Preprints.org

Article

Not peer-reviewed version

Rethinking Convolutional Semantics for Image Caption Generation Beyond Recurrent Paradigms

Noah Macdonald , Sofia Leblanc , Landen Whitaker , Arman Chowdhury *

Posted Date: 22 October 2025

doi: 10.20944/preprints202510.1744.v1

Keywords: image captioning; semantic alignment; neural attention

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Rethinking Convolutional Semantics for Image Caption Generation Beyond Recurrent Paradigms

Noah Macdonald, Sofia Leblanc, Landen Whitaker and Arman Chowdhury *

Dalhousie University, Canada
* Correspondence: achowdhury1@dal.ca

Abstract

The task of automatically generating natural language descriptions for images has become a cornerstone in bridging visual perception and linguistic understanding. While Recurrent Neural Networks (RNNs) and their variants such as LSTMs have long dominated the decoder component in image captioning systems, recent explorations suggest that Convolutional Neural Networks (CNNs) can serve as viable alternatives. However, the capability of CNN-based decoders to fully capture temporal and semantic dependencies in language has not been comprehensively assessed. In this paper, we introduce **VISCON** (Visual-Semantic Convolutional Network), a new convolutional decoder framework designed to investigate the strengths and weaknesses of CNN-based architectures in caption generation. Our study conducts a rigorous analysis across multiple dimensions, including network depth, convolutional filter complexity, integration of attention mechanisms, the role of sentence length in training, and the effectiveness of data augmentation strategies. Experiments are carried out on two widely adopted benchmarks, Flickr8k and Flickr30k, where we perform extensive comparisons with RNN-based decoders. Unlike conventional wisdom from recurrent models, our findings reveal that deeper convolutional stacks do not necessarily yield performance improvements, and the utility of visual attention is significantly less pronounced in convolutional decoding pipelines. Moreover, we observe that VISCON maintains competitive accuracy only when trained with relatively short captions, whereas performance degrades sharply as sentence length increases, indicating difficulty in modeling long-range dependencies. Finally, despite showing comparable BLEU and METEOR scores under certain settings, convolutional approaches consistently underperform on CIDEr, raising questions about their capacity to model human-like semantic richness. This comprehensive analysis highlights the underexplored trade-offs in convolutional decoding and contributes new insights into designing future captioning systems that harmonize visual-semantic reasoning with efficient sequence modeling.

Keywords: image captioning; semantic alignment; neural attention

---

1. Introduction

The problem of automatically describing images with coherent natural language sentences lies at the intersection of computer vision and natural language processing. The goal of an image captioning system is to generate fluent textual descriptions that not only identify objects in the scene but also express their relationships, contexts, and implied semantics [1]. Humans excel at this task by leveraging prior knowledge and commonsense reasoning—for instance, recognizing that a crowded stadium implies an ongoing sports event, or that the presence of a double-decker bus strongly suggests a location such as London. Replicating such nuanced reasoning in computational models remains a profound challenge.

Over the past decade, deep learning has spurred significant advances in image captioning. Early encoder-decoder approaches [2,3] borrowed ideas from neural machine translation, where an image is encoded into a vector representation by a Convolutional Neural Network [4,5] and decoded into a sentence by a Recurrent Neural Network (RNN) such as LSTM [6,15]. Attention mechanisms, initially

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

proposed for sequence-to-sequence translation [10], were quickly adapted to captioning [7–9,11], enabling models to dynamically focus on salient visual regions while generating words. These developments established RNNs as the de facto decoding architecture, demonstrating strong capacity to capture temporal dependencies in language.

Despite the dominance of recurrent models, researchers began to explore CNNs as alternatives for sequence modeling [20]. Aneja et al. [21] extended this paradigm to caption generation by using CNNs as decoders. The convolutional approach promises advantages such as parallel computation and reduced sequential dependencies, but it also raises questions about its ability to model long-range linguistic structures and contextual coherence. Unlike RNNs, which inherently model sequence order, CNNs depend on receptive field expansion to capture dependencies—a design that may or may not suffice for natural language descriptions of complex scenes.

