One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data

Michal Golovanevsky[†], Eva Schiller[‡], Akira Nair[§], and Eric Han*
Department of Computer Science, Brown University,
Providence, RI 02912, USA
[†]*E-mail: michal_golovanevsky@brown.edu*
[‡]*E-mail: eva_schiller@brown.edu*
[§]*E-mail: akira_anir@brown.edu*
**E-mail: eric_j_han@brown.edu*
www.brown.edu

Ritambhara Singh
Department of Computer Science,
Center for Computational Molecular Biology,
Brown University,
Providence, RI 02912, USA
E-mail: ritambhara@brown.edu
www.brown.edu

Carsten Eickhoff
School of Medicine,
Institute for Bioinformatics and Medical Informatics,
University of Tübingen,
Tübingen, 72074, Germany
E-mail: c.eickhoff@acm.org
www.uni-tuebingen.de

Appendix A.

1. 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.

1.1. *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.[1] 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:

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

1.2. *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.

$$\mathcal{O}(n^2 \cdot d) \text{ for each pair of modalities}^1$$

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:

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

.

1.3. *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.

2. Simulation Dataset Details

We consider two classes: (1) 20 random feature values that sum up to 1.0, and, (2) 20 random feature values that are each less than 0.15. The threshold was chosen at 0.15 because if 0.10 was the threshold, the mean of the 20 values would be 0.05, and thus, the sum would also be very close to 1, on average. This would render the task too difficult, and there would not be a significant difference between the samples across the two labels. Setting the threshold to 0.2 would render the task too easy, as on average, the numbers are consistently greater in the second class and the classes could be differentiated using only one modality. Thus, we chose 0.15 as the threshold.

3. Compute Resources

For each experiment, we use one NVIDIA GeForce RTX 3090 GPU. For the MIMIC task, single-modality models ran for roughly 40 minutes, and multi-modal models ran for roughly 55 minutes on average. For the eICU, 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 TADPOLE 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.

Table A1. Average runtimes for different tasks and model types using one NVIDIA GeForce RTX 3090 GPU.

Task	**Models**	**Runtime (minutes)**
MIMIC	Unimodal	40
	Multimodal	55
eICU	Unimodal pre-trained	50
	Unimodal neural net	1
	Multimodal	60
TADPOLE	Unimodal	5
	Multimodal	15
Simulation	OvO (20 modalities)	2
	Cross and Self-attention	20

4. Significance Testing

We use a t-test to determine if there is a significant difference the performance metrics (AUROC, AUPRC, MAUC, BCA) 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 MIMIC IV and CXR dataset, we compare against self-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 single-attention and OvO attention. The p-value for the AUROC scores is 0.00363 and the p-value for AUPRC is 0.000948. For the TADPOLE challenge, we compare against cross-attention as it performed the second best after OvO. We get a p-value for MAUC scores of $2.09e^{-7}$ and a p-value of $8.19e^{-12}$ for BCA. Thus, we demonstrate a statistically significant difference in MAUC and BCA means between self-attention and OvO attention. Lastly, for the eICU dataset, we compare against cross-attention as it performed the second best after OvO. We get a p-value for AUROC scores of $2.00e^{-15}$ and a p-value of $8.24e^{-14}$ for AUPRC. Thus, we demonstrate a statistically significant difference in AUROC and AUPRC means between self-attention and OvO attention.

Table A2. Results of t-tests comparing the performance metrics between OvO attention and the next best-performing multimodal models across different datasets.

Dataset	**Comparison Model**	**Metric**	**p-value**
MIMIC	Self-attention	AUROC	0.00363
		AUPRC	0.000948
TADPOLE	Cross-attention	MAUC	2×10^{-7}
		BCA	8×10^{-12}
eICU	Cross-attention	AUROC	2×10^{-15}
		AUPRC	8×10^{-14}

5. Limitations

This paper's primary goal is to address one of the major challenges associated with multimodal datasets - computational resource demand and cost. While we demonstrate the scalability of the OvO attention mechanism and its efficiency in handling multiple modalities, we did not conduct experiments focused on interpretability, which is also a crucial aspect of multimodal learning in the biomedical space. The potential for OvO to reveal modality importance through the learninable W_i parameter is promising, but further experiments are required to explore this interpretability in a meaningful way. Additionally, while we compared OvO to popular attention mechanisms like self-attention and cross-attention, there are other attention variants that could serve as future baselines, though they are not widely used in multimodal models. Lastly, the publicly available multimodal datasets often limit the diversity of modalities we can test, with many datasets primarily consisting of tabular data.

References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017).