Adaptive Transformer with Sequence-Guided Decoders for Enhanced Vision Captioning

Chloe Mitchell , Lobry Hsu , Diego Fernández *

Posted Date: 21 January 2025

doi: 10.20944/preprints202501.1501.v1

Keywords: Image Captioning, Transformer Architecture, Geometry-Aware Attention, Sequence Modeling, LSTMs

 doi:10.20944/preprints202501.1501.v1

Article

Adaptive Transformer with Sequence-Guided Decoders for Enhanced Vision Captioning

Chloe Mitchell, Lobry Hsu and Diego Fernández *

Bond University
* Correspondence: diego.fer@bond.edu.au

Abstract: In recent years, Transformer architectures have been extensively applied to image captioning, achieving remarkable performance. The spatial and positional relationships between visual objects play a pivotal role in crafting meaningful and accurate captions. To further enhance image captioning with Transformers, this paper introduces the *Adaptive Geometry-Integrated Transformer* (AGIT). This novel model incorporates advanced geometry-aware mechanisms into both its encoder and decoder, enabling superior representation and utilization of spatial information. Specifically, the proposed framework comprises two key components: i) a geometry-enhanced self-attention module, termed the *Geometry Attention Refiner* (GAR), which explicitly integrates relative spatial relationships into the visual feature representations during encoding; and ii) a sequence-guided decoding mechanism powered by *Position-Sensitive LSTMs* (PS-LSTMs) to accurately model and maintain word-order semantics while generating captions. Experimental evaluations on the MS COCO and Flickr30k datasets demonstrate that AGIT outperforms state-of-the-art models in both accuracy and computational efficiency, setting a new benchmark in image captioning.

Keywords: image captioning; Transformer architecture; geometry-aware attention; sequence modeling; LSTMs

1. Introduction

Image captioning, a critical task in computer vision [1], involves generating coherent and descriptive textual summaries for images. This process requires not only recognizing the visual elements within an image but also understanding their spatial arrangements, interactions, and contextual relationships. By bridging the gap between visual content and natural language, image captioning has wide-ranging applications, including assisting visually impaired individuals, enhancing image retrieval systems, and improving multimedia content accessibility [2].

Historically, early image captioning approaches were predominantly categorized into retrieval-based and template-based methods [7]. Retrieval-based methods relied on matching input images to existing datasets of captions through similarity metrics [8,9]. These approaches, though efficient, often struggled to generate novel captions and heavily depended on the quality of the dataset. Template-based approaches, on the other hand, utilized predefined syntactic templates combined with visual concept detections to construct captions [10–12]. While these methods provided structured outputs, their reliance on handcrafted rules and feature engineering limited flexibility and scalability.

With advancements in artificial intelligence, neural network-based methods have emerged as a transformative force in image captioning [13–15]. The introduction of encoder-decoder architectures [18] marked a significant leap, enabling the generation of captions by mapping image features to textual sequences. Subsequent innovations, such as multimodal learning [16,17], attention mechanisms [19,20], and compositional architectures [21], have further enhanced the capability of neural models to generate contextually accurate and semantically rich descriptions. Attention mechanisms, in particular, allow models to focus on specific image regions while generating captions, mimicking human attention patterns [26].

 doi:10.20944/preprints202501.1501.v1

Transformers, characterized by their self-attention mechanisms and parallel processing capabilities, have become a dominant framework in image captioning [27,28]. A standard Transformer-based model processes visual features extracted from Convolutional Neural Networks (CNNs) through an encoder, which maps these features into intermediate representations. The decoder, often leveraging Recurrent Neural Networks (RNNs) or alternative architectures, then generates captions sequentially [4,16,18]. This architecture has demonstrated remarkable success due to its ability to model long-range dependencies and efficiently process multimodal data.

Despite these advancements, one critical aspect often overlooked in existing models is the explicit representation and utilization of geometric and positional relationships between objects in an image. For instance, captions such as "a boy standing on a skateboard" and "a boy holding a skateboard" convey different meanings, yet without spatial awareness, models may fail to distinguish between them. While some methods incorporate geometric information at the encoder level, such as the Geometry Self-Attention (GSA) proposed in [24], they often neglect detailed positional reasoning during decoding.

