MCAT-HViT: Multimodal Co-Attention Hierarchical Visual Transformer for Survival Prediction

Saporita Alessia, Berselli Elena, Jelali Abeer

March 5, 2024

Abstract

We develop a multimodal co-attention hierarchical transformer able to predict whether ovarian cancer patients will survive more than 42 months. The threshold has been set empirically to 42 months based on our data distribution. However it can be tailored according to one's specific requirements. The survival is predicted via: 1) formulating both WSIs and genomic inputs as embedding-like structures, 2) using co-attention mechanism that learns pairwise interactions between instance-level histology patches and genomic embeddings, 3) fusing modalities by processing the embeddings with a transformer encoder that applies multi-head self-attention to all input tokens, thus allowing cross-modality information transfer and 4) reducing the noise contained in histology patches by randomly set to zero the visual embedding with probability chosen to be 30

1 Introduction

Survival prediction is a challenging weakly-supervised task in computational pathology that involves modeling complex interactions within the tumor microenvironment in gigapixel whole slide images (WSIs). In pathology, WSIs exhibit enormous heterogeneity and can be as large as 150,000 pixels. Depending on the problem, labels for slide-level classification may be: 1) localized in a small pixel region that occupies a tiny proportion of the total image or 2) spanning the entire composition of a WSI and dependent on the interactions of its components.

Due to the enormous gigapixel resolutions of WSIs, many approaches adopt a two-stage multiple instance learning-based (MIL) approach for tractable representation learning of WSIs, in which: 1) instance-level feature representations are extracted from randomly sampled image patches in the WSI, and then 2) global aggregation schemes are applied to the bag of instances to obtain a WSI-level representation for subsequent supervision.

Though often approached as a weakly-supervised task using only gigapixel WSIs, survival outcome prediction is traditionally framed as a multimodal learning task in which genomic information can be used as an additional modality for supervision or integration. In the current state-of-the-art, survival prediction faces an additional challenge due to the large data heterogeneity gap between WSIs and genomics: WSIs represented as bags containing tens of thousands of image patches as instances, while genomic features are often represented as tabular attributes. As a result, many approaches use late fusion mechanisms for feature integration, which prevents learning important multimodal interactions.

To address these challenges, we propose a weakly-supervised, multimodal learning framework called Multimodal Co-Attention Hierarchical Visual Transformer (MCAT-HViT) that learns a dense co-attention mapping between WSIs and genomics for survival outcome prediction by identifying informative instances using genomic features as queries formulated in an embedding space. Our co-attention transformation also reduces the space complexity of WSI bags, which enables the hierarchically combination of transformer layers to attend to the cross-modal interactions. The proposed method is evaluated on ovary cancer dataset.

2 Data Preparation

2.1 Data Collection

We collect data about ovary cancer by downloading WSIs and molecular data from NIH Genomic Data Commons Data Portal with the GDC Data Transfer Tool.

2.1.1 WSIs

We collect slide images acquired as tissue slide. CLAM is used to process WSIs: after segmenting the images and excluding holes, 256 x 256 patches without spatial overlapping are extracted just from the relevant portion of the slides and given as input to a pretrained truncated ResNet50 to encode raw image patches into 1024-dim feature vectors. The total amount of WSIs is 387 with an average number of patches equal to 9815. To correlate a patient and his corresponding image, we create a table as shown in Table 1

CASE ID	SLIDE ID	FILENAME
3b81ec7c-0934-4649-9231-9919dd26dd15	TCGA-24-1549	TCGA-24-1549-01A..af3133cc.svs

Table 1: WSI table structure.

2.1.2 Gene Expression Values

We select gene expression values belonging to the transcriptome profiling category acquired by the RNA-Seq technique. We filter out the protein coding genes according to HUGO Gene Nomenclature Committeey and then we normalize the values by applying a $\log_2$ transformation and a Standard Scaler. To correlate a patient and his corresponding mRNA values, we create a table as shown in Table 2 with an overall number of genes equal to 18941.

GENE ID	GENE NAME	CASE ID	...	CASE ID
ENSG00000000003.15	TSPAN6	34.669.924.782.174.000	...	4.225.675.965.560.760

Table 2: mRNA table structure.

