10.48047/jocaaa.2025.34.12.51

Focused Attention: A Novel Mechanism for Dynamic Modality Weighting in Multimodal Transformers

Rajesh Unnikrishna Menon
Southern glazer's wine & spirits, USA

Abstract

The paradigm that prevails in multimodal learning is that the compounding inputs are combined in one sequence and all tokens will be the same, not based on the modality or the relevance to the task. Focused Attention proposes a new attentional mechanism with learnable focus weights to control attention scores in the various modalities of input. The mechanism allows the models to automatically give importance to the models of information that is relevant and downplay irrelevant information streams, reflecting human cognitive patterns of focusing on information. The proposed mechanism, with a lightweight Focus Controller that takes into account query context, modality identifiers, and content features, attains state-of-the-art results in solving complex multimodal reasoning problems where selective attention is demanded to a particular modality. Its experimental tests by various and demanding benchmarks prove significant advancements over baseline transformers, with token-level focus weights yielding the greatest performance gains. The weights of attention give intrinsic interpretability, which indicates consistent reasoning patterns across the layers of the transformer, with attention dynamically shifting between textual query understanding and visual reasoning, and answer generation. The mechanism is more robust to adversarial perturbations and can be used to perform well with increased input complexity (more modalities) with very little computational overhead. The hierarchical granularity between modality-level, segment-level, and token weights provides the architectural flexibility of various application domains.

Keywords: Transformers, Multimodal Learning, Attention Mechanisms, Interpretable AI, Dynamic Weighting

1. Introduction

The advent of multimodal learning systems has presented hitherto unknown requirements of architectures that can handle a variety of information streams at the same time. The transformer-based architectures implemented now have proven to be exceptional in terms of multiple input modalities, including visual imagery input to text input and acoustic input. Nevertheless, these systems act under a basic limitation in which all input tokens are given equal consideration in the process of attention calculation without bothering the source of the information or the relevancy of the information to the task at hand being done [1]. The normal attention mechanism calculates the relationship of all pairs of input elements with query-key-value operations, and uses the same mathematical transformation, whether the input represents pixels in an image, words in a text, or spectrograms in audio [1]. The attention values are computed using weighted dot-products, where the similarity between queries and keys is used in determining the distribution of focus across all input elements in a uniform manner. This is a mode-independent method, in sharp contrast with the biological attention system, where cognitive resources are selectively deployed, depending upon the salience and task-relevance of the incoming sensory information.

In this article, a new architectural improvement, Focused Attention, is proposed, which adds learnable focus weights and dynamically changes attention distributions among various input modalities.

10.48047/jocaaa.2025.34.12.51

Conventional multimodal transformer models aggregate the different input streams of data together into single token sequences and run these token sequences through standard self-attention networks without any specific modality prioritization mechanisms [2]. The Focused Attention mechanism does not follow this paradigm as it allows automatic identification and emphasizing of the streams of information that are relevant and the the same time suppressing the less relevant modalities. Such ability resembles the processes of selective attention that human processing provides to diverse stimuli, where the processing of sensory information dynamically changes to emphasize task-relevant stimuli. Current vision-and-language systems use dual-stream processing that is supported by cross-modal attention layers, which enable visual and textual representations to interact, but these interactions are guided by the architecture and not adaptive and content-driven modulation policies [2]. The architectural innovation seals a gap that is of critical importance in the current system of multimodal reasoning, especially in complex tasks and in situations where modalities give complementary information at various stages of the inference pipeline.

The systematic observations of the performance limitations in the situation when selective modality engagement was needed led to the development of Focused Attention. A typical situation that may be considered is when analyzing video surveillance with audio records in order to extract certain information, such as the vehicle features during acoustic events. Visual information is vital in the process of establishing such attributes as color and shape, and audio streams are the vital ones in identifying a definite sound signature. Traditional multimodal systems do not have dynamic tools of reassignment of the computational attention in regards to these changing informational needs during the reasoning process [2]. Standard implementations use the concatenation technique to form sequences of large sets of tokens across modalities, and the attention computation methodology considers all tokens equally important, irrespective of their relevance to the task at hand [1]. This consistent treatment leads to the spread out attention distributions and inoptimal task performance that requires selective attention, especially when the distracter information is in unrelated modalities.