To address these limitations, this paper introduces the *Adaptive Geometry-Integrated Transformer* (AGIT), which integrates geometric and positional reasoning at both encoding and decoding stages. The proposed framework leverages a *Geometry Attention Refiner* (GAR) to enhance the encoder's ability to capture spatial relationships explicitly. Additionally, it incorporates *Position-Sensitive LSTMs* (PS-LSTMs) in the decoder, enabling precise word-order modeling and semantic consistency in generated captions. These innovations ensure that AGIT not only captures the "where" and "how" of visual elements but also translates these insights into grammatically and contextually accurate textual descriptions.

2. Related Work

Recent advancements in encoder-decoder Transformer architectures, particularly those enhanced with attention mechanisms, have sparked significant interest in the field of image captioning. Numerous methodologies have been proposed to address the challenges of generating coherent and descriptive captions for images, including approaches that utilize soft and hard attention mechanisms [29], residual connections [30], and the meshed-memory Transformer [31].

2.1. Advancements in Transformer Architectures

Transformer architectures [24,27,32,33] have established themselves as a dominant framework in image captioning due to their flexibility and effectiveness. Typically, these architectures comprise an encoder and a decoder. The encoder extracts features from image regions, while the decoder generates textual descriptions based on these features. Both components employ multi-layer residual networks, which improve gradient flow and enable deeper models. A key feature of these architectures is the self-attention mechanism, which captures long-range dependencies and contextual relationships among input data [27,28].

A noteworthy enhancement in this framework involves the integration of geometric cues. For example, additional features such as object centers, sizes, and spatial relationships [24] have been incorporated into encoder inputs to provide richer contextual information. Positional encoding techniques, such as sinusoidal functions, have further improved the sequential nature of generated text by maintaining word order consistency [26]. These innovations have significantly improved the quality of generated captions.

2.2. Enhancements in Attention Mechanisms

The attention mechanism plays a pivotal role in image captioning, enabling models to focus on relevant parts of an image while generating captions. Beyond the basic attention mechanism [29], various extensions have been developed to address specific limitations. Multi-head attention [26] allows the model to attend to different parts of an image simultaneously, providing diverse

 doi:10.20944/preprints202501.1501.v1

perspectives. Gate-controlled attention [21] selectively filters information, enhancing the alignment between image regions and generated words.

Other notable attention mechanisms include fully attentive paradigms [28,34], which leverage comprehensive attention across all input tokens, and meshed-connection attention [31], which establishes dense connections between layers to improve information flow. Dual attention mechanisms [35] combine spatial and semantic attention, offering a more holistic approach to aligning image features with textual outputs. These advancements have contributed to the robustness and versatility of image captioning systems.

2.3. Future Directions in Image Captioning

Despite the significant progress achieved with Transformer-based architectures, there remains ample scope for further innovation. Future research could focus on designing more sophisticated network architectures that better capture the hierarchical relationships within images. For example, leveraging graph neural networks (GNNs) to model the structural relationships between objects could enhance spatial reasoning capabilities.

In addition, integrating richer feature representations, such as high-resolution spatial details and temporal dynamics, could further improve caption quality. Advanced geometric encoding strategies that incorporate not only relative positions but also three-dimensional spatial information could be explored. Similarly, more dynamic positional encoding methods that adapt to varying text lengths and content could enhance decoder performance.

Lastly, the incorporation of multi-modal learning techniques, which combine visual, textual, and auditory data, represents a promising avenue for expanding the capabilities of image captioning systems [36]. By embracing these directions, the field of image captioning can continue to evolve, addressing increasingly complex tasks and real-world applications.

3. Methodology

Our proposed method, *Adaptive Geometry-Integrated Transformer* (AGIT), generates grounded captions by dynamically attending to specific image regions at each step. It retains the foundational encoder-decoder architecture while introducing novel enhancements to improve performance. In the encoder, a Geometry Self-Attention Refiner (GSR) optimizes image representations by integrating spatial relationships. The decoder employs a Position-Aware Self-Attention mechanism to generate word sequences that accurately reflect the contextual and spatial intricacies of the image.

Task Definition.

Given an input image $\boldsymbol{I}$, we represent its appearance features as $(\boldsymbol{X}_A \in \mathbb{R}^{N\times d})$, where N is the number of image regions, and d is the feature dimension. The corresponding caption sequence is denoted as $(\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\})$, with T indicating the number of words.

3.1. Geometry Self-Attention Refiner

To enhance the encoding of spatial relationships, the Geometry Self-Attention Refiner incorporates object geometry into the traditional self-attention mechanism. This integration ensures that both visual and spatial information contribute to feature representation.

