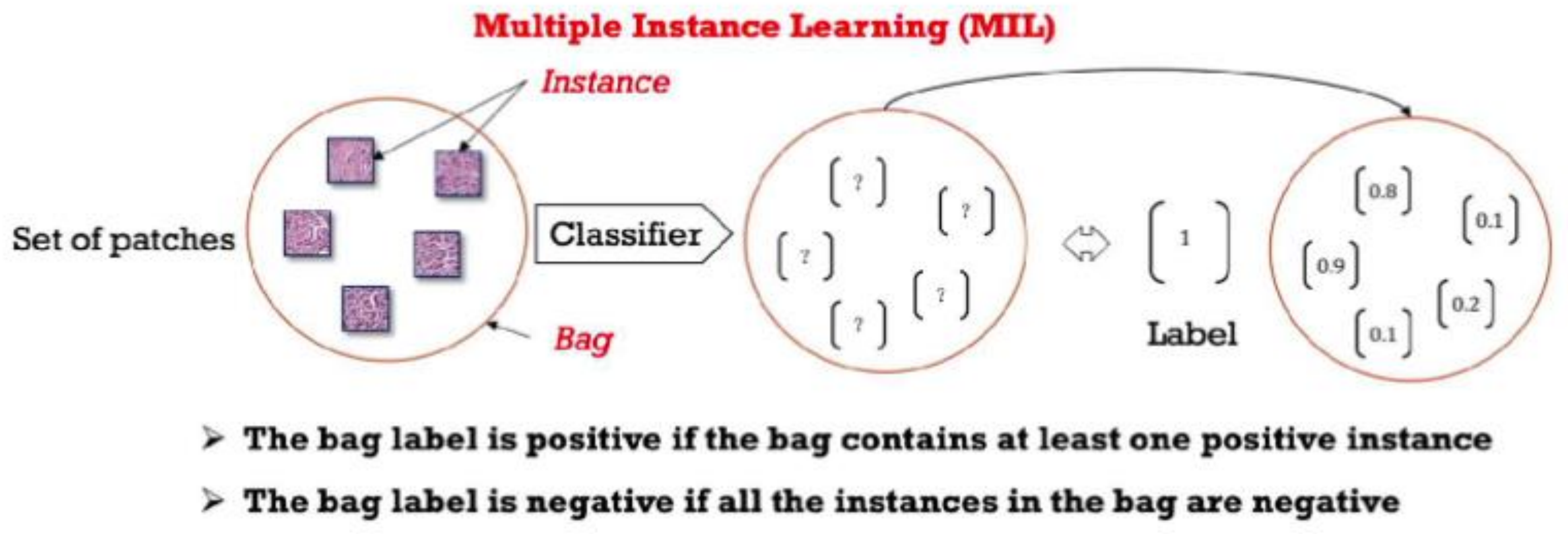
R.X. Xu[1,*], Y.C. Guo[1], S.H. Tu[1], X.Y. Wang[1] , J. Wang[1], Y.P. Zhao[1]
Supervisor: Dr. Jefferson Fong
[1] Computer Science and Technology Programme, Faculty of Science and Technology, BNU-HKBU United International College
**Corresponding Student Author Tel: +86-13691888248, E-mail: r130026173@mail.uic.edu.cn*

Abstract

Survival outcome prediction in computational pathology is a complex weakly-supervised and ordinal regression task, requiring the modeling of intricate interactions within the tumor microenvironment in gigapixel whole slide images (WSIs). Although recent advancements have treated WSIs as bags in multiple instance learning (MIL), representing entire WSIs remains a significant challenge due to 1) the computational burden of feature aggregation in large bags, and 2) the difficulty of integrating biological priors like genomic data. In response, we propose a multimodal co-attention based framework, which learns an interpretable, dense co-attention mapping between WSIs and genomic features in an embedding space. Drawing inspiration from Visual Question Answering (VQA), where word embeddings attend to salient objects in an image, histology patches are enabled to attend to relevant genes in predicting patient survival. Beyond visualizing these multimodal interactions, our approach also reduces the computational complexity of WSI bags, making Transformer layers a viable encoder backbone in MIL.

Introduction

The application of deep learning in the field of computational pathology has revolutionized the analysis and diagnosis of medical pathological images. By learning large amounts of pathological image data, these deep learning models can automatically identify disease types in pathological images, thus assisting physicians in making more accurate diagnoses. In our task, we proposed a multimodal deep learning approach to effectively process the gigapixel whole slide image, which is a kind of pathological image.