While RNN-based decoders have been studied extensively with respect to network depth, regularization, attention design, and sentence length, CNN-based decoders remain relatively underexplored. Several critical aspects—including the effect of deeper convolutional stacks, the interplay between attention mechanisms and convolutions, and the impact of training caption length distribution—have not yet been systematically investigated. Moreover, performance trade-offs across evaluation metrics such as BLEU, METEOR, ROUGE, SPICE, and CIDEr are poorly understood in this context. These gaps hinder a comprehensive understanding of when convolutional decoders may offer advantages and where they fundamentally lag behind recurrent methods.

The rise of Transformer-based models has further reshaped the landscape of sequence modeling, demonstrating the power of self-attention in capturing long-range dependencies. However, CNNs retain practical value due to their efficiency and inductive biases toward locality and compositionality. In multimodal learning, hybrid architectures that blend CNNs, RNNs, and Transformers are increasingly common. Thus, studying CNN decoders for captioning is not only a comparative exercise but also a step toward designing hybrid models that exploit the complementary strengths of different architectures.

In this paper, we make the following contributions:

- We introduce **VISCON**, a convolutional decoder framework, to systematically examine CNN-based captioning across multiple dimensions: network depth, data augmentation, attention integration, and sentence length effects.
- We conduct extensive experiments on Flickr8k [22] and Flickr30k [23], providing a controlled comparison with recurrent baselines.
- We reveal unique limitations of CNN-based decoders, including their difficulty in handling longer sentences and their weaker response to attention mechanisms compared with RNN decoders [28,29].
- Our study offers the first detailed evaluation of CNN captioning models with respect to metric-specific performance trade-offs, demonstrating that while VISCON can approximate RNN baselines in some metrics, it consistently struggles on CIDEr, indicating limited ability to capture high-level semantic richness.

Overall, this work seeks to clarify the underexplored dynamics of convolutional decoders in image captioning and to provide actionable insights for building next-generation systems that combine efficiency with semantic depth.

2. Related Work

Research on Image Caption Generation has evolved over multiple paradigms, each reflecting different assumptions about how visual and linguistic modalities can be connected. Broadly, the methods can be grouped into three classical families—*Retrieval-based*, *Template-based*, and *Deep Learning-based* approaches—with more recent explorations combining ideas across these boundaries. In what follows, we provide an extensive discussion of these categories, highlighting representative works,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

their advantages and limitations, and the motivations that inspired the development of our proposed framework, **VISCON**.

2.1. Retrieval-Based Approaches

Early attempts treated caption generation not as a process of novel sentence synthesis, but as a problem of retrieving suitable descriptions from a pre-defined set. These *Retrieval-based* methods assume that if one can find an image in a large database that is visually similar to the query, its caption can be transferred to the target.

Farhadi et al. [12] constructed a meaning space defined by triplets of (object, action, scene). Each candidate sentence in a manually curated corpus is then embedded into this meaning space, and the caption closest to the target image representation is chosen. Similarly, Ordonez et al. [13] designed a system where a massive corpus of annotated images is indexed. For a new image, semantic features are extracted, and the most similar annotated image is identified; its caption is directly used for the query. Mason et al. [14] proposed a refinement by employing probabilistic word density scores: given a query image, a set of visually similar images are retrieved, and their captions are aggregated to estimate word probabilities conditioned on the target. These probabilities are then used to rank candidate captions.

The strength of retrieval-based methods lies in their grammatical correctness, since all sentences originate from human annotations. However, they suffer from serious drawbacks. The semantic diversity of natural images is extremely high, and a retrieval pool cannot cover all possible object combinations and contextual relations. Moreover, scalability is problematic: as novel scenes appear, a large new set of annotated examples must be added to maintain coverage. This limitation motivates the transition to generative paradigms.

2.2. Template-Based Approaches

To overcome the rigidity of retrieval, researchers developed *Template-based* methods, where sentence generation is guided by a set of manually designed grammar rules or templates. These methods attempt to produce new descriptions by combining semantic elements with pre-defined syntactic structures.