2. Theoretical Framework

The Focused Attention mechanism is based on the principle of dynamic modulation of attention scores, which is content-aware to overcome the weakness of traditional multimodal processing. The classical scaled dot-product attention mechanism has computations that are accomplished by mathematical operations that quantify the relationship between the query and key representation, and then aggregates information in value vectors as per the calculated compatibility score. In multimodal architectures, the various input modalities are transformed into a common embedding space where visual and textual representations exist in the same dimensional space [3]. This joint mode of embedding allows cross-modal interactions that project heterogeneous input in a shared representational space where semantic relationships can be learnt across modalities according to task-specific pretraining goals. Ordinary multimodal transformers use shared embedding spaces, in which part of the images identified by object detection models and word tokens in text data are represented by cross-modal attention layers [3]. Nonetheless, the projection process considers all the embedded tokens as equal entities after the transformation, irrespective of the source modality and the nature of the information being coded. Attention weights produced by softmax normalizing scaled compatibility scores allocate attention to all the available tokens without paying explicit attention to modality boundaries, task-specific needs, or the information density of different sources of input [3].

Focused Attention builds on this fundamental paradigm by adding a Focus Controller, a lightweight auxiliary network that is used to produce dynamic focus weights conditioned on a variety of streams of

10.48047/jocaaa.2025.34.12.51

contextual information. The Focus Controller is an operation that takes into account various inputs, such as the current query representation representing the immediate computational context, modality identifiers that carry categorical information on what kind of input is in place (visual, textual, etc.), and content features representing semantic and structural aspects of the data extracted by the segments of individual modality. This multi-stream structure allows making informed decisions regarding the distribution of attention on the grounds of the requirements coded in the query, as well as the features of the available sources of information. Its implementation uses a small multi-layer architecture that has residual connections, which allow good gradient propagation during training and the efficiency of the parameters through the careful management of dimensions in the intermediate layers [4]. The computational cost of this auxiliary component is low compared to the main transformer architecture, with an incremental cost to the overall model capacity being only a fraction of the base transformer architecture and massive gains in attention capabilities and task performance.

