

Multimodal Artificial Intelligence for Integrating Textual and Visual Information in Automated Classification

Dr. Faith Achieng

Assistant Professor, Institute of Machine Learning Research, Kenya

Abstract

Multimodal artificial intelligence has become an important area of computational research because contemporary information systems frequently contain multiple forms of data describing the same entities or events. Textual descriptions provide semantic information through words and linguistic relationships, whereas images represent objects, visual characteristics, and spatial arrangements. Conventional classification systems commonly process these information sources independently, potentially overlooking complementary relationships between language and visual content. Integrating textual and visual representations offers opportunities to develop classification systems capable of examining information that may be incomplete or ambiguous when considered through a single modality. The present paper develops a hypothetical methodological framework for investigating multimodal artificial intelligence in automated classification. The proposed approach combines natural language processing, convolutional neural networks, feature representation, and multimodal fusion. Four classification configurations are considered: text-only classification, image-only classification, early feature fusion, and late decision fusion. A fictional digital product catalog containing 3,600 paired textual descriptions and product images is introduced to demonstrate the intended research design. The proposed classification categories include apparel, electronic products, and home-related products.

The methodology incorporates textual preprocessing, visual feature extraction, training and validation procedures, modality integration, and comparative classification assessment. Two illustrative analytical tables describe the hypothetical dataset distribution and performance differences among the selected classification configurations. Classification accuracy and macro-averaged F1-score are considered primary evaluation measures. All numerical values are manually constructed and do not represent completed experiments, actual image processing, or verified classification outcomes. The methodological discussion examines cross-modal alignment, information complementarity, feature dimensionality, computational complexity, and dataset reliability. The paper concludes that multimodal artificial intelligence offers a relevant methodological foundation for integrating textual and visual information, although genuine classification improvements must be established through reproducible model development and independently verified experimental evidence.

Keywords: Multimodal Artificial Intelligence, Text Classification, Visual Information, Automated Classification, Multimodal Fusion, Natural Language Processing, Computer Vision, Deep Learning, Feature Integration, Information Representation.

1. Introduction

Artificial intelligence has become an important component of modern information processing because computational systems increasingly encounter data represented through different formats and communication channels. Textual documents, digital images, structured records, and interactive media frequently describe overlapping aspects of individual entities or events. Conventional analytical systems often examine these information sources independently, thereby restricting their ability to incorporate complementary evidence. Multimodal artificial intelligence provides a methodological approach for processing information across several modalities and developing integrated representations. Such capabilities are relevant to automated classification problems involving digital catalogs, document repositories, media archives, and other information-intensive environments.

Textual information contains linguistic descriptions, contextual relationships, and semantic expressions that can contribute to automated classification. Natural language processing techniques transform written content into computational representations through tokenization, feature extraction, and statistical or neural modeling. However, textual descriptions may contain ambiguity, incomplete details, and expressions that require additional context. Images provide another source of information by representing physical characteristics, colors, shapes, textures, and spatial arrangements. Nevertheless, visual information may also be difficult to interpret when objects possess similar appearances or when relevant category distinctions depend on semantic characteristics not directly visible within an image.

Multimodal classification attempts to address selected limitations of individual information sources by combining textual and visual evidence. A product image may reveal physical characteristics, while its accompanying description provides information about intended use, material composition, or product category. When these sources contain complementary information, an integrated classifier may develop more informative representations than a system relying on one modality alone. However, successful integration requires careful consideration of feature compatibility, modality reliability, and the relationships between corresponding observations. Combining information indiscriminately can introduce unnecessary complexity and may reduce prediction reliability when the available sources contain inconsistent or misleading information.

Deep learning has expanded the methodological possibilities for multimodal classification through advances in visual feature extraction and distributed language representation. Convolutional neural networks can identify image characteristics, while neural language models can represent relationships among textual expressions. Multimodal architectures may combine these representations through feature concatenation, shared embedding spaces, or decision-level aggregation. These strategies differ in how they preserve information from individual modalities and how their predictions are integrated. Consequently, their comparative usefulness depends on dataset characteristics, the classification objective, computational resources, and the quality of cross-modal correspondence within the available observations.

