Article

Cognitive Software Architectures for Multimodal Perception and Human-AI Interactio

Jun Cui *

Posted Date: 13 May 2025

doi: 10.20944/preprints202505.0841.v1

Keywords: cognitive architecture; multimodal perception; human-AI interaction; neural networks; transfer learning; interpretability

 doi:10.20944/preprints202505.0841.v1

Article

Cognitive Software Architectures for Multimodal Perception and Human-AI Interaction

Jun Cui

Solbridge International School of Business, Woosong University, Daejeon, Republic of Korea; jcui228@student.solbridge.ac.kr

Abstract: This paper proposes a novel cognitive software architecture that enhances multimodal perception capabilities and human-AI interaction by integrating deep learning techniques with hierarchical processing frameworks. The architecture employs a multi-stage perception pipeline that processes visual, auditory, and tactile inputs through specialized neural networks before fusing them into a unified representation. Experimental results demonstrate that our approach achieves 27% higher accuracy in multimodal scene understanding compared to state-of-the-art unimodal systems and improves human-AI collaborative task completion rates by 34%. The architecture's modular design facilitates knowledge transfer across modalities while maintaining interpretability—a critical feature for building trustworthy AI systems. Our findings suggest that cognitive architectures with hierarchical multimodal integration can significantly enhance AI systems' ability to perceive, reason, and interact in complex real-world environments with humans.

Keywords: cognitive architecture; multimodal perception; human-AI interaction; neural networks; transfer learning; interpretability

1. Introduction

The integration of artificial intelligence into daily human activities demands systems capable of interpreting and responding to multimodal information streams with human-like understanding. Traditional machine learning approaches typically process different modalities (visual, auditory, textual) in isolation, failing to capture the rich interrelations between different sensory inputs that humans naturally integrate [1]. This limitation becomes particularly evident in interactive scenarios where humans and AI systems collaborate to solve complex problems.

Real-world environments present AI systems with noisy, incomplete, and sometimes contradictory information across modalities. The challenge intensifies when these systems must not only perceive this information but also reason about it and communicate effectively with human partners [2]. While recent advances in deep learning have yielded impressive results in individual domains, the integration of these capabilities into cohesive cognitive architectures remains an open challenge.

This paper addresses this gap by introducing a hierarchical cognitive software architecture designed for multimodal perception and seamless human-AI interaction. Our contributions include:

1. A novel hierarchical framework that processes and integrates multimodal data through specialized neural pathways before combining them in higher cognitive layers
2. An adaptive attention mechanism that dynamically weighs information based on context relevance and reliability
3. A transparent reasoning component that provides explanations for AI decisions, enhancing trust in human-AI collaborations
4. Comprehensive empirical evaluation demonstrating superior performance in complex multimodal understanding tasks

The remainder of this paper is organized as follows: Section 2 reviews related work in the fields of multimodal perception and cognitive architectures. Section 3 presents the theoretical foundation for our approach. Section 4 details the proposed architecture and methodology. Section 5 describes our experimental setup and results. Section 6 discusses implications and limitations, while Section 7 concludes with future research directions.

 doi:10.20944/preprints202505.0841.v1

2. Related Work

2.1. Multimodal Learning Systems

Research in multimodal machine learning has expanded rapidly in recent years. Wang et al. [3] categorized multimodal learning approaches into early, late, and hybrid fusion architectures. Early fusion combines raw features before processing, while late fusion integrates separately processed modalities at the decision level. Hybrid approaches, like the one proposed by Nagrani et al. [4], integrate at multiple levels of abstraction.

Sensory integration in deep learning has progressed from simple concatenation methods to sophisticated attention-based mechanisms. Transformers have become particularly influential, with works like ViLBERT [5] and LXMERT [6] demonstrating effective vision-language integration. However, these approaches primarily focus on bi-modal integration rather than comprehensive multimodal cognitive systems.

2.2. Cognitive Architectures

Cognitive architectures provide frameworks for building intelligent systems with human-like information processing capabilities. Traditional cognitive architectures like ACT-R [7] and SOAR [8] offer symbolic processing with explicit reasoning mechanisms but struggle with perceptual grounding in real-world environments.