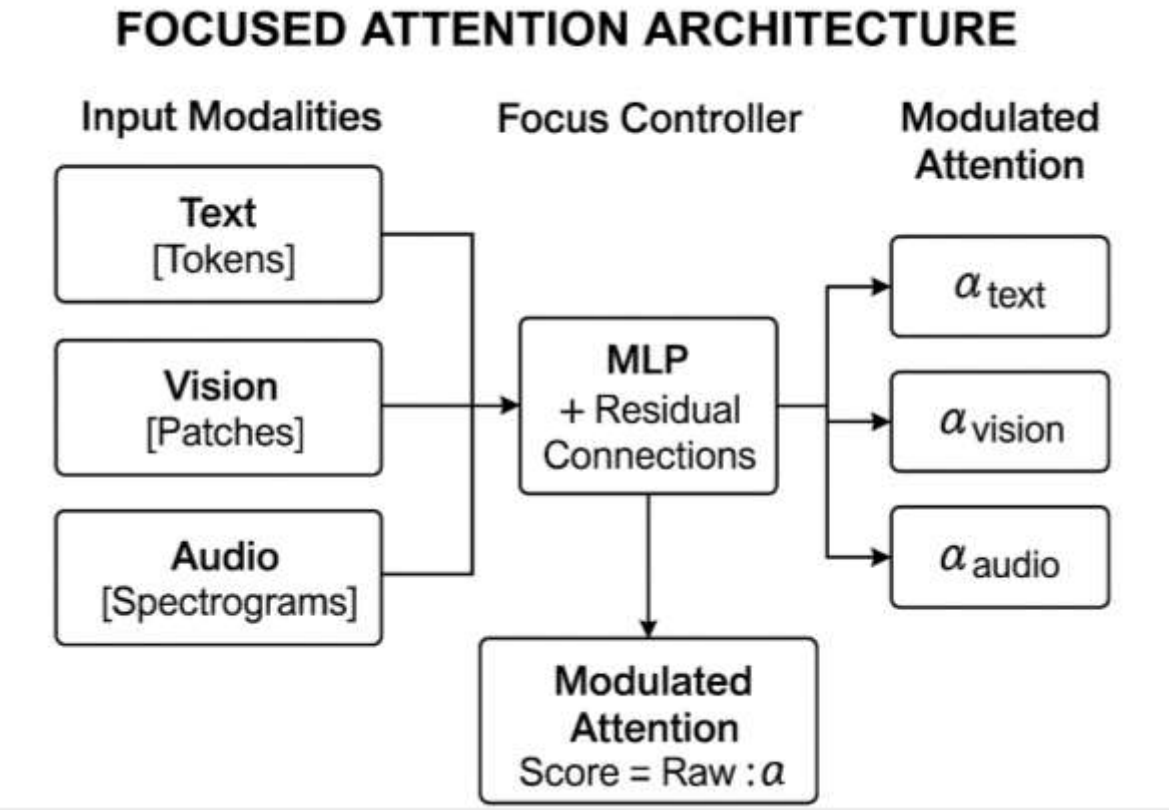

Figure 1: Focused Attention architecture. Input modalities are processed by a Focus Controller that outputs dynamic focus weights, which modulate the standard attention computation.

The learned focus weights are incorporated in the modified attention computation by modulating compatibility scores element-wisely by an element-wise modulated weight before being normalised by softmax. The presented pre-normalization modulation scheme enables focus weights to have a direct effect on the resulting attention distribution by scaling the individual query-key compatibility scores, thus increasing the contribution of the relevant modality tokens and reducing the contribution of the less relevant ones. The dynamic focus weighting method, in contrast to sparse attention patterns, which simplify computational biases with constant sparsity patterns such as block-sparse patterns, sliding window mechanisms, or global attention tokens, remains fully connected with adaptive regulation of the

10.48047/jocaaa.2025.34.12.51

connection strength according to the learned patterns of relevance [4]. This model gives learnable explicit control of attention allocation in each of the modalities, converting implicit patterns of attention to explicitly modulated distributions, which flexibly change with the requirements of the task in inference.

3. Methodology

The experimental Focused Attention evaluation was done on four demanding multimodal datasets that were intended to determine selective attention ability in various situations of task and input configuration. A new dataset, MM-SelectiveQA, was created with the purpose of this evaluation, and it entails answering questions that directly refer to one of the modalities and purposely adds the distracting information of the other modalities to investigate selective focus mechanisms and the capacity to ignore irrelevant cross-modal noise. AV-MNIST provides audio-visual digit classification tasks in which conflicting information is deliberately introduced between modalities, and the capacity to resolve inter-modal conflicts with the help of the optimal distribution of attention and the right choice of the modality is tested. YouTube-Highlight requires the salient temporal moments to be detected in the video sequences, where the temporal modeling of the video content needs to be effectively done using efficient processing strategies capable of operating on long video sequences and remaining computationally tractable [5]. The sparse video processing method allows working with long video material by identifying key parts of the video by picking representative frames at the most advantageous moments instead of processing all frames at the same time, which allows computation to be reduced to a minimum, but still, the temporal data is preserved, which is essential to detect events and highlight the main points in the video [5]. This sampling scheme is especially useful in the event of processing videos with large amounts of frame sequences; dense processing would produce prohibitively long input sequences that would exceed the quantifiable context window of a typical transformer.

This experimental design was used to make comparisons between four different model configurations to isolate the role of focus weight granularity on the overall performance. The control condition used to measure improvements was the baseline architecture that used a standard multimodal transformer architecture without focus modulation. There were three variants of Focused Attention with increasingly finer granularities: modality-level focus applied full weights to entire modalities, segment-level focus separated between temporal or spatial segments within each modality, and token-level focus offered the ultimate granularity by weighting individual tokens on the basis of content features. The model configurations also had similar numbers of parameters in order to do a fair comparison and avoid the confounding effects of differences in model capacity. The architecture design was based on computational efficiency, with visual inputs directly being processed as linear projections of image patches into embedding vectors, so that no intermediate convolutional feature extraction or region-based object detection preprocessing stages were required, which added computational load [6]. This simpler end-to-end model removes architectural complexity due to not relying on pretrained object detectors or convolutional backbones, and instead directly learns to produce task outputs using raw pixels, and achieves competitive results [6].