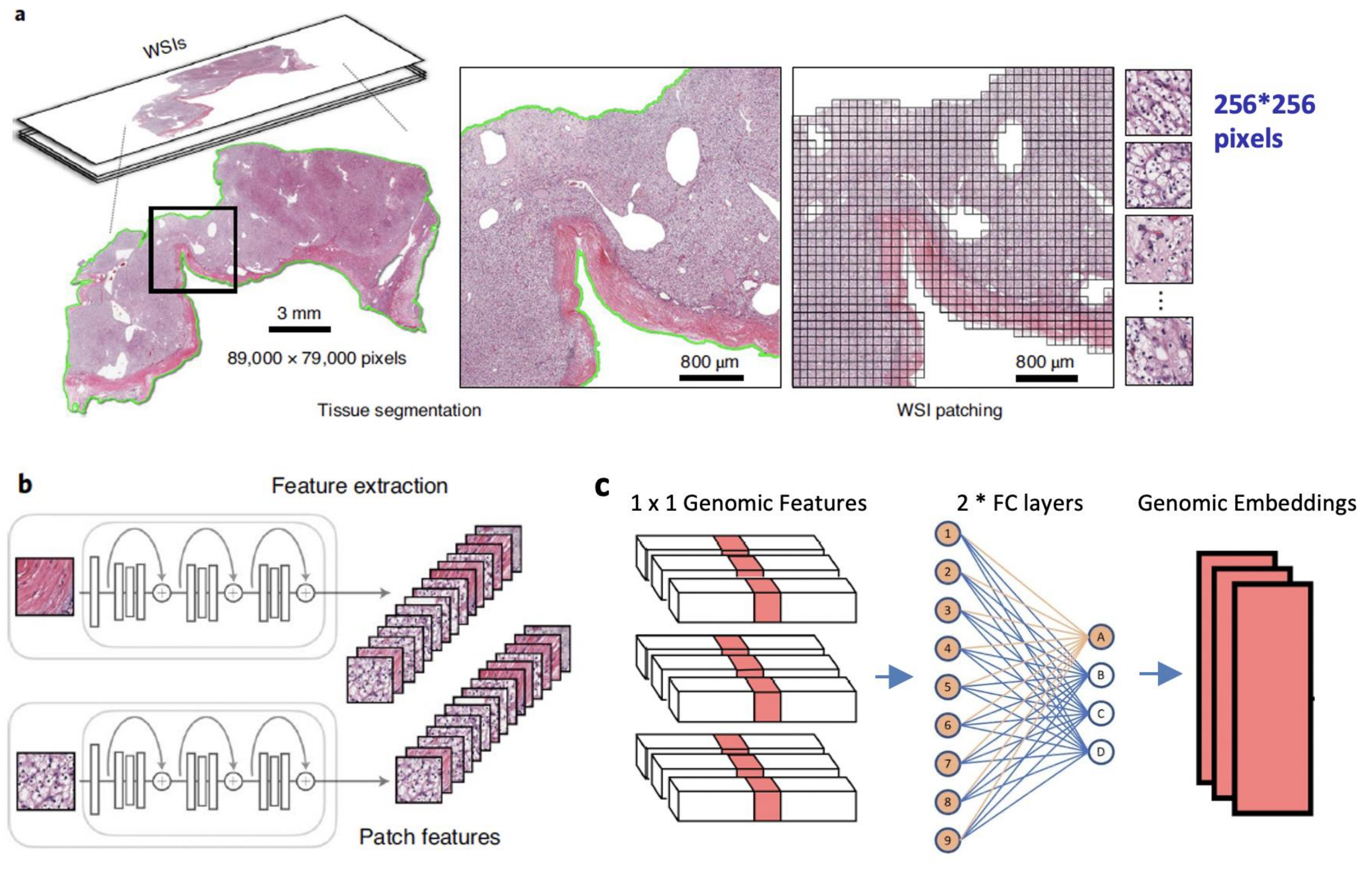

Multiple Instance Learning (MIL)
Instance
Set of patches
Classifier
Bag
?
1
Label
0.8
0.1
0.9
0.2
0.1
➢ The bag label is positive if the bag contains at least one positive instance
➢ The bag label is negative if all the instances in the bag are negative