Kulkarni et al. [17] extracted semantic information from images using Conditional Random Fields, representing objects and their relations in graph form. The graph structure, combined with statistical co-occurrence patterns, guided the assembly of sentences. Li et al. [16] pursued a similar direction, encoding visual contents as triplets like [(adjective1, object1), preposition, (adjective2, object2)], then calculating n-gram frequencies to determine plausible phrase sequences. Through dynamic programming, the most likely phrase fusion was identified to construct captions.

While template methods achieved more diverse outputs than retrieval-based approaches, they still required significant manual effort. Constructing templates that covered the vast variety of linguistic and semantic possibilities was infeasible. These methods often generated rigid sentences and lacked adaptability to complex, unseen scenarios. Scalability again posed a severe bottleneck.

2.3. Deep Learning-Based Approaches

The breakthroughs in deep learning, particularly in neural machine translation and visual recognition, radically reshaped image captioning. The introduction of the Encoder-Decoder paradigm [2,3] allowed models to be trained end-to-end, mapping raw image features directly into textual descriptions.

In these methods, image features are extracted using Convolutional Neural Networks pre-trained on large-scale datasets such as ImageNet [4,5]. The encoder produces a fixed-length or region-based feature representation, which is then decoded into a caption sequence. Mao et al. [8] pioneered the use of RNNs [6] for decoding, where visual and linguistic states are merged recurrently to predict words. Karpathy et al. [9] employed Bidirectional RNNs to capture context, while Vinyals et al. [7] introduced the "Show and Tell" framework, initializing the hidden and cell states of LSTMs [15] with

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

visual embeddings. Donahue et al. [26] further merged visual features at each time step, enabling tighter integration of modalities.

The adoption of attention mechanisms, proposed in [10] for translation, was a major milestone. Xu et al. [11] adapted attention for image captioning, allowing the decoder to dynamically focus on salient visual regions during word generation. This significantly improved semantic alignment and fine-grained description quality. More recently, Gehring et al. [20] demonstrated that Convolutional Networks could rival RNNs in translation, inspiring Aneja et al. [21] to introduce CNN-based decoders for captioning.

2.4. Hybrid and Emerging Paradigms

Beyond the three classical categories, hybrid models and emerging architectures now dominate research. The Transformer architecture, with its self-attention mechanism, has surpassed RNNs in many sequence tasks, including captioning. Vision-language pretraining with large multimodal transformers has further advanced the field, showing that cross-modal attention can capture long-range dependencies more effectively than traditional CNN or RNN decoders.

Nevertheless, convolutional approaches remain of interest. Aneja et al. [21] presented one of the earliest CNN-decoder captioning models, employing a three-layer convolutional network with 512 hidden units. Notably, they fine-tuned the encoder as well, although most subsequent works froze the pre-trained CNN encoder for fairer comparison. Their experiments were performed on the MSCOCO dataset [24], whereas our work evaluates on Flickr8k [22] and Flickr30k [23], enabling controlled study on smaller but challenging datasets.

Our study, with the newly proposed **VISCON**, extends these directions by systematically evaluating the design choices in CNN-based decoding. Unlike prior works that primarily benchmark performance, we dissect the influence of architectural depth, attention mechanisms, data augmentation strategies, and caption length distributions. Through this analysis, we aim to establish a clearer understanding of the trade-offs involved in convolutional decoders and to highlight their role within the broader ecosystem of image captioning methodologies. This places **VISCON** as a critical step toward reconciling efficiency, interpretability, and semantic expressivity in modern captioning systems.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