Incorporating Geometry Features. The geometry features of each object are represented as $\boldsymbol{X}_g \in \mathbb{R}^{N\times 5}$, with each row containing:

$$(x_{min}, y_{min}, x_{max}, y_{max}, S) \tag{1}$$

 doi:10.20944/preprints202501.1501.v1

where (x_{min}, y_{min}) and (x_{max}, y_{max}) are the coordinates of the bounding box, and S represents the normalized area of the bounding box. These features are embedded into a higher-dimensional space through a non-linear transformation:

$$\boldsymbol{X}_G = \text{ReLU}(\boldsymbol{X}_g \boldsymbol{W}_G + \boldsymbol{b}_G) \tag{2}$$

where $\boldsymbol{W}_G$ and $\boldsymbol{b}_G$ are learnable parameters.

Combining Appearance and Geometry. The appearance features $\boldsymbol{X}_A$ and geometry features $\boldsymbol{X}_G$ are combined to compute enhanced queries and keys:

$$\boldsymbol{Q}' = [\boldsymbol{X}_A \boldsymbol{W}_{Q_A}; \boldsymbol{X}_G \boldsymbol{W}_{Q_G}] \tag{3}$$
$$\boldsymbol{K}' = [\boldsymbol{X}_A \boldsymbol{W}_{K_A}; \boldsymbol{X}_G \boldsymbol{W}_{K_G}] \tag{4}$$

where $\boldsymbol{W}_{Q_A}$, $\boldsymbol{W}_{Q_G}$, $\boldsymbol{W}_{K_A}$, and $\boldsymbol{W}_{K_G}$ are learnable weights. The attention scores are calculated as:

$$\boldsymbol{\Omega}' = \frac{\boldsymbol{Q}' \boldsymbol{K}'^T}{\sqrt{d}} \tag{5}$$

The refined attention output is given by:

$$\text{Attention}_G(\boldsymbol{X}) = \text{softmax}(\boldsymbol{\Omega}')\boldsymbol{V}_A \tag{6}$$

Gate-Controlled Refinement. To further refine the attention output, a Gate-Controlled Linear Unit (GLU) modulates the attention features:

$$\boldsymbol{G} = \sigma(\boldsymbol{W}_G \boldsymbol{X}_A + \boldsymbol{b}_G) \tag{7}$$
$$\boldsymbol{Output} = \boldsymbol{G} \odot \boldsymbol{Attention}_G(\boldsymbol{X}) \tag{8}$$

where σ denotes the sigmoid function, and $\odot$ represents the element-wise product.

3.2. Position-Aware Decoder

The decoder integrates positional encoding into the self-attention mechanism, ensuring the sequential coherence of generated captions.

LSTM-Based Positional Encoding. At each time step t, the input to the LSTM includes the word embedding $\boldsymbol{w}_t$ and the pooled visual features $\boldsymbol{v}$:

$$\boldsymbol{x}_t = [\boldsymbol{w}_t; \boldsymbol{v}], \quad \boldsymbol{v} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_A^{(i)} \tag{9}$$

The LSTM updates its hidden state as:

$$\boldsymbol{h}_t, \boldsymbol{c}_t = \text{LSTM}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{c}_{t-1}) \tag{10}$$

Decoder Self-Attention. Using the refined features $\boldsymbol{X}^r$ and the LSTM encoding $\boldsymbol{h}_t$, the decoder computes word distributions as:

$$P(\boldsymbol{y}_t | \boldsymbol{y}_{1:t-1}) = \text{softmax}(\boldsymbol{W}_o \boldsymbol{h}_t + \boldsymbol{b}_o) \tag{11}$$

where $\boldsymbol{W}_o$ and $\boldsymbol{b}_o$ are learnable parameters.

By integrating geometry-aware attention and position-sensitive decoding, AGIT achieves superior captioning performance, as demonstrated in our experiments.

 doi:10.20944/preprints202501.1501.v1

4. Experiments

4.1. Datasets and Metrics

We evaluate our proposed AGIT model using two widely-adopted datasets: MS COCO [37] and Flickr30k [38]. MS COCO is among the largest and most comprehensive datasets for image captioning, comprising 123,287 images with five human-annotated captions per image. To ensure consistency in evaluation, we use the well-known "Karpathy" split [16], which includes 113,287 images for training, 5000 for validation, and 5000 for testing. Similarly, Flickr30k consists of 31,783 images, each annotated with five captions. For both datasets, captions exceeding 16 words are truncated, and all sentences are converted to lowercase to establish experimental vocabularies of 9,487 words for MS COCO and 7,000 words for Flickr30k.