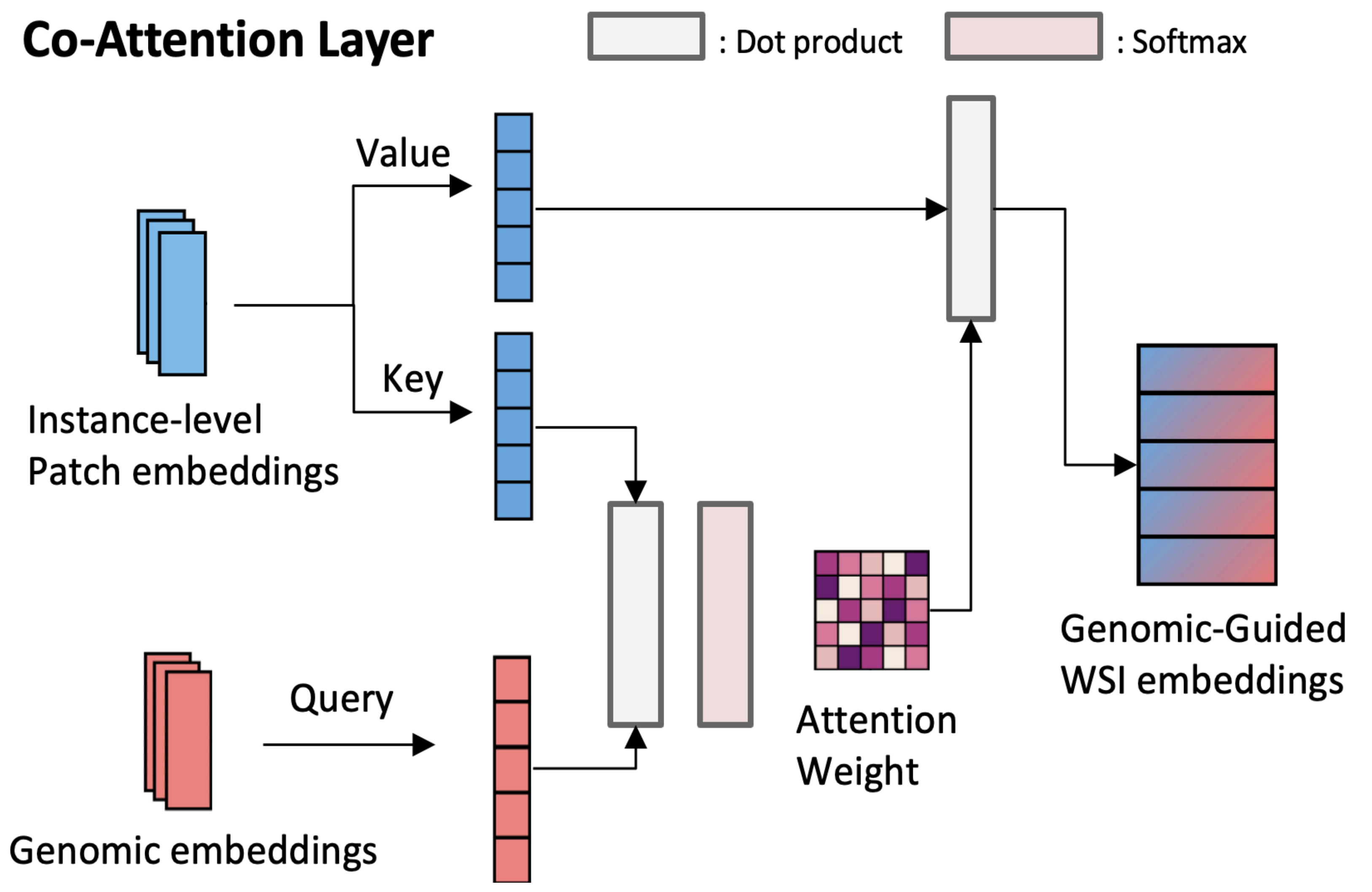
Multi-instance learning is a machine learning paradigm characterized by the fact that the training data is organized into instance bags, each of which contains several instances, at least one of which is positive. This learning method is particularly useful when only a small portion of a set of instances is positive, while the others are negative. In pathology, multi-instance learning has important applications when processing whole slide images.

In addition to directly processing the WSIs, we introduce another input, the genes, which make our model a multimodal one. These genetic data contain RNA and CNV sequences, which can be interpreted as textual information. We use Visual Question Answering (VQA) to better understand. A VQA system takes as input an image and a free-form, open-ended, natural-language question about the image and produces a natural-language answer as the output.