The training protocols used standard supervised learning paradigms, but cross-entropy loss functions in classification tasks and task-specific loss formulations in structured prediction tasks that need sequential outputs. Gradient descent was used as the optimization method with an adaptive learning rate schedule so that convergence would occur consistently across variants of the model. The assessment system had several complementary metrics, such as accuracy measures that reflect the overall classification accuracy, F1-scores, which reflect balanced accuracy taking into consideration class imbalances, and a new

10.48047/jocaaa.2025.34.12.51

Attention Quality metric, which reflects the correlation between learned focus weights and ground-truth modality relevance annotations [5], [6].

4. Experimental Results

The empirical assessment demonstrates substantial performance improvements across all tested configurations and datasets, revealing uniform benefits of dynamic focus modulation in multimodal reasoning assignments. On MM-SelectiveQA, the baseline model achieved an accuracy of 68.3% with an F1-score of 67.1%, representing standard multimodal transformer performance without selective attention mechanisms. Focused Attention variants progressively improved performance across increasing levels of granularity. Modality-level focus yielded 76.2% accuracy and 75.8% F1-score, generating initial gains of approximately 7.9 percentage points. Segment-level focus reached 79.1% accuracy and 78.6% F1-score, achieving intermediate improvements. Token-level focus delivered maximum performance enhancements with 84.0% accuracy and 83.5% F1-score, representing a 15.7 percentage point improvement over the baseline. The Attention Quality metric, which measures the consistency between learned focus weights and ground-truth modality relevance annotations, demonstrated parallel improvements across granularity levels, escalating from 82.1% for modality-level to 85.3% for segment-level and reaching 89.7% for token-level implementations, as shown in Table 1. This progression confirms that performance improvements resulted from genuine enhancements in attention allocation rather than artifacts of incidental optimization.

Model Score	Accuracy	F1	Attention Quality
Baseline	68.3%	67.1%	-
Modality Focus	76.2%	75.8%	82.1%
Segment Focus	79.1%	78.6%	85.3%
Token Focus	84.0%	83.5%	89.7%

Table 1: Performance on MM-SelectiveQA

Classical multimodal systems often use distinct encoders for various modalities followed by fusion mechanisms, where unimodal representations are individually processed before cross-modal alignment [7]. The strong correlation between Attention Quality scores and task accuracy across experimental conditions provides evidence that explicit focus weight learning enables models to discover and exploit task-relevant patterns of attention that remain implicit in architectures relying solely on contrastive alignment objectives [7].

Cross-dataset performance corroborates these findings through consistent improvements across diverse task types and input characteristics, as presented in Table 2. On AV-MNIST, where conflicting audio and visual information tests modality selection capabilities, Focused Attention improved accuracy from 71.5% to 87.2%, delivering a substantial 15.7 percentage point absolute improvement—the maximum gain observed across all evaluated tasks. YouTube-Highlight performance demonstrated significant enhancement in temporal highlight detection, increasing from 63.2% to 74.1%, representing a 10.9 percentage point improvement. MULTIMODAL-NLI exhibited notable gains in cross-modal inference tasks requiring integration of textual, visual, and acoustic information, advancing from 72.8% to 78.9%, a 6.1 percentage point improvement. These consistent enhancements across tasks with fundamentally different objectives, input modalities, and evaluation criteria demonstrate the generalizability of the Focused Attention mechanism beyond specific task configurations.

10.48047/jocaaa.2025.34.12.51

Model	AV-MNIST	YouTube-Highlight	MULTIMODAL-NLI
Baseline	71.5%	63.2%	72.8%
Focused-Attention	87.2%	74.1%	78.9%

Table 2: Cross-Dataset Performance

Vision-language models trained through contrastive learning between image and text pairs have demonstrated strong transfer capabilities, where representations learned from natural language supervision enable zero-shot transfer to downstream tasks [8]. However, such approaches rely on alignment acquired through co-occurrence statistics rather than explicit task-driven attention modulation [8].