The present paper examines multimodal artificial intelligence through a hypothetical comparative framework for automated classification involving paired textual and visual information. A fictional digital product catalog is introduced to illustrate how product descriptions and images could be combined for category prediction. Four classification approaches are considered: text-only modeling, image-only modeling, early feature fusion, and late decision fusion. The proposed methodology emphasizes data preparation, multimodal representation, independent testing, and performance

assessment. No actual images or textual records were processed, and the numerical examples are included exclusively to demonstrate the organization of future experimental research.

2. Objectives of the Study

The primary objective of this study is to examine the methodological potential of multimodal artificial intelligence for integrating textual and visual information in automated classification systems. The investigation focuses on information representation, fusion strategies, predictive accuracy, and model reliability. Three specific objectives establish the analytical direction of the proposed research framework.

2.1 Examination of Textual and Visual Information Representation

The first objective is to examine how textual descriptions and digital images can be represented through computational features suitable for automated classification. Textual information contains semantic relationships expressed through words, while images contain spatial and visual characteristics. The proposed study considers how natural language processing and convolutional feature extraction may transform these information sources into structured representations. Particular attention is given to the relationship between modality-specific features and the information required to distinguish observations belonging to different classification categories.

The investigation further aims to identify circumstances in which textual and visual representations provide complementary or overlapping information. A written description may distinguish product functionality, whereas an image may reveal physical characteristics associated with the same category. However, these relationships are not necessarily consistent across observations. Consequently, the proposed framework examines the need for suitable feature normalization, semantic alignment, and representation quality. This objective emphasizes the importance of preserving meaningful information from both modalities before combining them within a common classification procedure.

2.2 Comparative Evaluation of Multimodal Fusion Strategies

The second objective focuses on establishing a comparative framework for text-only, image-only, early-fusion, and late-fusion classification approaches. These configurations represent alternative strategies for processing modality-specific information and combining predictive evidence. Early fusion integrates representations before classification, whereas late fusion combines outputs from separately developed models. The proposed comparison examines how these approaches may influence classification accuracy and macro-averaged F1-score under consistent evaluation conditions involving paired textual and visual observations.

A further objective involves examining whether improved classification performance can be attributed to the availability of complementary information rather than additional computational complexity. Multimodal models may contain more parameters or employ more extensive feature representations than unimodal baselines. Consequently, a fair comparison should consider model configuration, validation procedures, and resource requirements. This objective emphasizes the importance of independent testing and consistent data partitioning before concluding that integrated textual and visual information provides meaningful predictive advantages over individual information sources.

2.3 Identification of Multimodal Classification Constraints

The third objective is to identify methodological challenges associated with integrating textual and visual information within automated classification systems. These include incomplete descriptions, inconsistent image quality, mismatched data pairs, redundant features, and unequal modality reliability. Such conditions may influence the relationships learned by classification algorithms and reduce predictive consistency. The proposed study examines how preprocessing, data verification, and suitable fusion mechanisms could address selected difficulties while maintaining meaningful correspondence between the textual and visual components of each observation.

The investigation also considers the computational and evaluation requirements associated with multimodal learning. Integrating multiple feature spaces may increase memory consumption, training complexity, and susceptibility to overfitting. Prediction reliability may also vary when one modality is missing or contains contradictory information. Consequently, this objective emphasizes transparent modeling procedures, appropriate validation, and the assessment of modality-specific contributions. These considerations provide a foundation for evaluating multimodal classification beyond a comparison of aggregate performance scores.

3. Review of Literature

3.1 Foundations of Textual and Visual Representation Learning

Smeulders et al. (2000) examined content-based image retrieval and discussed the relationship between visual features and semantic interpretation. Krizhevsky et al. (2012) demonstrated the effectiveness of deep convolutional architectures for image recognition. Pennington et al. (2014) introduced GloVe as an approach for learning distributed word representations from linguistic information. Kim (2014) investigated convolutional neural networks for sentence classification. These contributions established important foundations for developing computational representations of visual and textual information through distinct but potentially complementary learning procedures.

