A COMPUTATIONAL COMPLEXITY ANALYSIS FOR MULTIMODAL INTEGRATION SCHEMES

In this section, we present the step-by-step details of the computational complexity analysis presented in Section 3.3. The analysis is done with respect to the size of the input modalities associated with the three paradigms used in our experimental setting: early fusion followed by self-attention, cross-modal attention, and One-Versus-Others (OvO) Attention.

A.1 EARLY FUSION

The early fusion approach involves first combining the modalities and then processing the concatenated sequence with the self-attention mechanism.

Step 1: Concatenation of Modalities.

Let k be the number of modalities and n be the feature-length of each modality.

$$\text{Total length after concatenation} = k \times n$$

The complexity for this operation is linear:

$$\mathcal{O}(k \cdot n)$$

Step 2: Compute Queries, Keys, and Values.

The self-attention mechanism derives queries (Q), keys (K), and values (V) for the concatenated sequence (length $k \cdot n$) using linear transformations with representation dimension, d. The complexity of each transformation operation is:

$$\mathcal{O}(k \cdot n \cdot d)$$

Step 3: Compute Attention Scores.

Attention scores are computed by taking the dot product of queries and keys. The self-attention mechanism has quadratic complexity with respect to the sequence length and linear complexity with respect to the representation dimension d (43). Thus, given the concatenated sequence's length of $k \cdot n$ and the dimension of the keys and queries d, the complexity of this step is:

$$\mathcal{O}((k \cdot n)^2 \cdot d) = \mathcal{O}(k^2 \cdot n^2 \cdot d)$$

Step 4: Calculate the Weighted Sum for Outputs.

For each of the $k \cdot n$ positions in the concatenated sequence, we compute the softmax of the attention scores to produce the attention weights. These weights are then multiplied with their corresponding d-dimensional values to compute the weighted sum, which becomes the output. The computational complexity of these operations is:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

When combining all steps, the dominating terms in the computational complexity stem from the attention scores' computation and the weighted sum, culminating in an overall complexity of:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

A.2 CROSS-MODAL ATTENTION

For cross-modal attention, each modality attends to every other modality.

Step 1: Compute Queries, Keys, and Values for Inter-Modal Attention.

From a given modality, compute a query (Q), and from the remaining $k - 1$ modalities, compute keys (K) and values (V). Keys, queries, and values are obtained using linear transformations with representation dimension d. The complexity of each transformation operation is:

$$\mathcal{O}(n \cdot d) \text{ for each query, key, value set}$$

Considering all modalities:

$$\mathcal{O}(k \cdot (k-1) \cdot n \cdot d)$$

The term $k \cdot (k-1)$ comes from the number of pairwise permutations of k, given by ${}_kP_2 = \frac{k!}{(k-2)!} = k(k-1)$.

Step 2: Calculate Attention Scores for Inter-Modal Attention.
The queries and keys from different modalities are used to compute attention scores, which represent how much one modality should attend to another.

$$\mathcal{O}(n^2 \cdot d) \text{ for each pair of modalities (43)}$$

Considering all modalities:

$$\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d)$$

Step 3: Calculate the Weighted Sum for Outputs.
For every modality interaction, calculate the softmax of the attention scores to obtain the attention weights. These weights are then used in conjunction with the values vector to derive the weighted sum for the output:

$$\mathcal{O}(n^2 \cdot d) \text{ for each pair of modalities}$$

Considering all modalities:

$$\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d)$$

When evaluating all steps together, the dominating factors in computational complexity arise from the computation of attention scores and the weighted sum. Thus, the collective complexity for cross-modal attention, where each modality attends to every other, equates to:

$$\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d) = \mathcal{O}((k^2 - k) \cdot n^2 \cdot d)$$

For the complexity of cross-modal attention, the dominant term is k^2. The $k-1$ term effectively becomes a constant factor in relation to k^2. As k tends toward larger values, the difference between k^2 and $k^2 - k$ diminishes. This is a consequence of the principles of big O notation, which focuses on the fastest-growing term in the equation while dismissing constant factors and lower-order terms. As a result, for asymptotic analysis, the complexity

$$\mathcal{O}(k^2 - k) \cdot n^2 \cdot d$$

can be simplified to:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

.