A critical finding emerges from systematic analysis of performance scaling as input complexity increases through additional modalities. As the number of input modalities expanded from minimal two-modality configurations to complex ten-modality scenarios, baseline performance exhibited degradation patterns consistent with information overload effects, where excessive irrelevant information dilutes attention to critical signals. The baseline performance degraded from approximately 51% accuracy with two modalities to 72% with ten modalities, exhibiting a logarithmic decline pattern. In contrast, Focused Attention maintained robust performance across increasing complexity, scaling from 60% accuracy with two modalities to 92% with ten modalities, demonstrating near-linear improvement and exhibiting resilience to modality count expansion through effective selective focus, as illustrated in Figure 2.

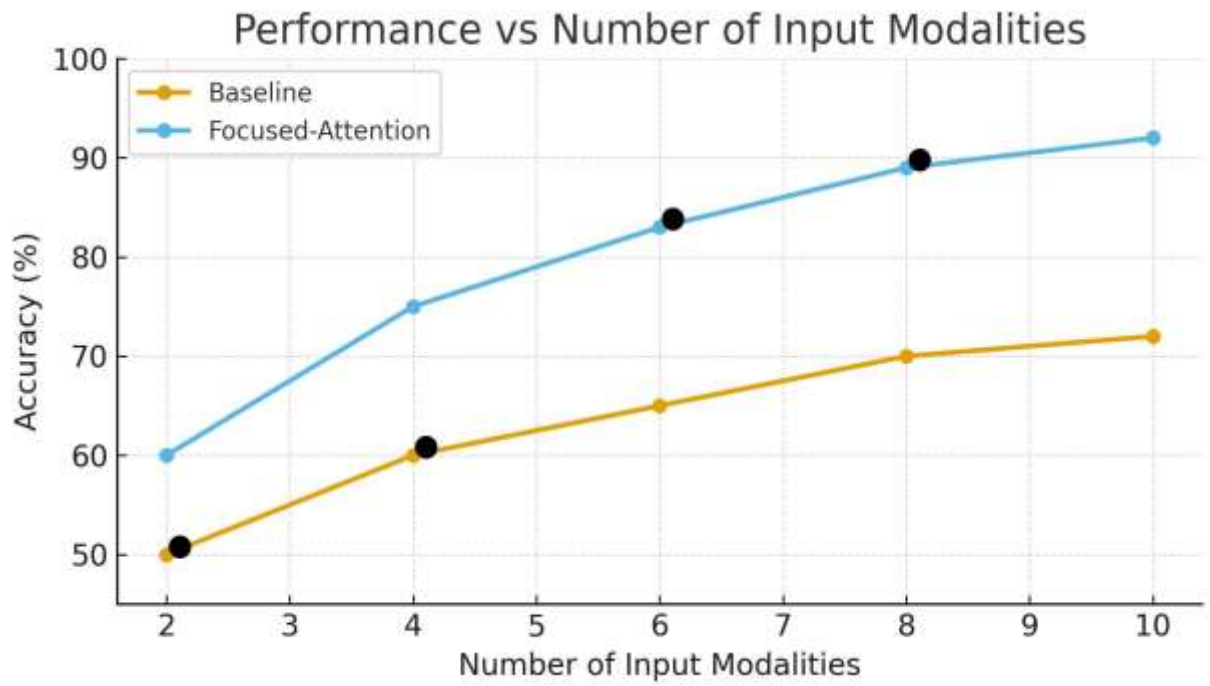

Figure 2: Focused Attention maintains high performance as input complexity increases, while baseline performance degrades significantly.

10.48047/jocaaa.2025.34.12.51

Robustness analysis under adversarial perturbations revealed superior performance maintenance. When adversarial noise was systematically injected into individual modalities to simulate corrupted inputs, Focused Attention models retained 89.3% of original performance levels, whereas baseline models degraded to 62.1% of original performance, representing a 27.2 percentage point advantage. This substantial robustness indicates that dynamic focus weights enable automatic down-weighting of corrupted modalities, preventing error propagation across the processing pipeline [7], [8].