2.1.3 Methylation Values

We select methylation values extracted from illumina human methylation 27, a microarray platform designed for assessing DNA methylation levels at specific sites in the human genome. We filter out the protein coding genes according to HUGO Gene Nomenclature Committeey and then we normalize with a Standard Scaler. To correlate a patient and the gene corresponding to the methylation value, we create a table as shown in Table 3 with an overall number of genes equal to 14055.

DNA LOCATION	GENE NAME	CASE ID	...	CASE ID
cg00005847	HOXD3	0.278284904613036	...	0.612975805275754

Table 3: Methylation table structure.

2.1.4 Labels

We collect our labels from clinical data. We select three fields: vital status (Dead/Alive), days to last follow-up and days to death in months. We encode Dead status as 0 and Alive status as 1, for Alive patients days to death value is set to None. To correlate a patient and his clincial data, we create a table as shown in Table 4

FIELD	CASE ID	CASE ID	...	CASE ID
vital status	Alive	Alive	...	Dead
days to death			...	46.52
days to last followup	6.0	11.26	...	46.52

Table 4: Label table structure.

2.2 Dataset

The dataset used in our experiments is obtained from the union of the previously described tables and has the following structure 5.

CASE ID	SLIDE ID	SURV MONTHS	STATUS	LABEL	RNA SEQ	...	MET
TCGA-09-1673	TCGA-09-..83.csv	2.97	1	1	4.712029	...	-3.477141

Table 5: Dataset structure.

The field SURVIVAL MONTHS refers to days to last follow-up for Alive patients and to days to death for Dead ones. The LABEL field values is set according to a threshold (in months): to 0 if a patient died before it or 1 if a patient survived for at least those months. Our dataset distribution of survival months for Alive and Dead patients is shown in Figure 2 and Figure 3. In our experiment, the threshold has value 42 months: in the case of ovary cancer, it is a reliable period not to consider the pathology critical. After this filtering, the overall number of patients is 258: 117 with label equal to 1 and 141 with label equal to 0, as shown in Figure 1. To be able to use or not methylation values during training, we create two datasets: one only with RNA values (18917 columns), one with both RNA and methylation values (28095 columns).

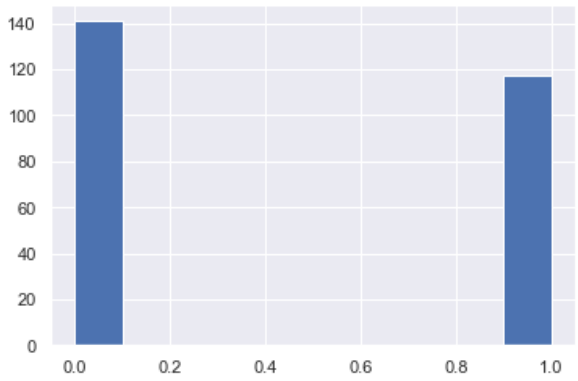


Figure 1: Label Distribution

3 Model

To evaluate different approaches in the resolution of the previously described problem, we modified the below described state-of-the-art methods for our purpose. The goal is to construct a multimodal architecture that performs binary classification, predicting whether the patient survives more than 42 months or not.

3.1 SNN

SNN is self-normalizing neural network which enable high-level abstract representations. While batch normalization requires explicit normalization, neuron activations of SNN automatically converge towards zero mean and unit variance. This convergence property of SNN allows to: 1) train deep networks with many layers, 2) employ strong regularization schemes, and 3) make learning highly