A.3 One-Versus-Others (OvO) Attention Complexity

Step 1: Averaging of "Other" Modalities.
Let k be the number of modalities and n be the feature-length of each modality. For each modality m_i, averaging over the other $k-1$ modalities results in a complexity of:

$$\mathcal{O}(n)$$

Given that this needs to be computed for all k modalities:

$$\mathcal{O}(k \cdot n)$$

Step 2: Calculate Attention Scores with Shared Weight Matrix W.
The modality vector m_i and the average of "other" modalities, $\frac{\sum_{j \neq i}^{n} m_j}{n-1}$, are used to compute attention scores, which represent how much one modality should attend to the others. Multiplication with the weight matrix W (with representation dimension d) and the dot product with the summed modalities lead to:

$$\mathcal{O}(n^2 \cdot d)$$

Considering this operation for all k modalities:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

Step 3: Calculate the Weighted Sum for Outputs.
For every modality interaction, calculate the softmax of the attention scores to obtain the attention weights. These weights are then used in conjunction with the m_i vector (analogous the values (V) vector) to derive the weighted sum for the output:

$$\mathcal{O}(n^2 \cdot d) \text{ for each pair of modalities}$$

Considering all modalities:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

When evaluating all steps together, the dominating factors in computational complexity arise from the computation of attention scores. Thus, the collective complexity for cross-modal attention, where each modality attends to every other, equates to:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

In summary, One-Versus-Others (OvO) Attention exhibits a computational complexity that grows linearly with respect to the number of modalities ($\mathcal{O}(k \cdot n^2 \cdot d)$). In contrast, both early fusion through self-attention and cross-attention approaches demonstrate quadratic growth with respect to the number of modalities ($\mathcal{O}(k^2 \cdot n^2 \cdot d)$). This makes OvO a more scalable option for multimodal integration.

B TCGA Modality Descriptions and Detailed Pre-processing

CNV defines the varying number of repeats of genetic fragments found in a human genome. The number of repeats of specific genetic fragments influences gene expression levels and has been associated with the progression of different cancers (32). Any genomic regions missing CNV values or only having one unique value across all cases were removed. DNA methylation represents the amount of condensation of genetic regions due to the chemical alteration imposed by methyl groups. This condensation generally represses gene activity near the genetic region. Any genomic regions with missing values were removed. Clinical data includes information such as the patient's diagnosis, demographics, laboratory tests, and family relationships. Categorical features were isolated and a coefficient of variation test was run to determine highly variable features. Features with a coefficient of variation higher than 70 were kept for analysis, along with the target variable. These features were converted into numerical format using one-hot-encoding. Gene expression data is collected through RNA-sequencing. Levels of gene expression are recorded by detecting the amounts of transcripts found for each gene. These levels can be used to determine the molecular mechanisms underlying cancer. Transcriptomic data was filtered to only include protein-coding genes and measured in fragments per kilobase of exon per million mapped fragments (FPKM). Imaging - TCGA collects pathology slide images of tissues sampled from the tumor. This modality provides visual information about the malignant region and can help with diagnosis and treatment planning. The image data was filtered only to include DX images, which result from a single X-Ray exposure, rotated to landscape view, then cropped to the median aspect ratio of 1.3565. We filtered for patients that had all five modalities, and we also only chose the patients that were still alive, to create a more balanced number of patients between cancer types (338 colon cancer patients, 329 kidney, 301 lung, 228 liver, and 226 stomach patients, after the filtering). The task we created is to classify each patient's cancer type. For all modalities, features with missing values were dropped. For CNV, DNA Methylation, and gene expression data, feature reduction was performed using a random forest classifier, only on training data, ensuring the test was not seen by the random forest. Using the validation set, we determined the best number of estimators (out of 50, 100, 150).

C Hyperparameter Tuning

For each experiment, we used the validation accuracy to determine the best hyperparameters. We tuned the learning rate (0.01 - $1 * 10^{-8}$), batch size (16, 32, 64, 128), epochs (200 epochs with early

stopping if validation accuracy did not increase for 5 epochs), and number of attention heads for the OvO and pairwise cross-modal attention models (1, 2, 4, 8, 16). For the neural network encoders, we tuned the number of linear layers ranging from 1 to 4. Similarly, for the convolutional neural network, we tuned the number of convolution layers ranging from 1 to 4.

