ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Literature Review: Recent Advances in Computer Vision and Language AI

Suresh Babu Rajasekaran

NVIDIA

ABSTRACT

This comprehensive literature review examines the latest breakthroughs in computer vision and natural language processing (NLP), two rapidly evolving fields with applications across search, human-computer interaction, robotics, and more. It synthesizes key findings, trends, limitations, and open challenges from cutting-edge research at their intersection. The dramatic progress driven by deep neural networks is analysed in depth, along with issues like generalization, context handling, reasoning, uncertainty, and human-centric evaluation. Although remarkable advances have been made, especially in computer vision, core problems remain to be addressed. This review provides a thorough overview of the state-of-the-art, reflecting the most recent innovations, and promising future directions in this dynamic research domain.

***Corresponding author**

Suresh Babu Rajasekaran, NVIDIA.

Received: June 12, 2023; **Accepted:** July 25, 2023; **Published:** August 21, 2023

Keywords: Computer Vision, Natural Language Processing, Deep Learning, Multimodal Learning, Visual Question Answering, Scene Understanding, Context Modelling, Generalization, Reasoning, Human-AI Interaction

Introduction

Computer vision and natural language processing (NLP) have seen explosive growth, with transformative impact across science and industry. Using data-driven deep learning methods, computer vision has achieved super-human performance on tasks like image classification while NLP models can generate remarkably coherent text [1, 2]. Integrating these modalities to perform sophisticated joint visual and linguistic understanding remains an open challenge. This literature review synthesizes current progress and limitations, providing a comprehensive overview of the state-of-the-art at the intersection of computer vision and NLP.

Theoretical Background

Multimodal machine learning aims to model interactions between diverse modalities like vision, language, acoustics, etc. [3]. By learning alignments between textual, visual, and speech data, multimodal systems could unlock new capabilities in areas like visual context reasoning, natural language generation, human-AI interaction, and more. Potential real-world applications range from assistive technologies to intelligent surveillance, search, robotics, autonomous vehicles, and beyond [4]. Tackling these challenges may also spur development of more robust and holistic evaluation methodologies for AI systems.

Computer Vision

Computer vision has achieved remarkable advances in problems like image classification object detection, semantic segmentation and more [5-7]. Convolutional neural networks (CNNs) now dominate, offering superior representation learning abilities over earlier hand-crafted features. Various CNN architectures pre-trained on huge labelled datasets like ImageNet can encode transferable visual features [8]. Recently, Transformer models have also been adapted for computer vision, achieving promising results in tasks like image classification and object detection [9-11].

Unsupervised pre-training has become increasingly popular, with models like BEiT, MAE and Mask Feat (Wei et al., 2021) matching or exceeding supervised pre-training performance [12, 13]. Loss functions based on contrastive learning, predicting masked patches, and other self-supervision signals have proven effective. Transfer learning from these generative models provides benefits across down-stream tasks.

Natural Language Processing

NLP has progressed rapidly, with neural models reaching new milestones in translation question answering dialogue systems and other tasks [14-16]. Pre-trained language models like ELMo, BERT, GPT-3 and T5 have been especially impactful [2, 17-19]. By pre-training on vast unlabelled corpora using objectives like masked language modelling and text generation, they develop general linguistic representations that transfer effectively. Performance on benchmark NLP datasets has substantially increased through their use.

However, fundamental challenges remain around relational reasoning, interpretability, and grounding language in real-world knowledge [20].Combining textual understanding with computer vision provides an avenue for developing more human-like language intelligence.

Vision and Language Integration

There has been growing interest in unified modelling of vision and language (Baltrušaitis et al., 2019). Relevant tasks include image captioning visual question answering visual, video description, embodied agents, grounding textual concepts in images and more. Large datasets like COCO, Flickr30k, and VQA have enabled data-driven progress. Multimodal Transformer models like ViLBERT, LXMERT, VL-BERT, UNITER and OSCAR have proven effective by learning cross-modality representations [21-31]. Other approaches integrate CNN image features into language model architectures via attention mechanisms. However, issues around model interpretability, bias, and spurious correlations remains a concern [32].

Robustly grounding language in rich visual contexts is still difficult, as is leveraging temporal/causal reasoning and common sense knowledge. Generalization across domains and capturing nuanced semantics and higher-order relationships between vision and language remains an open challenge [33]. Developing more human-centric evaluations and multimodal datasets is an active area of inquiry.

Recent Advances

Cutting-edge techniques at top conferences provide insight into current progress and limitations. In computer vision, self-supervised models matching supervised pre-training (BEiT, MAE, MaskFeat) demonstrate the power of contrastive self-supervision (Wei et al., 2021) [12, 13]. Object detectors combining CNNs and Transformers (DETR, Deformable DETR) show advantages of attention for vision (Zhu et al., 2020) [11].