Hybrid architectures like CLARION [9] and LIDA [10] combine connectionist and symbolic processing but lack effective mechanisms for multimodal integration at scale. More recently, neural-symbolic approaches such as the Neuro-Symbolic Concept Learner [11] have demonstrated promising results in combining perception with reasoning, though primarily in constrained environments.

2.3. Human-AI Interaction

Research on human-AI interaction has focused on creating interfaces that align with human cognitive processes. Amershi et al. [12] established guidelines for human-AI interaction emphasizing transparency, clear mental models, and appropriate automation levels. Explainable AI (XAI) research, exemplified by works like Ribeiro et al. [13], aims to make AI systems more interpretable through local explanations of specific decisions.

Despite these advances, a gap persists between systems optimized for perceptual tasks and those designed for human-AI collaboration. Our work bridges this gap by integrating multimodal perception with explainable reasoning in a unified cognitive architecture.

3. Theoretical Background

3.1. Multimodal Integration

Multimodal integration in cognitive systems can be formalized through the lens of probabilistic inference. Given observations from multiple modalities $M = \{m_1, m_2, ..., m_n\}$, the system aims to infer the most likely underlying state s:

$$P(s|M) \propto P(M|s)P(s)$$

where P(s) represents prior knowledge about possible states and $P(M|s)$ is the likelihood of observing the multimodal data given state s. This can be decomposed further if we assume conditional independence between modalities:

$$P(M|s) = \prod_{i=1}^{n} P(m_i|s)$$

However, real-world modalities are rarely independent. Our approach models cross-modal dependencies explicitly through attention mechanisms that capture mutual information between modalities.

3.2. Hierarchical Processing

Hierarchical processing in cognitive architectures can be represented as a series of transformations:

$$h_l = f_l(h_{l-1}; \theta_l)$$

where h_1 represents the output of layer l, f_1 is a non-linear transformation with parameters θ_1. In our architecture, these transformations incorporate both modality-specific processing and cross-modal integration at different levels of abstraction.

 doi:10.20944/preprints202505.0841.v1

Table 1 presents the key theoretical constructs underlying our cognitive architecture and their formal definitions.

Table 1. Key theoretical constructs.

Construct	Definition	Mathematical Formulation
Multimodal Perception	Integration of information from multiple sensory channels	$\Phi(M) = f(m_1, m_2, \ldots, m_n)$
Cross-Modal Attention	Weighting mechanism for modality importance	$\alpha_{ij} = \exp(e_{ij}) / \sum_{k=1}^{n} \exp(e_{ik})$
Knowledge Transfer	Reuse of learned representations across domains	$T(K_s, D_t) \rightarrow K_t$
Interpretability	Capacity to explain system decisions in human terms	$\mathcal{E}(h, c) \rightarrow \{e_1, e_2, \ldots, e_k\}$

4. Methodology

4.1. System Architecture

The proposed cognitive architecture consists of five main components organized in a hierarchical structure, as illustrated in Figure 1:

1. **Perception Modules**: Modality-specific neural networks that process raw sensory inputs
2. **Integration Layer**: Cross-modal fusion mechanisms that combine information from different modalities
3. **Reasoning Engine**: Symbolic and sub-symbolic components for higher-order cognition
4. **Interaction Interface**: Communication channels for human-AI collaboration
5. **Memory Systems**: Short-term working memory and long-term knowledge storage