Methodology

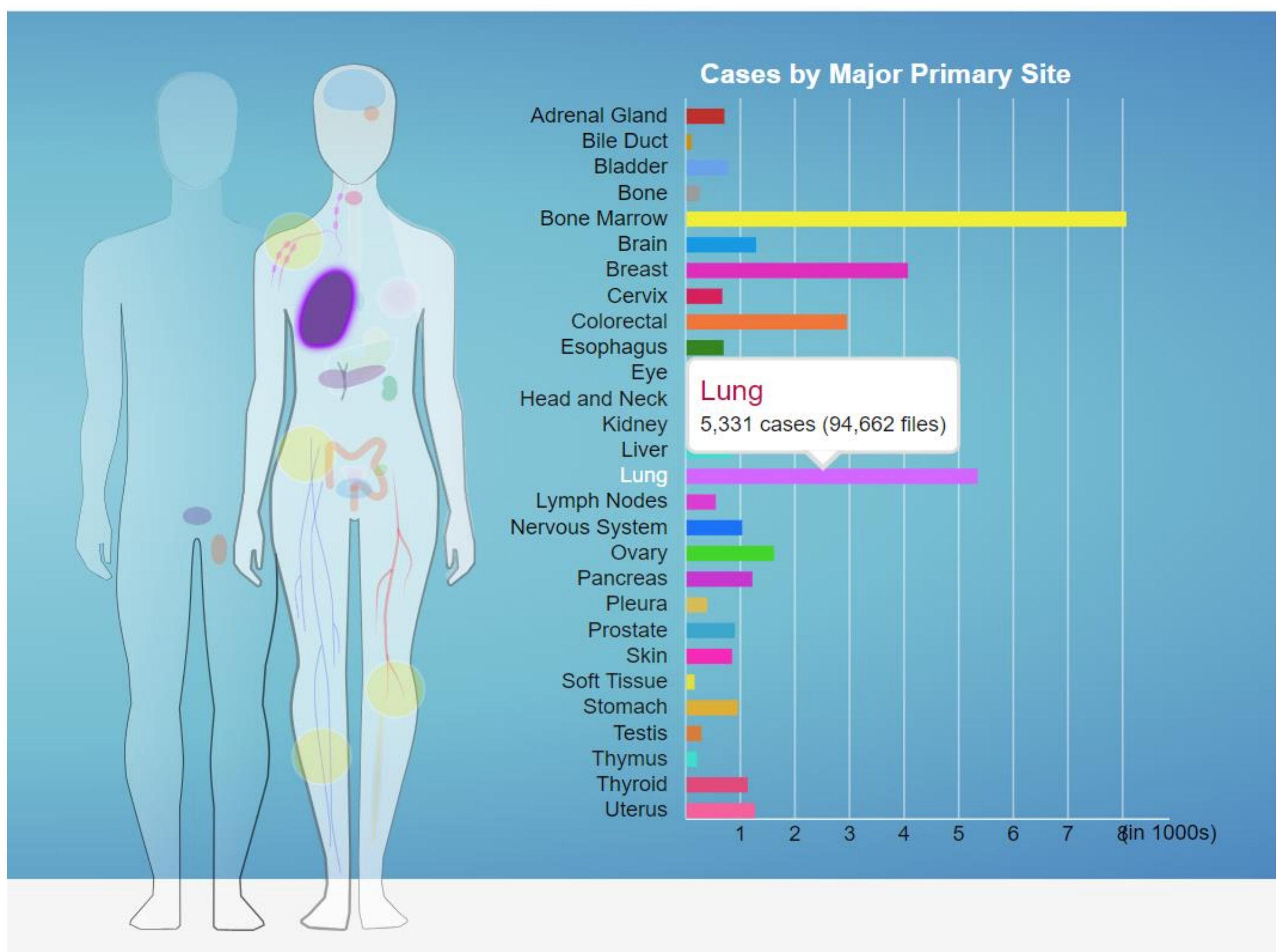
➢ **Model Work Flow:**

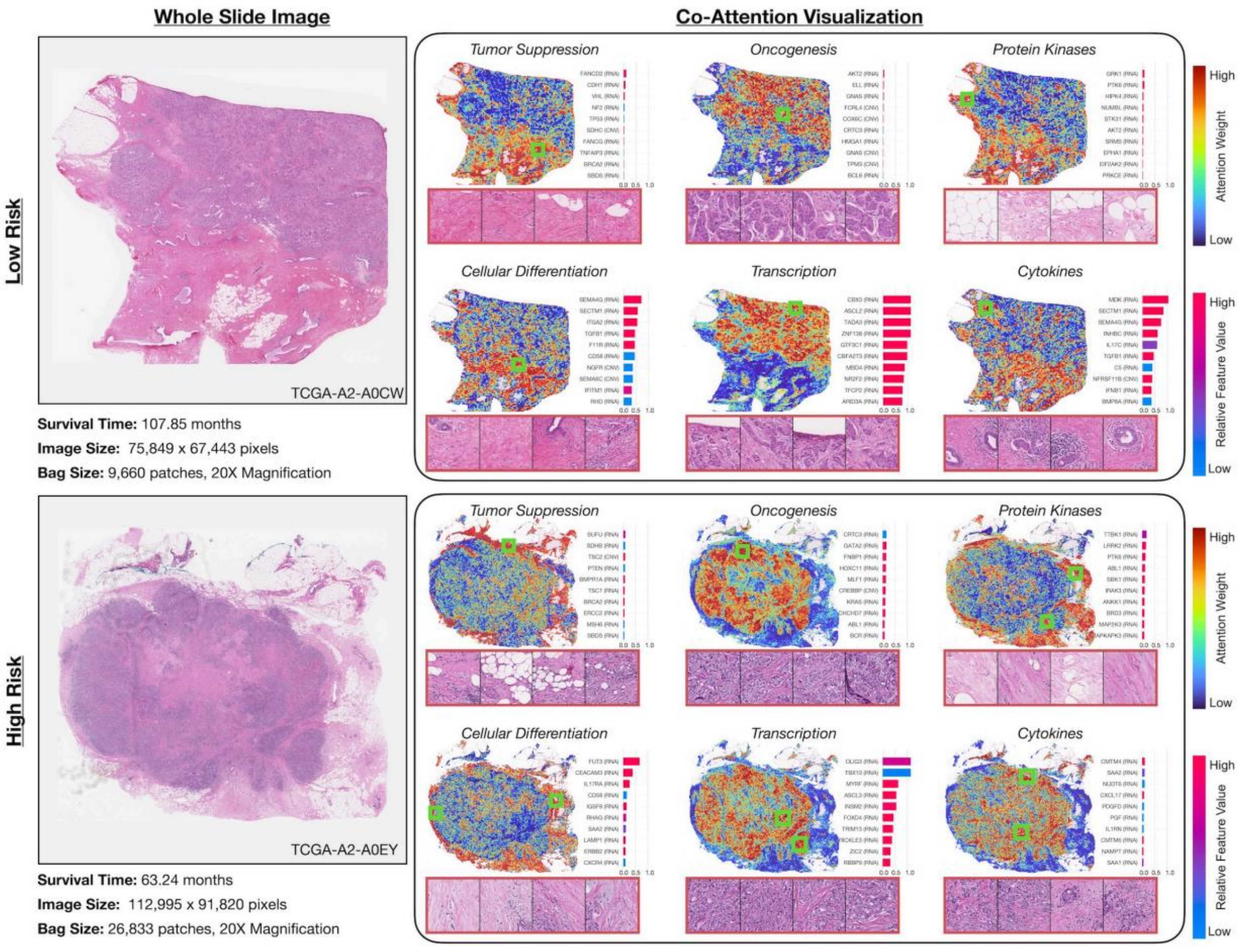




Experiment

To validate our proposed method, we used the Lung cancer datasets from The Cancer Genome Atlas (TCGA), a public cancer data consortium that contains matched diagnostic WSIs and genomic data with labeled survival times and censorship statuses. After the experiment in 5-fold cross-validation in the TCGA Lung Cancer dataset, we approach around 0.97 on LUAD and around 0.96 on LUSC.





Survival Time: 107.85 months
Image Size: 75,849 x 67,443 pixels
Bag Size: 9,660 patches, 20X Magnification

Survival Time: 63.24 months
Image Size: 112,995 x 91,820 pixels
Bag Size: 26,833 patches, 20X Magnification

Conclusion

In this work, we proposed Multimodal Co-Attention Transformer for survival outcome prediction in pathology. Our method formulates both gigapixel WSIs and genomic features as permutation-invariant sets, from which we develop more sophisticated feature aggregation strategies in MIL via transformer attention. A limitation in our current study is that we used a previously-curated gene set with potentially overlapping biological functional impact. Future work would focus on investigating early fusion of WSIs with more fine-grained, distinct biological gene sets, and further quantification of phenotype-genotype correspondences.