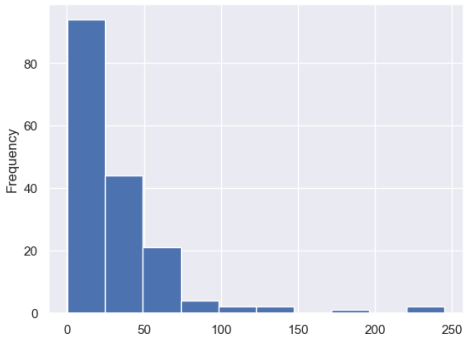


Figure 2: Survival Months Distribution for Alive Patients

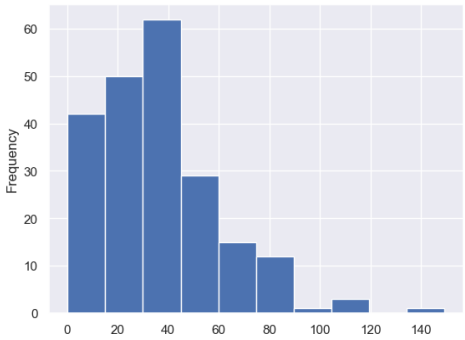


Figure 3: Survival Months Distribution for Dead Patients

robust. It is a unimodal baseline for only genomic features which has been used previously for survival outcome prediction in the TCGA.

3.2 MCAT

Multimodal Co-Attention Transformer (MCAT), for weakly-supervised and multimodal learning using WSIs and genomics for survival outcome prediction. Multiple Instance Learning (MIL) is a weakly-supervised learning task that operates on set-based data structures also known as "bags", in which each bag is an unordered (permutation-invariant) set of instances that can be of varying size with incomplete instance-level labels. For single-label classification, given a bag containing d-dimensional instances with label Y, the goal is to learn a permutation-invariant function F that predicts the bag label without detailed knowledge of the instances. MCAT develops a set-based neural network architecture F that integrates histological and genomic data to estimate the hazard function, which measure the probability of the patient surviving after time point t. Instance-level feature representations are extracted from small image patches in the WSI using all available tissue information across multiple WSIs for large-scale training. For all WSIs, the tissue-containing image regions are patched into a set of non-overlapping 256x256 patches and are fed as input into a ResNet-50 CNN + FC layer (pretrained on ImageNet) that extracts k-dim feature embeddings. For M total histology patches across all WSIs, the extracted patch embeddings are packed into a bag. To obtain more expressive, embedding-like feature representations for genomic data, genes are categorized into N different sets with similar biological functional impact. In MCAT implementation, N = 6 functional categories are used to define the following genomic embeddings: 1) Tumor Supression, 2) Oncogenesis, 3) Protein Kinases, 4) Cellular Differentiation, 5) Transcription, and 6) Cytokines and Growth. Each genomic category is fed to a FC layer to obtain genomic embeddings, which are packed into a bag data structure. Due to the data heterogeneity gap between gigapixel WSI and genomic features, the Genomic-Guided Co-Attention (GCA) layer is used to learn a dense co-attention mapping between bag representations of WSIs and genomics that directly models pair-wise interactions between the feature embeddings of the two modalities. An important detail of GCA is that the bag size of Q is much smaller than K, V . As a result, the query bag aggregates the bag containing M instance-level patch embeddings into a bag containing N WSI-level embeddings which makes the cost of applying subsequent self-attention layers quadratic with respect to N instead of M. Eventually, two MIL Transformers are constructed to aggregate feature embeddings in the resulting bag. In this process, self-attention is used to model complex, long-range feature interactions between genomic-guided visual concepts that would otherwise be intractable using the original WSI bag with large M. Then, the global attention pooling adaptively computes a weighted sum of all embeddings within each respective set to finally construct bag-level features. As a final step, those bag-level features are concatenated and fed as input to several FC layers to obtain the final predicton.

3.3 MeTra

MeTra is a multimodal Medical Transformer that can learn from imaging data, non-imaging data, or a combination of both, to perform survival prediction. In MeTra, chest radiographs are first processed by a vision backbone (ViT) to extract high-level image features. Additionally, clinical parameters are projected into the latent representation using a linear layer to match the dimensionality of the image tokens. To fuse imaging and non-imaging data efficiently, the latent representations of both backbones are concatenated to form the latent representation. The self-attention mechanism used inside transformers to process the input sequence does not consider the order of the elements in the sequence. To address this issue, a learnable position encoding token is added to the embeddings of both modalities. The resulting multimodal representation is processed with a transformer encoder, where the multi-head self-attention layers allow cross-modality information transfers. A multi-layer perceptron is applied to the output to form the final prediction that quantifies the likelihood of survival of the patient.