LeCun et al. (2015) provided a broader explanation of deep learning principles and hierarchical representation development. Simonyan and Zisserman (2015) examined deep convolutional architectures for image recognition, while He et al. (2016) introduced residual learning for increasingly deep visual networks. Socher et al. (2013) investigated semantic composition through recursive neural representations. Together, these studies demonstrate how textual and visual information can be transformed into learned feature spaces, providing methodological foundations for subsequent attempts to integrate different modalities.

3.2 Development of Multimodal and Cross-Modal Learning

Ngiam et al. (2011) investigated multimodal deep learning and demonstrated approaches for learning representations from multiple information sources. Srivastava and Salakhutdinov (2012) examined multimodal learning through deep Boltzmann machines and investigated shared representations of textual and visual information. Frome et al. (2013) developed visual-semantic embeddings connecting image recognition with linguistic information. These contributions established important approaches for combining different information types and demonstrated that cross-modal representation learning can extend beyond independent processing of individual modalities.

Karpathy et al. (2014) examined fragment-based image-sentence embeddings and investigated relationships between visual regions and linguistic descriptions. Gong et al. (2014) studied image-

sentence representations using weakly annotated photographic collections. Karpathy and Fei-Fei (2015) investigated visual-semantic alignment for generating image descriptions. Vinyals et al. (2015) explored neural image caption generation, while Antol et al. (2015) introduced visual question answering as a multimodal reasoning task. These contributions expanded the methodological possibilities for aligning language and visual information.

3.3 Dataset Construction and Multimodal Evaluation Reliability

Russell et al. (2008) developed LabelMe as a resource for image annotation and visual recognition research. Torralba and Efros (2011) investigated dataset bias and demonstrated the importance of evaluating visual recognition systems across different data sources. Hodosh et al. (2013) examined image-sentence relationships through ranking-based evaluation. These studies highlighted the importance of annotation quality, dataset characteristics, and appropriate performance assessment. Their contributions are relevant to multimodal classification because unreliable data correspondence can influence both training and evaluation.

Lin et al. (2014) introduced the Microsoft COCO dataset for object recognition within complex visual contexts. Cawley and Talbot (2010) investigated model-selection bias and its influence on performance estimates. Snoek et al. (2012) examined Bayesian optimization for machine learning configuration. These contributions demonstrate the importance of structured datasets, transparent optimization, and independent evaluation. Their methodological implications extend to multimodal systems because increasing feature complexity can introduce additional opportunities for overfitting and misleading comparisons.

4. Research Methodology

4.1 Research Design and Multimodal Classification Framework

The present paper adopts a hypothetical comparative quantitative research design for investigating automated classification using paired textual and visual information. Four classification configurations are considered: a text-only Support Vector Machine, an image-only Convolutional Neural Network, an early-fusion model, and a late-fusion model. The proposed classification problem involves three digital product categories. Each record would contain one product image and an accompanying textual description. This framework has not been implemented experimentally and is intended exclusively for methodological illustration.

The proposed analytical procedure consists of dataset preparation, textual preprocessing, image transformation, feature extraction, multimodal integration, model optimization, and independent classification assessment. The unimodal approaches would provide separate text and image baselines. The early-fusion configuration would combine text and image representations before classification, whereas late fusion would integrate separately generated prediction probabilities. All modeling decisions would rely on development information. The independent test dataset would remain unavailable during parameter optimization and fusion-weight selection.

4.2 Hypothetical Dataset and Modality-Specific Preprocessing

For illustrative purposes, the study assumes a fictional dataset containing 3,600 paired product records. Each record includes a short product description, a corresponding image, and one predefined category label. The proposed categories are apparel, electronics, and home-related products, with 1,200

observations assigned to each category. These records are manually specified as part of the study design. No actual product descriptions, image files, commercial databases, or computationally generated sample records were collected or processed.

The proposed textual preprocessing procedure would include tokenization, consistent handling of punctuation, and construction of relevant word-based features. Image preparation would involve standardized dimensions, pixel normalization, and training-only augmentation where appropriate. A text classifier would use TF-IDF representations, while the visual classifier would employ convolutional features. For early fusion, learned textual representations and visual features would be normalized before concatenation. Product-level data partitioning would ensure that closely related image–description pairs did not appear across different evaluation subsets.