In NLP, models leveraging self-supervision over images and text (CLIP, ALIGN) make promising steps towards grounding language in vision [34, 35]. Large multimodal models like FLAVA achieve strong performance by pre-training on diverse vision, language, speech, and audio data [36]. However, issues like bias, safety, and robust evaluation persist.

There are also emerging efforts to build more human-like multimodal agents using virtual environments and simulations (Yan et al., 2020) [37]. Still translating such progress to real-world domains remains challenging.

Discussion

This review highlights remarkable innovation at the intersection of computer vision and NLP, while also revealing persistent gaps. Fundamental problems around contextual reasoning, uncertainty, generalization, and human collaboration are largely still open. Potential research directions include leveraging neuro-symbolic AI techniques combining neural networks and declarative knowledge, developing causal reasoning capabilities, and increased focus on human-centric benchmarks and safety.

Conclusion

In conclusion, this literature review has provided a comprehensive overview of the state-of-the-art in combined computer vision and NLP systems. Although progress has accelerated, core challenges around semantic understanding, reasoning, and human interaction remain open. As research continues rapidly, multimodal AI systems may gradually approach human-level visual and linguistic intelligence through innovations in representation learning, context modelling, and human-AI collaboration.

References

1. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 770-778.
2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, et al. (2020) Language models are few-shot learners. Advances in neural information processing systems 33: 1877-1901.
3. Baltrušaitis T, Ahuja C, Morency LP (2018) Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41: 423-443.
4. Srivastava N, Salakhutdinov RR (2012) Multimodal learning with deep boltzmann machines. Advances in neural information processing systems 25.
5. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25.
6. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.
7. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition 3431-3440.
8. Deng J, Dong W, Socher R, Li LJ, Li K, et al. (2009) Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition 248-255.
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. (2017) Attention is all you need. Advances in neural information processing systems 30.
10. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929.
11. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, et al. (2020) End-to-end object detection with transformers. In European Conference on Computer Vision Springer, Cham 213-229.
12. Bach FR, Binder A, Montavon G, Klauschen F, Müller KR, et al. (2021) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one 10: e0130140.
13. He K, Fan H, Wu Y, Xie S, Girshick R (2021) Masked autoencoders are scalable vision learners. arXiv preprint arXiv: 2111.06377.
14. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations ICLR 2015.
15. Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: Unanswerable questions for SQuAD. arXiv https://arxiv.org/abs/1806.03822.
16. Adiwardana D, Luong MT, So DR, Hall J, Fiedel N, et al. (2020) Towards a human-like open-domain chatbot. arXiv preprint 1: 1-38.
17. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, et al. (2018) Deep contextualized word representations. arXiv https://arxiv.org/abs/1802.05365.
18. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 4171-4186.
19. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, et al. (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research

21: 1-67.

20. Bender EM, Koller A (2020) Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 5185-5198.
21. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition 3156-3164.
22. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, et al. (2015) Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision pp 2425-2433.
23. Das A, Kottur S, Gupta K, Singh A, Yadav D, et al. (2017) Visual dialog. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 326-335.
24. Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, et al (2015) Sequence to sequence-video to text. In Proceedings of the IEEE international conference on computer vision. 4534-4542.
25. Das A, Datta S, Gkioxari G, Lee S, Parikh D, et al. (2018) Embodied Question Answering. In CVPR 1-10.
26. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, et al. (2015) Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision 2641-2649.
27. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems 32.
28. Tan H, Bansal M (2019) Lxmert: Learning cross-modality encoder representations from transformers. arXiv https://arxiv.org/abs/1908.07490.
29. Su W, Zhu X, Cao Y, Li B, Lu L, et al. (2019) Vl-bert: Pre-training of generic visual-linguistic representations. arXiv https://arxiv.org/abs/1908.08530.
30. Chen YC, Li L, Yu L, Kholy AE, Ahmed F, et al. (2020) Uniter: Universal image-text representation learning. In European Conference on Computer, Vision Springer Cham 104-120.
31. Li X, Yin X, Li C, Zhang P, Hu X, et al. (2020) Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision Springer, Cham 121-137.
32. Agrawal A, Lu J, Antol S, Mitchell M, Zitnick CL, et al. (2016) Analyzing the behavior of visual question answering models. arXiv preprint 13.
33. Kiela D, Divvala S, Giryes R, Singh A (2020) Opportunities in visual semantics.
34. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. (2021) Learning transferable visual models from natural language supervision. In International Conference on Machine Learning PMLR 8748-8763.
35. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. (2021) Learning transferable visual models from natural language supervision. In International Conference on Machine Learning PMLR 8748-8763.
36. Alayrac JB, Recasens A, Schneider R, Arandjelović R, Ramapuram J, et al. (2022) Flava: A foundational language and vision alignment model. arXiv preprint 2204.02967.
37. Shridhar M, Thomason J, Gordon D, Bisk Y, Han W, et al. (2020) ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 10740-10749.