3.4 MCAT-ViT

The first attempt is inspired both by MeTra and MCAT, as shown in Figure 4 We borrow the initial structure of MCAT to learn feature embeddings of both WSI and genomic data and to reduce the dimensionality of histological data representations by exploiting the GCA layer. Since the dimensionality of the two modalities match, we implement a Vision Transformer to allow cross-modality information transfer as in MeTra. The last step of the architecture is a MLP performing the survival prediction.

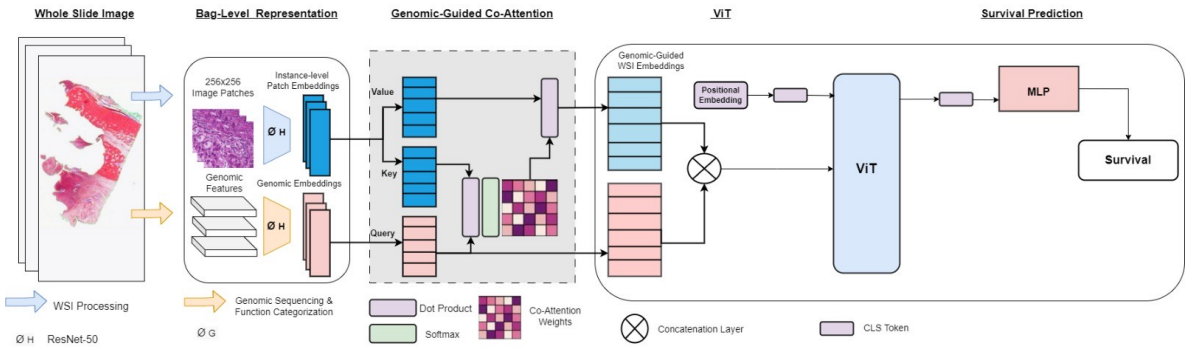


Figure 4: MCAT-ViT Architecture

3.5 MCAT-HViT

The second architecture we implement is a Hierarchical Vision Transformer, whose backbone is based on MCAT and MeTra, as shown in Figure 5. Transformer layers can be combined hierarchically to attend to the cross-modal interactions. A common practice is that multimodal inputs are encoded by independent Transformer streams and their outputs are concatenated and fused by another Transformer. This kind of hierarchical attention is an example of late interaction/fusion. Starting from the MCAT structure, instead of simply applying a MLP, to better model the correlation between the two modalities, we perform the following steps: 1) concatenate the feature embeddings, 2) add a learnable positional encoding token and 3) feed the resulting representation into a ViT. With respect to the previous architecture, after the GCA layer we apply a unimodal transformer encoder to learn intra-modality connections, while the last multimodal ViT is used to better learn inter-modality relations.

4 Training

During training, we used Adam optimization with a learning rate of 2 x 10^{-4}, and weight decay of 1 x 10^{-5}. Due to samples having varying bag sizes, we use a batch size of 1, with 32 gradient accumulation steps and 20 epochs for training. As in MeTra, we exploit visual dropout to reduce the noise contained