To comprehensively evaluate the performance of AGIT, we utilize five standard metrics: BLEU [39], METEOR [40], ROUGE [41], CIDEr [42], and SPICE [43]. These metrics collectively measure various aspects of the generated captions, including n-gram precision, semantic alignment, and overall coherence.

4.2. Implementation Details

Our implementation employs a pre-trained Faster R-CNN model with a ResNet-101 backbone [4] to extract 36 features for each image, with each feature vector being 2048-dimensional. To optimize memory usage, these features are projected into a 512-dimensional space. The LSTM hidden state size is set to 1024, while the input dimensions for the Geometry Self-Attention Refiner (GSR) and self-attention modules are both configured to 512. Furthermore, the multi-head self-attention mechanism utilizes eight heads, and both the encoder and decoder are composed of three layers.

Dropout regularization is applied during training, with a dropout rate of 0.5 for the LSTM and 0.1 for the self-attention layers. During the cross-entropy training phase, the learning rate is initialized to (5×10^{-4}) and decays by a factor of 0.8 every three epochs. For CIDEr optimization, the learning rate is set to (2×10^{-5}) and follows the same decay schedule. All models are trained using the Adam optimizer with a batch size of 50. For inference, we employ beam search with a beam size of five across all experiments.

4.3. Ablation Experiments

To evaluate the contributions of individual components within AGIT, we conduct a series of ablation studies on the MS COCO dataset. Our baseline is a *vanilla Transformer* model [24], which lacks geometric features and employs sinusoidal position encodings.

Effect of Geometry Self-Attention Refiner

To assess the impact of the GSR module, we integrate it into the baseline model. The GSR incorporates explicit geometric relationships into the encoder, refining raw image representations. As shown in Table 1, the CIDEr score improves significantly from 109.0 to 115.1. This enhancement demonstrates that without geometric relations, the baseline model is prone to misinterpreting irrelevant regions. The GSR effectively addresses this issue, enabling the model to identify precise spatial contexts and generate geometry-aware captions.

Table 1. Ablation study results on the MS COCO "Karpathy" test split [16].

Model	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Base	75.0	32.8	27.3	55.5	109.0	20.6
Base+GSR	76.9	35.6	28.1	57.0	115.1	21.4
Base+Position-LSTM	76.5	34.5	28.0	56.8	114.9	21.3
Full: AGIT	**77.5**	**37.8**	**28.5**	**57.6**	**119.8**	**21.8**

 doi:10.20944/preprints202501.1501.v1

Effect of Position-Aware LSTM Decoder

Next, we replace the sinusoidal position encoding of the baseline model with the position-aware LSTM decoder. This module enhances the decoder's capacity to incorporate sequential information. As indicated in Table 1, the CIDEr score increases by 5.9 points, underscoring the effectiveness of position-aware encoding in guiding the decoder to focus on semantically relevant regions.

Geometry Queries and Keys

To evaluate the efficiency of incorporating geometric features into queries and keys, we compare two strategies: addition and concatenation. Table 2 shows that concatenating appearance and geometric features yields superior performance. For instance, the BLEU-1 score increases from 75.0 (baseline) to 77.5 with concatenation, compared to 76.0 with addition.

Table 2. Comparison of strategies for combining geometric and appearance features.

Strategy	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Add	76.0	35.1	27.2	56.0	116.4	20.7
Concatenate	**77.5**	**37.8**	**28.5**	**57.6**	**119.8**	**21.8**

Role of Gated Linear Units (GLUs)

GLUs refine the outputs of self-attention layers by selectively emphasizing important features. Table 2 indicates that integrating GLUs into the encoder achieves the best results, while applying them to both encoder and decoder leads to diminishing returns. This observation aligns with findings in [20], highlighting the importance of balanced module integration.

4.4. Comparison with State-of-the-Art Models

We compare AGIT against state-of-the-art models on MS COCO and Flickr30k datasets. These comparisons include SCST [3], Up-Down [4], ORT [27], AoANet [20], and others. Table **??** illustrates that AGIT consistently outperforms these methods, achieving substantial improvements in BLEU-4, METEOR, and CIDEr scores.