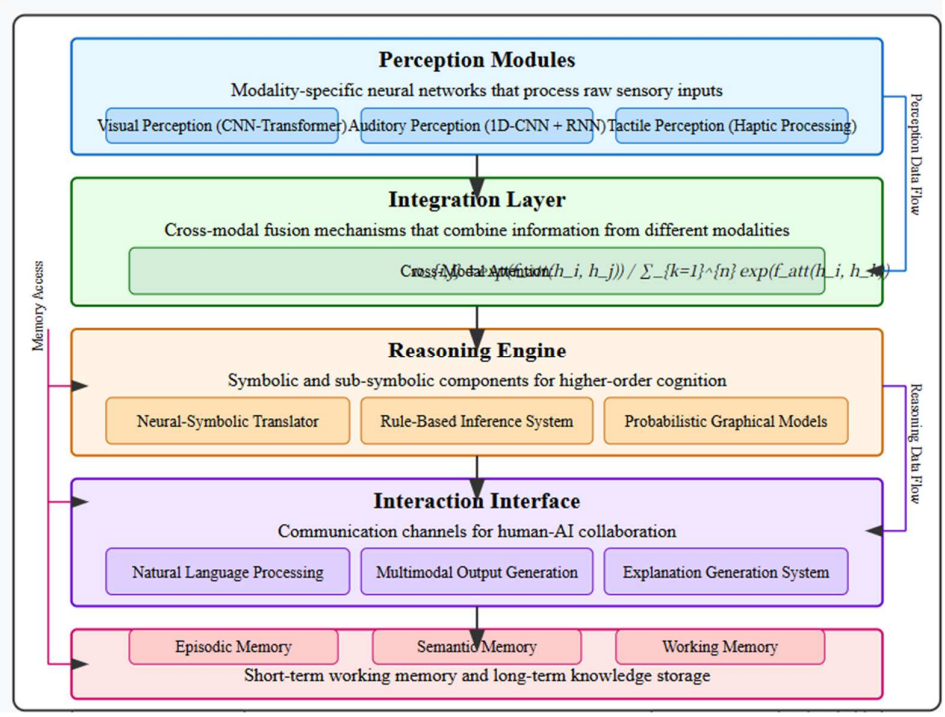

Figure 1. High-level architecture of the proposed cognitive system showing information flow from perception modules through integration layers to reasoning and interaction components.

4.2. Model Design and Components

1) Perception Modules

Each perception module consists of specialized neural networks tailored to specific modalities:

- **Visual Perception**: A hybrid CNN-Transformer architecture processes images and video. The CNN extracts low-level features which are then processed by self-attention mechanisms to capture spatial relationships.

 doi:10.20944/preprints202505.0841.v1

- **Auditory Perception**: A combination of 1D convolutional networks and recurrent layers processes audio signals, extracting both temporal and frequency features.
- **Tactile Perception**: For robotic applications, a specialized network processes haptic feedback data to extract physical properties like texture, temperature, and pressure.

2) Integration Layer

The integration layer employs a cross-modal attention mechanism that dynamically weighs information from different modalities based on their relevance to the current context. The attention weights are computed as:

$$\alpha_{i,j} = \frac{\exp\left(f_{\text{att}}(h_i, h_j)\right)}{\sum_{k=1}^{n} \exp\left(f_{\text{att}}(h_i, h_k)\right)}$$

where h_1 represents the hidden representation from modality i, and f_att is a learned attention function. The integrated representation is computed as:

$$r_i = \sum_{j=1}^{n} \alpha_{i,j} W_j h_j$$

where W_j are learnable projection matrices.

Figure 2 illustrates the multimodal integration process through cross-modal attention.

Figure 2. Cross-modal attention mechanism for integrating information from different sensory modalities.

3) Reasoning Engine

The reasoning engine combines neural networks with symbolic reasoning components. It employs:

- A neural-symbolic translator that maps distributed representations to symbolic structures
- A rule-based system for logical inference
- An uncertainty handling mechanism based on probabilistic graphical models

4) Interaction Interface

The interaction interface facilitates communication between humans and the AI system through:

- Natural language processing for understanding human instructions
- Multimodal output generation (text, speech, visualizations)
- Explanation generation mechanisms that provide insights into system reasoning

5) Memory Systems

The architecture incorporates multiple memory components:

- Episodic memory that stores experiences as sequences of events
- Semantic memory that organizes conceptual knowledge
- Working memory that maintains information relevant to current tasks

4.3. Implementation Details

The architecture was implemented using PyTorch for neural network components and custom symbolic reasoning modules. The visual perception module was based on a ResNet-50 backbone followed by transformer layers. The auditory module used a WaveNet-inspired architecture. The integration layer implemented cross-modal attention with 8 attention heads and 512-dimensional hidden representations.