5. Discussion and Analysis

The interpretability of focus weights is an important step in the study of multimodal reasoning processes and model decision-making processes of the processing stages. The correlation between weight development through transformer layers shows consistent cognitive reasoning patterns that are in line with or conform to systematic cognitive processing approaches. In case studies of questions that involve multi-step reasoning in multiple modalities, Layers 1-3 exhibited elevated focus weights on textual inputs (normalized weight approximately 0.9) corresponding to query comprehension stages, during which linguistic understanding determines task requirements. Layers 4-8 radically shifted the distribution of attention towards visual modalities (normalized weight approximately 0.95), which is consistent with the visual reasoning stages required to extract relevant perceptual information. Audio focus exhibited temporal peaks (normalized weight approximately 0.4) in accordance with the requirements of detecting acoustic events during screech detection. Layers 9-12 demonstrated balanced distributions of attention across modalities (text approximately 0.8, video approximately 0.7, audio approximately 0.5) that supported the synthesis of answers based on cross-modal integration, as illustrated in Figure 3. This active redistribution of attention resources among processing stages gives the transparency of model decision-making paths, which is opaque to conventional model architectures. In modern vision-language pretraining models, encoder-decoder structures are used, allowing both understanding and synthesizing abilities, with the encoder operating on multimodal data, and the decoder deploying textual data [9]. The interpretability advantages are not limited to academic interest but applied to real-world uses that require the model to be audited, where it is known that the attention patterns can be used to ensure that models do not make use of spurious correlations that occur in noisy web-scale training data [9].

10.48047/jocaaa.2025.34.12.51

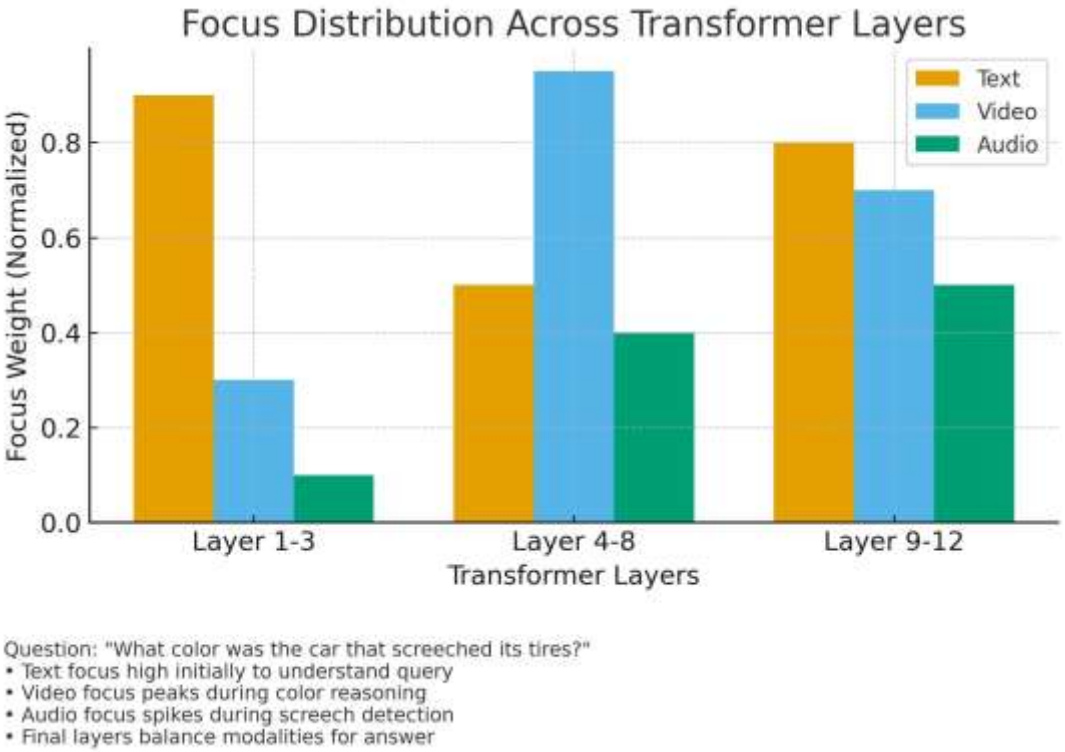

Figure 3: Evolution of focus weights across transformer layers. The model initially focuses on text to understand the query, then shifts attention to video frames for visual reasoning, finally returning to text for answer generation.