4.5. Caption Text Comparisons

To qualitatively demonstrate the advantages of our proposed AGIT model, we present qualitative comparisons of captions generated by AGIT, the baseline 'Vanilla Transformer' model, and ground truth captions (GT1-GT3) for six selected images. It is evident that AGIT produces captions with a more comprehensive semantic structure and accurate positional relations. For instance, in the top-left image, AGIT correctly identifies that "people" are "under umbrellas" and "in front of the store," whereas the baseline fails to capture such nuanced spatial relationships. Similarly, for the middle-right image, AGIT appropriately describes the "dog laying under the chair" instead of incorrectly stating "laying on the chair".

The ability of AGIT to produce captions with precise geometric and positional relationships is primarily attributed to its Geometry Self-attention Refiner (GSR) module. This module explicitly integrates spatial correlations among image regions into the object feature representations. Furthermore, the position-LSTM plays a pivotal role in guiding the decoder to attend to relevant visual objects during each decoding step. Collectively, these enhancements enable AGIT to describe almost all key objects within an image, such as "people," "umbrellas," and "restaurant" in the aforementioned example, while also capturing their spatial interrelations. As a result, AGIT consistently generates reliable and spatially-aware captions that align closely with human interpretation.

 doi:10.20944/preprints202501.1501.v1

5. Conclusions

In this paper, we introduce the Attention-based Geometry-Integrated Transformer (AGIT), a significant extension of the Transformer architecture specifically designed for image captioning tasks. AGIT effectively refines image representations by embedding geometric features of visual objects into their respective region encodings. Additionally, the position-LSTM module enhances the decoder by accurately encoding the sequential order of caption words.

Our ablation experiments reveal the effectiveness of the GSR module for geometric feature integration and the position-LSTM for positional encoding. Each module independently contributes to substantial performance improvements when integrated with a baseline model. Furthermore, comparative experiments conducted both offline and online validate the superiority of the AGIT framework over state-of-the-art methods. AGIT consistently outperforms competitors on benchmark datasets such as MS COCO and Flickr30k, highlighting its robustness and reliability in generating spatially-aware and semantically rich image captions.

6. Future Work

Future work will focus on extending AGIT to handle multimodal inputs, such as integrating audio or temporal video data for broader applications like video summarization and scene understanding. Enhancing the geometric representation by incorporating 3D spatial information, such as depth or point clouds, is another promising direction, especially for domains requiring precise spatial reasoning like robotics and medical imaging. Efficiency improvements through techniques like model pruning or quantization will enable deployment on resource-constrained devices. Further, exploring self-supervised or semi-supervised learning paradigms could expand AGIT's applicability to scenarios with limited labeled data. Finally, improving explainability by visualizing attention mechanisms and providing user-friendly controls will increase trust and adaptability in critical applications.

References

1. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: European conference on computer vision, Springer, 2010, pp. 15–29.
2. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision 123 (1) (2017) 32–73.
3. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7008–7024.
4. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
5. M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10575–10584. http://doi.org/10.1109/CVPR42600.2020.01059.
6. J. Zhang, K. Li, Z. Wang, X. Zhao, Z. Wang, Visual enhanced glstm for image captioning, Expert Systems with Applications 184 (2021) 115462. https://doi.org/10.1016/j.eswa.2021.115462.
7. S. Bai, S. An, A survey on automatic image caption generation, Neurocomputing 311 (2018) 291–304.
8. V. Ordonez, G. Kulkarni, T. L. Berg, Im2text: Describing images using 1 million captioned photographs, in: Proceedings of the Advances in Neural Informa- tion Processing Systems (NIPS), 2011, pp. 1143–1151.
9. A. Gupta, Y. Verma, C. V. Jawahar, Choosing linguistics over vision to describe images, in: In Twenty-Sixth National Conference on Artificial Intelligence, 2012, pp. 606–612.
10. R. Socher, L. Fei-Fei, Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 966–973.