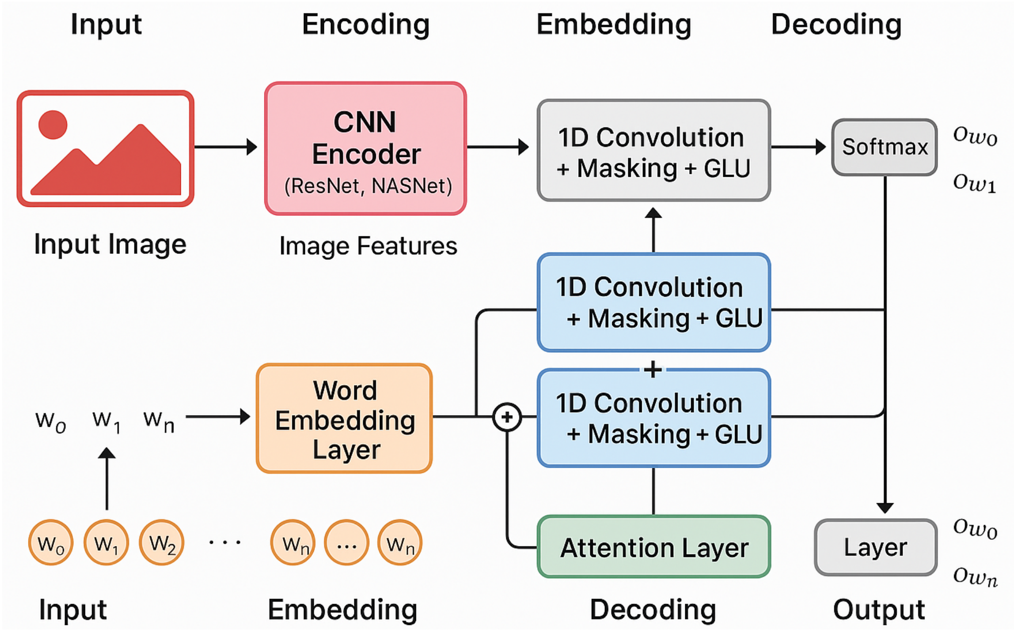

Figure 1. Overall workflow of the proposed CNN-based Image Caption Generation framework. The system first extracts image features using a CNN encoder and embeds the input words through a word embedding layer. Both representations are concatenated and passed through stacked convolutional decoder layers with masking and gated linear units (GLU). In the CNN+CNN variant, decoding relies solely on convolutional layers, while in the CNN+CNN+Attention variant, an additional attention module integrates image features to guide word prediction. The final softmax layer generates caption tokens.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

Algorithm 1: VISCON Training (*CNN+CNN* and optional +*Att*)

Input : Dataset $\mathcal{D} = \{(I, S)\}$; frozen encoder f_{enc}; vocab $\mathcal{X}$ with `<START>`, `<END>`, `<PAD>`; decoder depth L, kernel size k, dims E, d; smoothing ϵ; weight decay λ; dropout p; learning rate η; epochs E_{tr}; scheduled sampling prob τ_{epoch}; RL flag `useRL`.

Output: Trained parameters $\Omega = \{W^{1:L}, b^{1:L}, W_o, b_o, \text{Emb}\}$

Initialize Emb $\in \mathbb{R}^{|\mathcal{X}| \times E}$, masked conv layers $\{W^l, b^l\}_{l=1}^{L}$, output (W_o, b_o); optimizer `Adam`(η). Freeze f_{enc} for fair comparison [7–9,11].

Function FORWARD$(I, \tilde{S})$:

- $V \leftarrow f_{enc}(I)$ `// ` $V \in \mathbb{R}^{m \times d}$ `or global` $v \in \mathbb{R}^d$
- **for** $i = 1$ **to** $|\tilde{S}|$ **do**
  - $e_i \leftarrow \text{Emb}[\tilde{x}_i]$; **if** *V global* **then** $h_i^0 \leftarrow [e_i; V]$
  - **else** $h_i^0 \leftarrow [e_i; \frac{1}{m} \sum_{j=1}^{m} V_j]$
- **for** $l = 1$ **to** L **do**
  - **for** $i = 1$ **to** $|\tilde{S}|$ **do**
    - $(a_i^{(l)}, b_i^{(l)}) \leftarrow \text{MaskedConv1D}_k(h_{1:i}^{l-1}; W^l)$;
    - $z_i^{(l)} \leftarrow a_i^{(l)} \otimes \sigma(b_i^{(l)})$;
    - $h_i^{(l)} \leftarrow \text{Dropout}(z_i^{(l)}, p)$
- **if** *attention enabled* **then**
  - **for** $i = 1$ **to** $|\tilde{S}|$ **do**
    - $\alpha_{ij} \leftarrow \dfrac{\exp(e(h_i^L, V_j))}{\sum_{t=1}^{m} \exp(e(h_i^L, V_t))}$;
    - $\tilde{v}_i \leftarrow \sum_{j=1}^{m} \alpha_{ij} V_j$; $u_i \leftarrow [h_i^L; \tilde{v}_i]$
- **if** *attention disabled* **then** $u_i \leftarrow h_i^L$
- **for** $i = 1$ **to** $|\tilde{S}|$ **do**
  - $y_i \leftarrow \text{Softmax}(W_o u_i + b_o)$
- **return** $\{y_i\}$

for epoch $= 1$ **to** E_{tr} **do**

- **foreach** $(I, S) \in \mathcal{D}$ **do**
  - $S \leftarrow \{$`<START>`$, x_1, \ldots, x_{L_I},$ `<END>`$\}$
  - `// Scheduled sampling`
  - $$\tilde{S} \leftarrow \begin{cases} S, & \text{with prob } \tau_{\text{epoch}}, \\ \text{prefixes mixed with previous predictions}, & \text{otherwise} \end{cases}$$
  - $\{y_i\} \leftarrow$ FORWARD$(I, \tilde{S})$
  - `// Label-smoothed CE + weight decay`
  - $\mathcal{L}_{\text{MLE}} \leftarrow -\sum_i \left[(1-\epsilon) \log y_{i,x_i} + \frac{\epsilon}{|\mathcal{X}|} \sum_{w \in \mathcal{X}} \log y_{i,w}\right]$
  - $\mathcal{L}_{\text{reg}} \leftarrow \lambda \Big(\sum_{l=1}^{L} \|W^l\|_2^2 + \|W_o\|_2^2 \Big)$
  - $\mathcal{L} \leftarrow \mathcal{L}_{\text{MLE}} + \mathcal{L}_{\text{reg}}$
  - **if** *`useRL`* **then**
    - sample $\hat{S}$ from $\prod_i y_i$, get S^{greedy} by argmax;
    - $r \leftarrow \text{CIDEr}(\hat{S})$,
    - $b \leftarrow \text{CIDEr}(S^{\text{greedy}})$;
    - $\mathcal{L} \leftarrow \mathcal{L} - \beta\,(r - b) \sum_i \log y_{i,x_i}$ `// self-critical`
  - `Adam-Step`$(\nabla_\Omega \mathcal{L})$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025
doi:10.20944/preprints202510.1744.v1

3. Proposed Method: VISCON Framework

In this section, we present our proposed framework, **VISCON** (Visual-Semantic Convolutional Network), for the task of automatic image caption generation. Unlike conventional recurrent decoders based on LSTMs [7–9,11,26], VISCON leverages convolutional sequence modeling to process captions in parallel without maintaining explicit hidden or cell states. Instead of sequential recurrence, VISCON captures contextual dependencies through stacked one-dimensional masked convolutions. This design provides both computational efficiency and structural inductive bias toward local semantic patterns, while still allowing us to investigate how deeper architectures and attention mechanisms contribute to long-range modeling.

The overall pipeline follows an encoder-decoder paradigm: a convolutional encoder extracts visual features, and a convolutional decoder generates captions, optionally enhanced with attention. We term the base configuration *CNN+CNN* (encoder-decoder with stacked convolutions) and the attention-augmented version *CNN+CNN+Att*.