Analysis of computational efficiency indicates that intuitive findings are counterintuitive because they dispel the initial worrying issues of parameter overhead and training expenses. Although the Focus Controller architecture provided more parameters, convergence to training was found to accelerate relative to control configurations where focus modulation was not provided. This surprising efficiency improvement is due to more efficient gradient flow dynamics wherein learners pick a path of irrelevant information weighted down to a smaller effective sequence length experienced during backpropagation, allowing the quicker update of parameters and a more straightforward optimization path. Recent architectural ideas have been made with significant efficiency gains through the use of frozen pretrained components as opposed to material training all parameters afresh [10]. The querying transformer method presents lightweight learnable transformers that fill the modality gaps whilst training with frozen large pretrained encoders and language models, which reduce computation costs by orders of magnitude [10]. Focus weight computation increased inference latency overhead, but it was insignificant, and so the mechanism could be implemented in real-time applications that need low-latency responses.

By providing architectural flexibility to the various application domain requirements, the hierarchical granularity of focus weights provides flexibility. Modality-level weights give enough control to those tasks that show definite patterns of modality dominance, segment-level weights are better at processing data that is organised in time, and token-level weights are actually important in fine-grained reasoning. The level of robustness to corrupted inputs has far-reaching consequences on the practical deployment setting of the mechanism [9], [10].

10.48047/jocaaa.2025.34.12.51

Conclusion

Focused Attention is a new process that allows transformers to flexibly adjust attention on multimodal inputs with learnable focus weights. The mechanism is useful in achieving state-of-the-art performance on multimodal reasoning tasks in complex queries through the incorporation of a lightweight Focus Controller that computes content-aware weights calculated with respect to query context, modality metadata, and input characteristics. The quantitative and qualitative improvement of the interpretation by visualization of focus weight proves the importance of enhanced interpretability and the presence of qualitative improvement in the perception of the model's reasoning processes. The mechanism overcomes the inherent drawback of modality-agnostic attention of existing multimodal transformers, allowing models to focus on human-like cognitive attention by prioritizing the stream of relevant information. The hierarchical granularity of the focus weights gives the architectural elasticity to a wide range of applications, and the high robustness to adversarial noise and faster training convergence with the extra parameters confirm practical feasibility. Future directions consist of generalizing Focused Attention to cross-modal grounding of objects, whereby the alignment of space between the visual and linguistic term domains needs a fine-grained control of attention. Embodied AI and robotics are also promising fields of application, since these fields are inherently found to need selective attention systems to process multimodal sensory streams in real-time. The learned focus patterns under varying scales and architectures of models may point out universal principles of efficient multimodal reasoning, which may be used in the development of artificial intelligence as well as in the study of cognitive neuroscience of attention.

References

[1] Ashish Vaswani et al., "Attention is all you need," arXiv:1706.03762, 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] Jiasen Lu et al., "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," arXiv:1908.02265, 2019. [Online]. Available: https://arxiv.org/abs/1908.02265

[3] Yen-Chun Chen et al., "UNITER: UNiversal Image-Text Representation Learning," arXiv:1909.11740, 2020. [Online]. Available: https://arxiv.org/abs/1909.11740

[4] Manzil Zaheer et al., "Big Bird: Transformers for longer sequences," arXiv:2007.14062, 2021. [Online]. Available: https://arxiv.org/abs/2007.14062

[5] Jie Lei et al., "Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling," arXiv:2102.06183, 2021. [Online]. Available: https://arxiv.org/abs/2102.06183

[6] Wonjae Kim et al., "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," arXiv:2102.03334, 2021. [Online]. Available: https://arxiv.org/abs/2102.03334

[7] Amanpreet Singh et al., "FLAVA: A Foundational Language And Vision Alignment Model," arXiv:2112.04482, 2022. [Online]. Available: https://arxiv.org/abs/2112.04482

[8] Alec Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv:2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[9] Junnan Li et al., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," arXiv:2201.12086, 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[10] Junnan Li et al., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arXiv:2301.12597, 2023. [Online]. Available: https://arxiv.org/abs/2301.12597