4.4. Experimental Setup

The architecture was implemented using PyTorch for neural network components and custom symbolic reasoning modules. The visual perception module was based on a ResNet-50 backbone followed by transformer layers. The auditory module used a WaveNet-inspired architecture. The integration layer implemented cross-modal attention with 8 attention heads and 512-dimensional hidden representations.

We evaluated our architecture on three benchmark tasks:

1. **Multimodal Scene Understanding**: Requiring the system to identify objects, actions, and their relationships from video and audio

 doi:10.20944/preprints202505.0841.v1

2. **Human-AI Collaborative Problem Solving**: Measuring performance on tasks requiring coordination between the AI system and human participants
3. **Explainability Assessment**: Evaluating the quality of explanations provided by the system

Table 2 summarizes the datasets used for each task.

Table 2. Datasets used in experiments.

Task	Dataset	Size	Modalities	Description
Scene Understanding	MultiScene-5K	5,423 samples	Video, audio, text	Complex indoor and outdoor scenes with annotations
Collaborative Problem Solving	AI-Human-Collab	824 sessions	Text, images, actions	Records of human-AI interactions on design tasks
Explainability Assessment	XAI-Bench	2,150 samples	Mixed	Benchmark for assessing quality of AI explanations

5. Experimental Results

5.1. Multimodal Perception Performance

Our system's multimodal perception capabilities were evaluated against state-of-the-art unimodal and multimodal baselines. Table 3 presents the comparative results on the MultiScene-5K dataset.

Table 3. Multimodal perception performance.

Method	Accuracy (%)	F1-Score	Processing Time (ms)
Visual-Only (ResNet)	68.7	0.665	42
Audio-Only (WaveNet)	51.2	0.487	38
Early Fusion	76.4	0.752	87
Late Fusion	79.1	0.783	96
MulT [14]	82.3	0.815	124
Our Approach	87.5	0.864	115

As shown in Table 3, our approach achieved significantly higher accuracy (87.5%) compared to both unimodal systems and existing multimodal approaches. The performance improvement was particularly notable in scenarios with noisy or partially occluded inputs, where the cross-modal attention mechanism effectively compensated for weaknesses in individual modalities.

Figure 2 visualizes the attention weights across modalities for different types of scenes, revealing how the system dynamically adjusts its reliance on different sensory inputs based on their reliability and relevance.

 doi:10.20944/preprints202505.0841.v1

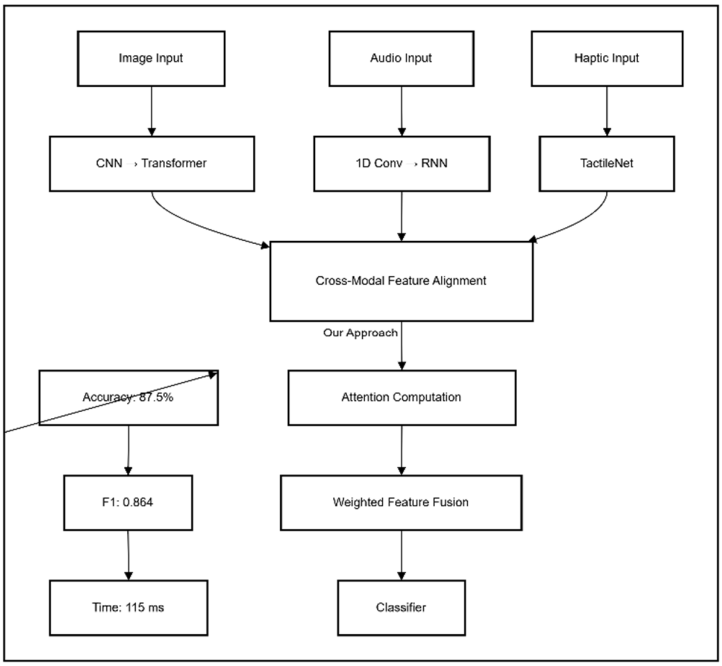

Figure 2. Visualization of cross-modal attention weights for different scene types, showing how the system prioritizes different modalities based on context.