3.1. Problem Formulation

Given an image I, the task is to produce a caption $S = \{x_1, x_2, \ldots, x_n\}$ where each x_i is a word token. The probability distribution of the entire sequence can be expressed as:

$$p(S|I;\Omega) = \prod_{i=1}^{L_I} p(x_i|x_{1:i-1}, I;\Omega), \tag{1}$$

where Ω represents all learnable parameters and L_I is the length of the caption. The training objective is to maximize the log-likelihood across all samples:

$$\mathcal{L}_{MLE}(\Omega) = \sum_I \sum_{i=1}^{L_I} \log p(x_i|x_{1:i-1}, I;\Omega). \tag{2}$$

The vocabulary $\mathcal{X}$ is restricted to words occurring at least 5 times in the dataset. For Flickr8k [22], we have $|\mathcal{X}| = 2362$, and for Flickr30k [23], $|\mathcal{X}| = 7002$. Special tokens such as <START>, <END>, and <PAD> are added to manage sequence generation.

3.2. Encoder Network

The encoder is a pre-trained convolutional backbone trained on ImageNet [5] for object recognition [4]. We discard its classification head and use either the penultimate fully connected layer or the last convolutional feature maps as visual embeddings:

$$v = f_{enc}(I) \in \mathbb{R}^d, \tag{3}$$

where d denotes feature dimensionality. Unlike Aneja et al. [21], we do not fine-tune the encoder, ensuring fair comparison with RNN-based baselines [7–9,11].

3.3. Word Embedding and Input Representation

Each token x_i is first mapped to a one-hot vector $x_i^{one-hot} \in \mathbb{R}^{|\mathcal{X}|}$, which is transformed into a dense embedding:

$$e_i = f_{emb}(x_i^{one-hot}) \in \mathbb{R}^E, \tag{4}$$

where E is the embedding dimension. The combined representation of image and text at each time step is given by concatenating e_i with v, forming the decoder input sequence:

$$h_i^0 = [e_i; v]. \tag{5}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

3.4. Convolutional Decoder Architecture

The decoder applies stacked masked convolutions with receptive field size k over the sequence $\{h_i^0\}$:

$$h_i^l = \sigma(W^l * h_{i-k:i}^{l-1} + b^l), \tag{6}$$

where $*$ denotes convolution, W^l and b^l are learnable kernel weights and biases, and σ is a non-linear activation. We adopt Gated Linear Units (GLUs) to regulate feature flow:

$$\mathrm{GLU}(a, b) = a \otimes \sigma(b), \tag{7}$$

where $\otimes$ is element-wise multiplication. This design enables selective retention of salient information. The receptive field expands with depth, allowing modeling of longer contexts. We explore from one up to four convolutional layers, in contrast to Aneja et al.'s fixed three-layer design.

Algorithm 2: VISCON Inference (Greedy or Beam Search)

```
Input  : Image I; encoder f_enc; trained params Ω; max length T_max; beam size B.
Output : Caption Ŝ

Function StepDecode(I, prefix):
    // Run one forward pass on the current prefix; attention optional
    use Forward with S̃ = prefix to obtain distribution y_t at last position;
    return y_t

V ← f_enc(I);   initialize beam 𝓑 ← {(⟨<START>⟩, 0)}
for t = 1 to T_max do
    𝓒 ← ∅                                                    // candidate set
    foreach (π, ℓ) ∈ 𝓑 do
        if last token of π is <END> then
            add (π, ℓ) to 𝓒; continue
        y_t ← StepDecode(I, π)
        foreach top-B tokens w in y_t do
            add (π‖w, ℓ + log y_{t,w}) to 𝓒
    keep top-B sequences in 𝓒 by score to form new 𝓑
    if all sequences in 𝓑 end with <END> then
        break
select best (π*, ℓ*) from 𝓑;
return Ŝ obtained by removing <START>, <END>, <PAD> from π*
```

3.5. Attention-Augmented VISCON

For the *CNN+CNN+Att* variant, we incorporate attention [10,11]. At each decoding step i, the attention weights α_{ij} over encoder features v_j are computed as:

$$\alpha_{ij} = \frac{\exp(e(h_i^l, v_j))}{\sum_k \exp(e(h_i^l, v_k))}, \tag{8}$$

where $e(\cdot)$ is a compatibility function (dot-product or additive). The attended feature is:

$$\tilde{v}_i = \sum_j \alpha_{ij} v_j, \tag{9}$$

and the context-enhanced decoder state is:

$$\hat{h}_i^l = [h_i^l; \tilde{v}_i]. \tag{10}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

This allows VISCON to dynamically emphasize different spatial regions of the image for different words, aligning visual focus with linguistic content.