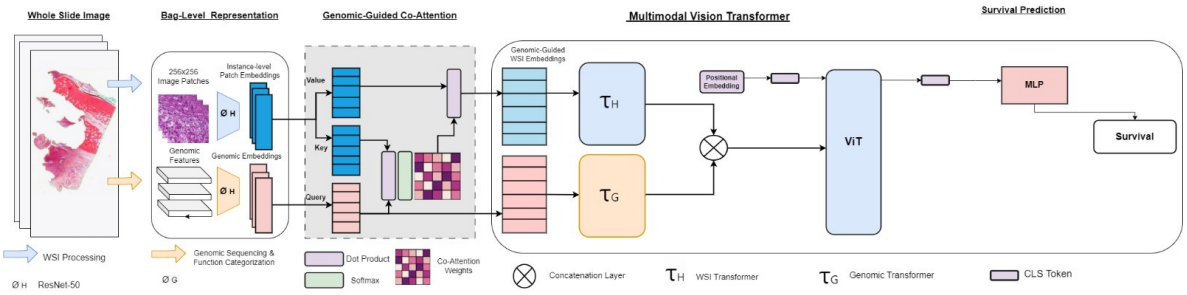
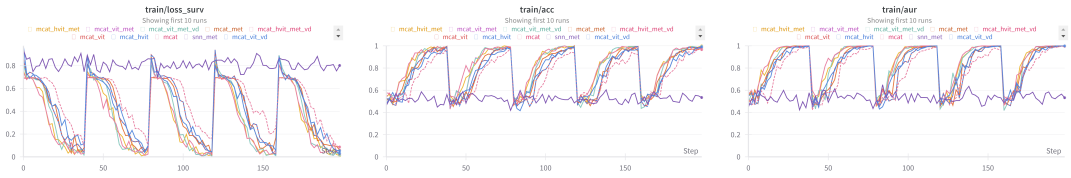
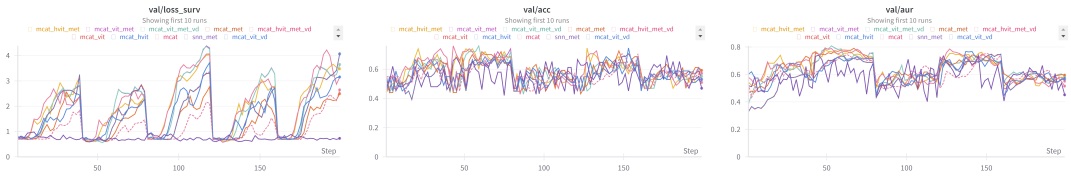


Figure 5: MCAT-HViT Architecture

in WSI images. We use Cross-Entropy Loss as a loss function. We train our proposed methods in a 5-fold cross-validation, applying the widely used splitting of 80-20. To evaluate our models, we use the metrics of Accuracy and AUROC since the dataset is not completely balances: without the AUROC, Accuracy would not be a suitable metric to measure the performance of the models. In Figure 7 we show the progress in training and validation by comparing all the tested architectures, while in Figure 9 are reported the comparisons just between SNN, MCAT-ViT and MCAT-HViT since those are the architecture achieving the best results.

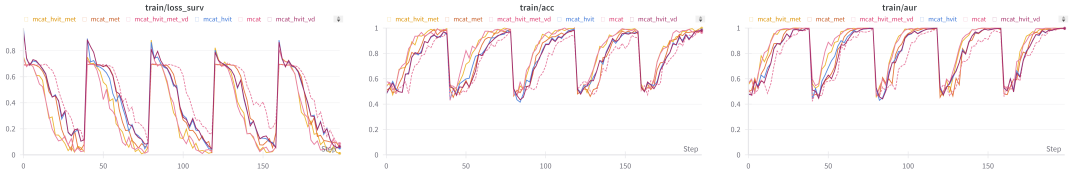


(a) Training loss (b) Training Accuracy (c) Training AUROC



(a) Validation loss (b) Validation Accuracy 5 (c) Validation AUROC

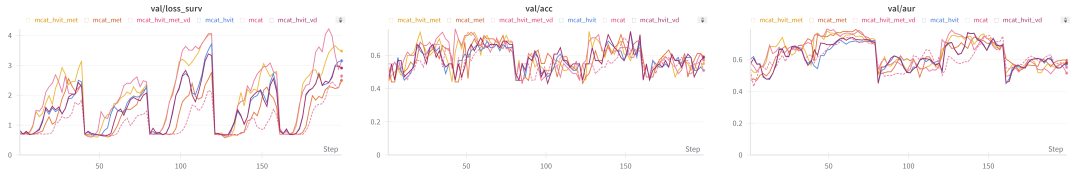
Figure 7: All architectures



(a) Training loss (b) Training Accuracy (c) Training AUROC

5 Results, Comparison and Conclusions

Using the same 5-fold cross-validation splits we evaluate SNN, MCAT, MCAT-ViT, and MCAT-HViT on the task of survival prediction. For all methods, we use the same instance-level feature extraction pipeline for bag construction of WSIs, as well as identical training hyperparameters and loss function for supervision. Table 6 shows the results of all methods, considering alternatively visual dropout