4.3 Fusion Architecture and Classification Evaluation

The text-only configuration would use a linear Support Vector Machine trained on TF-IDF features. The image-only configuration would employ a Convolutional Neural Network for extracting visual representations. The early-fusion model would combine text embeddings and image features before applying a shared classification layer. The late-fusion model would combine calibrated prediction probabilities from independently trained modality-specific classifiers. Fusion parameters and classification settings would be selected using validation information without access to independent testing outcomes.

The proposed evaluation would divide the hypothetical dataset into 70% training, 15% validation, and 15% independent testing observations. Stratified partitioning would preserve the intended category proportions across the three subsets. Classification accuracy and macro-averaged F1-score would serve as principal evaluation measures. A completed investigation should also examine class-specific errors, computational requirements, and sensitivity to incomplete modalities. Repeated testing and uncertainty estimates would be necessary before establishing meaningful differences among the selected classification configurations.

5. Analysis

5.1 Illustrative Distribution of Paired Textual and Visual Records

The distribution of classification categories represents an important consideration in multimodal learning because unequal category frequencies can influence training and performance interpretation. Table 1 presents the hypothetical allocation of paired textual and visual records across three digital product categories. Each record is assumed to contain a corresponding description and image. The fictional dataset is balanced across classes, providing a simplified foundation for comparative evaluation. These records do not represent actual observations collected from a digital commerce platform.

The illustrative distribution demonstrates how category proportions could be maintained across training, validation, and independent testing subsets. Balanced allocation reduces certain difficulties associated with class imbalance but does not guarantee representative or reliable learning. Differences in image quality, description length, and semantic consistency may still influence classification outcomes. Consequently, a completed investigation would require additional examination of dataset diversity and modality correspondence. The proposed table provides a methodological example rather than a description of actual product information.

Table 1: Hypothetical Distribution of Paired Textual and Visual Classification Records

Product Classification Category	Training Records	Validation Records	Test Records	Total
Apparel Products	840	180	180	1,200
Electronic Products	840	180	180	1,200
Home-Related Products	840	180	180	1,200
Total	2,520	540	540	3,600

The illustrative dataset contains equal class proportions, with 33.33% of observations assigned to each category. Its purpose is to demonstrate consistent data partitioning and the organization of modality-paired observations.

5.2 Illustrative Comparison of Unimodal and Multimodal Classification

Automated classification performance can be examined by comparing predicted categories with established reference labels. Text-only and image-only models provide useful baselines for identifying the contribution of individual information sources. Multimodal approaches introduce additional information through feature or decision integration. Table 2 presents hypothetical classification outcomes for four configurations. The constructed independent test set contains 540 observations. The numerical values have been manually selected and do not represent outputs produced by actual machine learning algorithms.

The hypothetical comparison assigns higher illustrative accuracy to the multimodal configurations, with late decision fusion receiving the largest numerical value. This example demonstrates how complementary textual and visual information might be assessed through comparative classification. However, the assigned differences do not establish that multimodal integration improves actual predictive performance. Genuine evaluation would require implemented models, representative paired observations, reproducible optimization, and independent statistical testing. Computational costs and the contribution of each modality would also require assessment.

Table 2: Hypothetical Comparative Classification Performance of Multimodal AI Models

Classification Configuration	Correct Predictions	Incorrect Predictions	Accuracy (%)	Macro-F1
Text-Only SVM	420	120	77.78	0.778
Image-Only CNN	441	99	81.67	0.817
Early Feature Fusion	468	72	86.67	0.867
Late Decision Fusion	477	63	88.33	0.883

The illustrative pattern provides a basis for examining whether integrating two modalities could contribute additional discriminative information. However, meaningful conclusions would require evaluation of model stability, prediction errors, and the effects of contradictory textual and visual inputs.

6. Discussion

6.1 Complementary Information and Multimodal Representation

Multimodal artificial intelligence provides a methodological approach for combining information that may be distributed across different representations. Textual descriptions can provide semantic details, while images contribute visual characteristics associated with the same observation. When these sources contain complementary information, their integration may help distinguish categories that are difficult to identify using one modality alone. However, multimodal effectiveness depends on whether the available information sources are accurate, relevant, and appropriately aligned with the intended classification outcome.