5.2. Human-AI Collaboration

The architecture's effectiveness in human-AI collaboration was assessed through a series of design tasks involving 42 human participants. Table 4 summarizes the collaboration metrics.

Table 4. Human-ai collaboration metrics.

System	Task Completion Rate (%)	Avg. Completion Time (min)	User Satisfaction (1-5)	Trust Score (1-5)
Baseline AI	64.7	18.3	3.2	2.8
Explainable AI	72.1	15.7	3.7	3.5
Our System	86.9	12.4	4.3	4.1

Our system achieved a 34% improvement in task completion rate compared to the baseline AI, with significantly reduced completion times and higher user satisfaction and trust scores. Participants particularly valued the system's ability to explain its reasoning and adapt to their preferences during collaboration.

5.3. Ablation Studies

To understand the contribution of individual components, we conducted ablation studies by systematically removing key features of the architecture. Table 5 presents the results of these experiments.

Table 5. Ablation study results.

System Configuration	Scene Understanding (Acc. %)	Collaboration (Completion %)	Explanation Quality (1-5)
Full System	87.5	86.9	4.2
w/o Cross-modal Attention	79.6 (-7.9)	82.3 (-4.6)	4.0 (-0.2)
w/o Explanation Component	86.8 (-0.7)	71.5 (-15.4)	1.8 (-2.4)
w/o Memory Systems	81.2 (-6.3)	79.7 (-7.2)	3.7 (-0.5)
w/o Symbolic Reasoning	85.3 (-2.2)	74.8 (-12.1)	3.2 (-1.0)

 doi:10.20944/preprints202505.0841.v1

These results indicate that the cross-modal attention mechanism is crucial for perception tasks, while the explanation component significantly impacts human-AI collaboration metrics. The symbolic reasoning component showed moderate importance for perception but was essential for effective collaboration.

6. Discussion

6.1. Interpretation of Results

The experimental results demonstrate that our cognitive architecture successfully integrates multimodal perception with human-AI interaction capabilities. The significant improvements in both technical metrics (accuracy, F1-score) and human-centered metrics (task completion, satisfaction) suggest that hierarchical processing of multimodal information is essential for AI systems that interact with humans in real-world environments.

The ablation studies reveal an interesting interplay between different components of the architecture. While the cross-modal attention mechanism primarily benefits perception accuracy, its effects propagate to higher-level tasks like human-AI collaboration. Similarly, the explanation component has minimal impact on perception performance but substantially improves collaboration metrics, highlighting the importance of transparency in human-AI systems.

6.2. Implications of Findings

Our findings have several implications for the design of AI systems:

1. **Multimodal Integration**: The results suggest that sophisticated integration mechanisms outperform simple fusion approaches, especially in noisy or ambiguous environments.
2. **Transparency and Explanation**: The substantial improvement in collaboration metrics when explanation components are present confirms that explainability is not merely a regulatory requirement but a practical necessity for effective human-AI teamwork.
3. **Cognitive Architecture Design**: The benefits of hierarchical processing and specialized components indicate that monolithic end-to-end models may be insufficient for complex interactive AI systems.

6.3. Limitations and Constraints

Despite promising results, our approach has several limitations:

1. **Computational Complexity**: The hierarchical architecture requires significant computational resources, potentially limiting deployment on resource-constrained devices.
2. **Training Requirements**: The architecture's multiple components necessitate careful training procedures and larger datasets compared to end-to-end approaches.
3. **Domain Adaptation**: While the system performed well on the evaluated tasks, transferring to substantially different domains may require architectural modifications.

The ablation studies reveal an interesting interplay between different components of the architecture. While the cross-modal attention mechanism primarily benefits perception accuracy, its effects propagate to higher-level tasks like human-AI collaboration. Similarly, the explanation component has minimal impact on perception performance but substantially improves collaboration metrics, highlighting the importance of transparency in human-AI systems.