D Compute Resources

For each experiment, we use one NVIDIA GeForce RTX 3090 GPU. For the Hateful Memes task, single-modality models ran for roughly 40 minutes, and multi-modal models ran for roughly 55 minutes on average. For the Amazon reviews task, the single modality pre-trained models ran for roughly 50 minutes, the single modality neural network ran for a minute, and the multi-modal models ran for approximately an hour on average. For the TCGA task, single-modality models ran for 5 minutes, while multi-modal models ran for roughly 15 minutes on average. In the simulation dataset, the maximum modalities was 20 which took our model, OvO, roughly 2 minutes to run, while the cross-modal attention baseline took about 20 minutes to run on average.

E Significance Testing

We use a t-test to determine if there is a significant difference in accuracy and F1-score means between OvO attention and the next best-performing multimodal model. Our sample size is 10 from each group, as we initialized the models with 10 random seeds. For the Hateful Memes dataset, we compare against cross-attention as it performed the second best after OvO. Using an $\alpha = 0.01$, we have evidence to reject the null hypothesis and conclude that there is a statistically significant difference in means between cross-attention and OvO attention. The p-value for the accuracy scores is $1.22e^{-8}$ and the p-value for F1-scores is $1.89e^{-5}$. For the Amazon reviews dataset, we compare against self-attention as it performed the second best after OvO. We get a p-value for accuracy scores of $4.77e^{-4}$ and a p-value of $4.87e^{-4}$ for F1-scores. Thus, we demonstrate a statistically significant difference in accuracy and F1-score means between self-attention and OvO attention. Lastly, for the TCGA dataset, we do not have evidence to reject the null hypothesis and cannot say that the accuracy and F1-score means were different between OvO and cross-attention since the p-values were greater than $\alpha = 0.01$ (p-value of 0.04 for accuracy means, and p-value of 0.02 for F1-score means). This demonstrates that although cross-attention performed slightly better than OvO, it was not statistically significant.

F Case study for attention vectors on Amazon reviews dataset

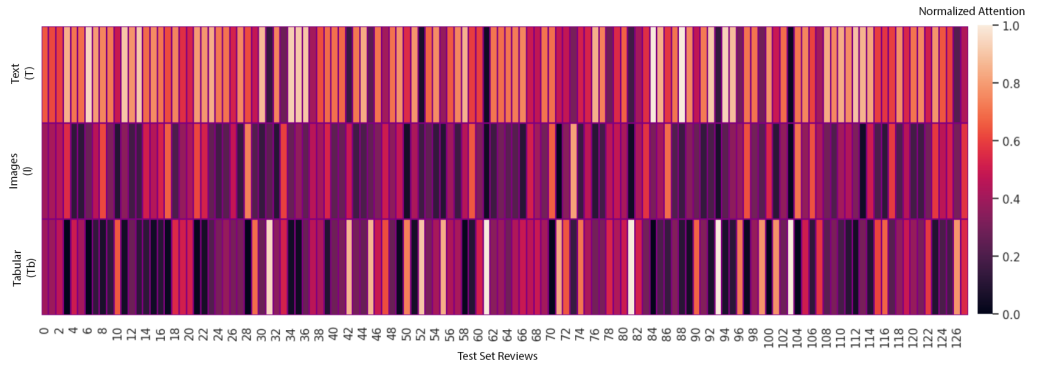

Figure 5: **Attention heatmap for Amazon reviews dataset**. Each attention vector is computed by averaging across the embedding dimension and across the 10 random seeds used to report our best model. The horizontal axis includes a sample of size 128 (batch size of the model) reviews from the test set, and the vertical axis includes the "main" modality from the attention score. The attention scores Text (T) vs. others, Images (I) vs. others, and Tabular (Tb) vs. others. This figure is consistent with the single-modality results from table 3.

Since our model, OvO, performed well on the Amazon reviews task and could be used for future sentiment analysis tasks reliably, we wanted to explore the attention scores on this task. Each attention context vector shown in Figure 5 is computed by averaging across the embedding dimension and across the 10 random seeds used to report our best model. The X-axis includes a sample of size 128 (batch size of the model) reviews from the test set and the y-axis includes the "main" modality from the attention score. The attention weights are Text (T) vs. others, Images (I) vs. others, and Tabular (Tb) vs. others. We observe that the text attention vector is the most highly scored, which is supported by the single-modality results from Table 3, where text was the highest performing single modality. Thus, we further demonstrate that OvO can be used to better understand modality importance, without the need for ablation studies. This is significant because knowing which modality is most important to decision-making can motivate future data collection efforts in diverse research environments and help make deep learning models more transparent.