The proposed framework illustrates this relationship through paired product descriptions and images. A visual representation may reveal an object's appearance, whereas accompanying text can describe features that are difficult to recognize visually. Nevertheless, the usefulness of integration varies according to the classification problem. Some categories may be distinguished adequately through one modality. Consequently, the proposed comparison emphasizes the need to establish whether multimodal information provides measurable benefits beyond the individual contributions of textual and visual representations.

6.2 Early and Late Fusion in Automated Classification

Early fusion integrates textual and visual features before the classification decision is produced. This approach allows the predictive model to learn relationships among modality-specific representations within a common feature space. However, concatenating features can increase dimensionality and computational requirements. The effectiveness of early fusion may also depend on suitable normalization and representation quality. Without these safeguards, one modality may dominate the combined feature space or introduce information that does not contribute meaningfully to classification.

Late fusion combines predictions generated by independently developed modality-specific classifiers. This strategy preserves separate modeling procedures and allows the contributions of different information sources to be adjusted during decision aggregation. However, reliable fusion requires appropriately calibrated probabilities and validation-based parameter selection. The hypothetical results assign favorable performance to late fusion, but they do not demonstrate empirical superiority. Future experiments would need to compare both fusion approaches under consistent computational conditions and examine their sensitivity to modality-specific errors.

6.3 Classification Reliability and Cross-Modal Generalization

Multimodal classification introduces evaluation challenges because textual and visual records may contain correlated information or repeated representations of the same entity. If related product images or descriptions appear across training and testing datasets, performance estimates may become misleading. Consequently, independent data partitioning must consider the relationships among multimodal records. The proposed framework emphasizes product-level separation and the prevention of

information leakage during representation learning, parameter optimization, and final classification evaluation.

Generalization also depends on whether the model remains reliable when textual descriptions or visual characteristics differ from those represented during training. A classifier developed using standardized product images may respond differently to photographs containing unusual backgrounds or lighting conditions. Similarly, abbreviated descriptions and unfamiliar terminology may influence textual predictions. Therefore, future research should investigate performance across varied image qualities, writing styles, and incomplete modality conditions. The present hypothetical results cannot establish such generalization properties.

7. Challenges

7.1 Cross-Modal Alignment and Information Consistency

Multimodal classification requires an appropriate relationship between textual descriptions and their corresponding visual observations. Errors in data pairing can introduce misleading associations and influence the representations learned by computational models. A product description may refer to characteristics that are not visible within its image, while an image may contain objects unrelated to the primary classification target. Consequently, the proposed framework emphasizes verification of modality correspondence and the identification of inconsistencies before model development.

Semantic alignment introduces additional difficulties because textual and visual information describe entities through different representational structures. Words express conceptual relationships, whereas images contain spatial and perceptual patterns. A common feature space does not automatically preserve every relevant relationship between these sources. Consequently, multimodal integration requires suitable representations and appropriately designed fusion procedures. Future experiments should examine how errors in one modality influence combined predictions and whether the classifier can maintain reliability when the sources provide conflicting information.

7.2 Computational Complexity and Feature Dimensionality

Multimodal learning can introduce substantial computational requirements because separate feature extraction procedures may be necessary for textual and visual information. Image encoders frequently process large feature representations, while language models introduce additional parameters and sequence-processing requirements. Combining these outputs may increase memory consumption and model complexity. Consequently, computational efficiency must be considered alongside classification performance. More extensive architectures do not necessarily provide sufficient improvement to justify their additional training and inference requirements.

Feature dimensionality presents another challenge because the combined representation may contain redundant or weakly informative information. A large feature space can increase the risk of overfitting when the available dataset contains relatively few observations. Appropriate feature normalization, dimensionality control, and regularization may help address these difficulties. However, excessive reduction can remove important cross-modal relationships. The proposed framework therefore emphasizes careful feature integration and the evaluation of computational costs across the complete classification pipeline.

7.3 Dataset Quality and Missing Modality Information

Multimodal datasets may contain incomplete textual descriptions, low-resolution images, inconsistent category labels, and repeated observations. These characteristics can influence classification reliability even when sophisticated modeling techniques are available. Annotation quality is particularly important because models learn relationships between paired information and predefined categories. Inconsistent labels may introduce contradictory training signals. Consequently, dataset construction requires careful verification of category definitions, modality correspondence, duplicate records, and the diversity of represented observations.