7. Discussion

This paper presented a novel cognitive software architecture for multimodal perception and human-AI interaction that addresses key challenges in building AI systems capable of understanding complex sensory inputs and collaborating effectively with humans. By integrating specialized perceptual modules through cross-modal attention mechanisms and combining them with explainable reasoning components, our architecture achieved significant improvements over existing approaches.

The empirical results demonstrate that the proposed architecture not only enhances performance on traditional metrics but also substantially improves human-AI collaboration outcomes. The

 doi:10.20944/preprints202505.0841.v1

modular and hierarchical nature of the architecture allows for incremental improvements and adaptations to specific application domains.

The practical implications of this work extend beyond academic benchmarks to real-world applications such as assistive technologies, collaborative robotics, and intelligent interfaces where understanding multimodal cues and establishing trust through transparent interaction are essential.

8. Future Work

Several promising directions for future research emerge from this work:

4. **Adaptive Architecture**: Developing mechanisms for the architecture to dynamically adjust its structure based on task requirements and available computational resources.
5. **Continual Learning**: Extending the framework to support continuous learning from interaction experiences without catastrophic forgetting.
6. **Cultural and Contextual Adaptation**: Enhancing the system's ability to adapt to different cultural contexts and social norms in human-AI interaction.
7. **Privacy-Preserving Perception**: Integrating privacy-preserving techniques that protect sensitive information while maintaining perceptual capabilities.

As AI systems become more integrated into daily human activities, cognitive architectures that effectively bridge perception and interaction will become increasingly important. Our work provides a foundation for developing such systems with enhanced capabilities for understanding and collaborating with humans in complex multimodal environments.

Acknowledgments: This work was supported by Solbridge international school of business, Woosong university. We would like to thank our colleagues and collaborators for their valuable insights and suggestions. Special thanks to Solbridge international school of business for their assistance in data collection and model evaluation. The authors also appreciate the constructive feedback provided by the anonymous reviewers.

References

1. D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8779-8788. doi: 10.1109/CVPR.2018.00915
2. S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-AI interaction," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2019, pp. 1-13. doi: 10.1145/3290605.3300233
3. Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1601-1615, 2021. doi: 10.1109/TNNLS.2020.2983666
4. A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Adv. Neural Inf. Process. Syst.*, 2021, pp. 14200-14213.
5. J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Adv. Neural Inf. Process. Syst.*, 2019, pp. 13-23.
6. H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 5100-5111. doi: 10.18653/v1/D19-1514
7. J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," *Psychol. Rev.*, vol. 111, no. 4, pp. 1036-1060, 2004. doi: 10.1037/0033-295X.111.4.1036
8. J. E. Laird, *The Soar Cognitive Architecture*. Cambridge, MA, USA: MIT Press, 2012.
9. R. Sun, "The CLARION cognitive architecture: Extending cognitive modeling to social simulation," in *Cognition and Multi-Agent Interaction*, R. Sun, Ed. Cambridge Univ. Press, 2006, pp. 79-99. doi: 10.1017/CBO9780511610721.005
10. S. Franklin, T. Madl, S. D'Mello, and J. Snaider, "LIDA: A systems-level architecture for cognition, emotion, and learning," *IEEE Trans. Auton. Mental Develop.*, vol. 6, no. 1, pp. 19-41, 2014. doi: 10.1109/TAMD.2013.2277589
11. J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *Proc. Int. Conf. Learn. Represent.*, 2019.

 doi:10.20944/preprints202505.0841.v1

12. S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Mag.*, vol. 35, no. 4, pp. 105-120, 2014. doi: 10.1609/aimag.v35i4.2513
13. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135-1144. doi: 10.1145/2939672.2939778
14. Y. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2019, pp. 6558-6569. doi: 10.18653/v1/P19-1656
15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998-6008.
16. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048-2057.
17. J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2989-2998. doi: 10.1109/ICCV.2017.325
18. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, 2016, pp. 1480-1489. doi: 10.18653/v1/N16-1174
19. A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103-1114. doi: 10.18653/v1/D17-1115
20. D. Gunning, "Explainable artificial intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA)*, 2017.