11. G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, T. L. Berg, Babytalk: Understanding and generating simple image descriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (12) (2013) 2891–2903. http://doi.org/10.1109/TPAMI.2012.162.
12. Y. Ushiku, M. Yamaguchi, Y. Mukuta, T. Harada, Common subspace for model and similarity: Phrase learning for caption generation from images, in: Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2668–2676. http://doi.org/10.1109/ICCV.2015.306.
13. A. Karpathy, A. Joulin, F. F. Li, Deep fragment embeddings for bidirectional image sentence mapping, in: Proceedings of the Twenty Seventh Advances in Neural Information Processing Systems (NIPS), Vol. 3, 2014, pp. 1889–1897.
14. L. Ma, Z. Lu, L. Shang, H. Li, Multimodal convolutional neural networks for matching image and sentence, in: Proceedings of IEEE International Conference on Computer Vision, 2015, pp. 2623–2631.
15. F. Yan, K. Mikolajczyk, Deep correlation for matching images and text, in: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3441–3450. http://doi.org/10.1109/CVPR.2015.7298966.
16. A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137.
17. A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (4) (2017) 664–676. http://doi.org/10.1109/TPAMI.2016.2598339.
18. O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
19. Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4651–4659. http://doi.org/10.1109/CVPR.2016.503.
20. L. Huang, W. Wang, J. Chen, X.-Y. Wei, Attention on attention for image captioning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4634–4643.
21. R. M. Oruganti, S. Sah, S. Pillai, R. Ptucha, Image description through fusion based recurrent multimodal learning, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3613–3617. http://doi.org/10.1109/ICIP.2016.7533033.
22. J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, A. L. Yuille, Learning like a child: Fast novel visual concept learning from sentence descriptions of images, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2533–2541. http://doi.org/10.1109/ICCV.2015.291.
23. M. Nabati, A. Behrad, Multimodal video-text matching using a deep bifurcation network and joint embedding of visual and textual features, Expert Systems with Applications 184 (2021) 115541, online. https://doi.org/10.1016/j.eswa.2021.115541.
24. L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, H. Lu, Normalized and geometry-aware self-attention network for image captioning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10324–10333. http://doi.org/10.1109/CVPR42600.2020.01034.
25. J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
26. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
27. S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image captioning: Transforming objects into words, in: Advances in Neural Information Processing Systems, 2019, pp. 11135–11145.
28. G. Li, L. Zhu, P. Liu, Y. Yang, Entangled transformer for image captioning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8928–8937.
29. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, 2015, pp. 2048–2057.
30. L. Gao, K. Fan, J. Song, X. Liu, X. Xu, H. T. Shen, Deliberate attention networks for image captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8320–8327.
31. M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10578–10587.

32. L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, J. Gao, Unified vision-language pre-training for image captioning and vqa., in: AAAI, 2020, pp. 13041–13049.
33. H. Chen, Y. Wang, X. Yang, J. Li, Captioning transformer with scene graph guiding, in: 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 2538–2542. http://doi.org/10.1109/ICIP42928.2021.9506193.
34. X. Zhu, L. Li, J. Liu, H. Peng, X. Niu, Captioning transformer with stacked attention modules, Applied Sciences 8 (5) (2018) 739.
35. L. Yu, J. Zhang, Q. Wu, Dual attention on pyramid feature maps for image captioning, IEEE Transactions on Multimedia (2021) 1–1http://doi.org/10.1109/TMM.2021.3072479.
36. S. Liu, Z. Ren, J. Yuan, Sibnet: Sibling convolutional encoder for video captioning, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (9) (2021) 3259–3272. http://doi.org/10.1109/TPAMI.2019.2940007.
37. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
38. P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78.
39. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.
40. S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
41. C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81.
42. R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
43. P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: ECCV, 2016.
44. W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, T. Zhang, Recurrent fusion network for image captioning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 499–515.
45. T. Yao, Y. Pan, Y. Li, T. Mei, Exploring visual relationship for image captioning, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 684–699.
46. X. Yang, K. Tang, H. Zhang, J. Cai, Auto-encoding scene graphs for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10685–10694.
47. J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks (m-rnn), arXiv preprint arXiv:1412.6632.
48. W. Cai, Q. Liu, Image captioning with semantic-enhanced features and extremely hard negative examples, Neurocomputing 413 (2020) 31–40. https://doi.org/10.1016/j.neucom.2020.06.112.
49. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. https://doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
50. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
51. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
52. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
53. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. https://doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
54. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
55. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.

56. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
57. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
58. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
59. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
60. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
61. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
62. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
63. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
64. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
65. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
66. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
67. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
68. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
69. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
70. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
71. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
72. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
73. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
74. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
75. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
76. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

77. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
78. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
79. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
80. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
81. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
82. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
83. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
84. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
85. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
86. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
87. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
88. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
89. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
90. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
91. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
92. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
93. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
94. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
95. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
96. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.

97. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
98. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
99. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
100. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
101. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.
102. Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. 2024. LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding Reasoning and Planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26428–26438.