(a) Validation loss (b) Validation Accuracy (c) Validation AUROC

Figure 9: MCAT-ViT, MCAT-HViT and SNN

equal to 0 or 0.3 and testing the performances with and without methylation data. Also, it shows the number of learnable parameters for each model and each configuration. In Figure 10 we show the confidence intervals for each model and configuration. Regarding the performance, the SNN performs slightly better than our proposed MCAT-ViT and MCAT-HViT. The reason is that the SNN implements a precise regularization technique, which is designed to improve performance, while MCAT architectures do not exploit such method. It is important to highlight this difference because, despite SNN only processes genomic data and does not employ WSIs, its performance is comparable with the MCAT-ViT and MCAT-HViT ones. The visual dropout is introduced as an attempt to reduce the WSIs'noise in both the MCAT-ViT and MCAT-HViT architectures, but its application does not significantly affect the performance. We explain this phenomenon with the capability of the networks to learn during training which are the most relevant and less noisy features to focus on. Another point is that our MCAT-ViT and MCAT-HViT performance improves when also methylation data are employed in the prediction, despite the confidence interval is slightly widened. Otherwise, without the inclusion of methylation data, the state-of-the-art MCAT performs better. However, those intervals are significantly smaller in the case of SNN and MCAT, meaning that the uncertainty in prediction for MCAT-ViT and MCAT-HViT is higher. Regarding the number of parameters, the SNN has more parameters than the MCAT networks. This is because SNN does not use a categorization into groups for the genomic features, while MCAT divides those features into groups and applies a smaller SNN for each group: in the end, the summation of the parameters of each SNN in the MCAT is less than the number of parameters for the entire SNN. Another observation is that we replace the MLP of MCAT with a ViT: this substitution causes the increment of the number of parameters for those architectures. In conclusion, the proposed architecture of MCAT-ViT and MCAT-HViT are comparable with the state-of-the-art networks in the field of survival prediction using multimodal learning. In particular, although there are no substantial differences, the changes applied let the proposed methods have better performance with respect to MCAT. However, MCAT-ViT and MCAT-HViT are still prone to refinement since their confidence intervals have greater fluctuation with respect to SNN and MCAT.

Model	Met	Params	Visual Dropout	Accuracy	AUROC
SNN	true	7,389,442		65.03	70.25
	false	6,039,874		64.70	71.30
MCAT	true	4,823,556		63.03	69.85
	false	4,455,428		63.92	69.32
MCAT-ViT	true	6,003,202	-	63.79	70.17
	true		0.3	63.79	70.14
	false	5,437,954	-	62.34	66.05
	false		0.3	60.78	66.00
MCAT-HViT	true	8,111,618	-	62.52	70.02
	true		0.3	62.53	70.83
	false	7,546,370	-	62.34	68.19
	false		0.3	62.35	68.57

Table 6: Results.

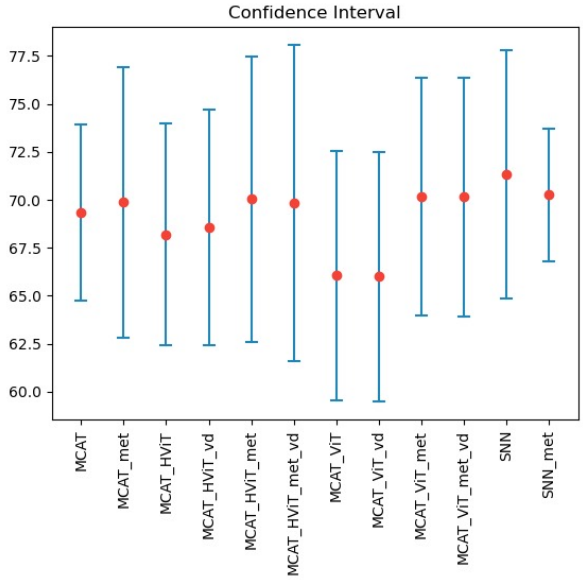


Figure 10: Confidence Intervals.