Missing modality information creates additional difficulties when some records contain only text or only images. A classifier designed exclusively for complete image–text pairs may be unable to process such observations without modification. Methods for handling unavailable modalities may include modality-specific fallback predictions or representations designed to tolerate incomplete inputs. However, their effectiveness requires independent evaluation. The proposed framework considers complete pairs for methodological simplicity and does not establish performance under missing-modality conditions. Future research should examine these limitations through controlled experiments.

8. Conclusion

Multimodal artificial intelligence offers a methodological foundation for integrating textual and visual information within automated classification systems. The present paper developed a hypothetical comparative framework incorporating text-only classification, image-only classification, early feature fusion, and late decision fusion. Its principal focus was the organization of modality-specific representations, classification procedures, and independent evaluation. The proposed approach demonstrates how complementary information sources could be examined within a common computational framework.

The illustrative dataset and performance tables demonstrate possible methods for reporting classification distributions and comparative predictive outcomes. However, all numerical values were manually constructed and do not represent completed experiments or verified classification results. Establishing genuine improvements would require representative paired datasets, implemented learning algorithms, independent testing, and appropriate statistical analysis. Consequently, the hypothetical comparisons cannot establish the superiority of multimodal fusion over individual textual or visual classification approaches.

Overall, the study identifies cross-modal alignment, feature dimensionality, computational efficiency, dataset reliability, and model generalization as important considerations in multimodal classification. Future empirical investigations should examine these factors through reproducible experimental procedures and varied information environments. Multimodal systems may provide useful capabilities when their component representations contribute complementary and reliable information. Nevertheless, meaningful conclusions regarding classification effectiveness must be supported by independently verified experimental evidence rather than illustrative numerical comparisons.

References

1. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>

2. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Cornia, M., & Cucchiara, R. (2024). The revolution of multimodal large language models: A survey. Findings of the Association for Computational Linguistics: ACL 2024, 13590–13618. <https://doi.org/10.18653/v1/2024.findings-acl.807>
3. Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: Universal image-text representation learning. In Computer Vision – ECCV 2020 (pp. 104–120). Springer. https://doi.org/10.1007/978-3-030-58577-8_7
4. Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). DeViSE: A deep visual-semantic embedding model. In Advances in Neural Information Processing Systems (Vol. 26). Curran Associates.
5. Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 4904–4916). PMLR. <https://proceedings.mlr.press/v139/jia21b.html>
6. Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 5583–5594). PMLR.
7. Kiela, D., Bhooshan, S., Firooz, H., Testuggine, D., & others. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. In Advances in Neural Information Processing Systems (Vol. 33).
8. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning (Vol. 162, pp. 12888–12900). PMLR.
9. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 19730–19742). PMLR. <https://proceedings.mlr.press/v202/li23q.html>
10. Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2020). VisualBERT: A simple and performant baseline for vision and language. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 3558–3568). Association for Computational Linguistics.
11. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In Computer Vision – ECCV 2020 (pp. 121–137). Springer. https://doi.org/10.1007/978-3-030-58577-8_8
12. Liang, P. P., Liu, Z., Zadeh, A., & Morency, L.-P. (2018). Multimodal language analysis with recurrent multistage fusion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 150–161). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1014>
13. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems (Vol. 32).

14. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 8748–8763). PMLR. <https://proceedings.mlr.press/v139/radford21a.html>
15. Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 5100–5111). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1514>
16. Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 6558–6569). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1656>
17. Vatsa, M., Jain, A., & Singh, R. (2024). Adventures of trustworthy vision-language models: A survey. Proceedings of the AAAI Conference on Artificial Intelligence, 38(20), 22650–22658. <https://doi.org/10.1609/aaai.v38i20.30275>
18. Wang, J., Hu, X., Gan, Z., Yang, Z., Dai, X., Wang, X., Li, J., & Liu, J. (2022). OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Proceedings of the 39th International Conference on Machine Learning (Vol. 162, pp. 23318–23340). PMLR.
19. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. National Science Review, 11(12), Article nwae403. <https://doi.org/10.1093/nsr/nwae403>
20. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1103–1114). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1115>