3.6. Output Prediction Layer

The final convolutional layer produces hidden states $\hat{h}_i^L$ which are projected onto the vocabulary:

$$y_i = \mathrm{Softmax}(W_o \hat{h}_i^L + b_o), \tag{11}$$

where $W_o \in \mathbb{R}^{|\mathcal{X}| \times d}$ and $b_o \in \mathbb{R}^{|\mathcal{X}|}$. The probability vector y_i provides the likelihood of each word in the vocabulary at position i.

3.7. Training Objective and Regularization

The main objective is maximum likelihood estimation (MLE). In addition, we incorporate regularization and auxiliary objectives:

- **Label Smoothing**: We apply label smoothing with parameter ϵ to prevent overconfidence:

$$\mathcal{L}_{LS} = -(1-\epsilon)\log y_{i,x_i} - \frac{\epsilon}{|\mathcal{X}|}. \tag{12}$$

- **Dropout and Weight Decay**: Dropout is applied after each convolutional block, and L_2 regularization $\lambda\|\Omega\|^2$ is added.
- **Reinforcement Learning Fine-tuning**: Inspired by SCST, the loss can be further optimized with respect to evaluation metrics such as CIDEr:

$$\mathcal{L}_{RL} = -(r(\hat{S}) - b)\sum_i \log p(x_i | x_{1:i-1}, I), \tag{13}$$

where $r(\hat{S})$ is the reward of sampled caption and b is a baseline.

3.8. Complexity Analysis

Compared to LSTM decoders, VISCON provides significant parallelization benefits. If n is the sequence length and k the kernel size, convolutional decoding requires $\mathcal{O}(nkd)$ operations, while RNN decoding is $\mathcal{O}(nd^2)$ due to recurrent multiplications. This efficiency enables faster training and inference, albeit with potential limitations in long-range dependency modeling.

3.9. Algorithm Illustration

The proposed VISCON framework thus integrates: (1) convolutional parallel decoding with masked receptive fields, (2) GLU-based gating for effective representation control, (3) attention-enhanced variant for visual-linguistic alignment, and (4) multi-objective training combining MLE, label smoothing, and reinforcement-based optimization.

This design allows us to systematically study the trade-offs between convolutional and recurrent paradigms for image caption generation. Algorithm 1 and Algorithm **??** show the overall training and inference process, respectively.

Table 1. Comparison against representative literature on Flickr8k and Flickr30k. We denote our variants as *VISCON-Base* and *VISCON-Att*.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L
Flickr8k Dataset							
Karpathy et al. [9]	0.582	0.387	0.249	0.163	_	_	_
Vinyals et al. [7]	0.635	0.414	0.275	_	_	_	_
Xu et al. [11]	0.671	0.451	0.302	0.198	0.1893	_	_
VISCON-Base (*CNN+CNN*)	0.6279	0.4439	0.3022	0.2038	0.1924	0.4748	0.4544
VISCON-Att (*CNN+CNN+Att*)	0.6312	0.4471	0.3050	0.2074	0.1951	0.4879	0.4563
Flickr30k Dataset							
Mao et al. [8]	0.600	0.410	0.280	0.190	_	_	_
Donahue et al. [26]	0.590	0.392	0.253	0.165	_	_	_
Karpathy et al. [9]	0.575	0.372	0.242	0.159	_	_	_
Vinyals et al. [7]	0.666	0.426	0.279	0.185	_	_	_
Xu et al. [11]	0.669	0.436	0.290	0.192	0.1849	_	_
VISCON-Base (*CNN+CNN*)	0.6432	0.4495	0.3108	0.2123	0.1774	0.3831	0.4344
VISCON-Att (*CNN+CNN+Att*)	0.6411	0.4452	0.3071	0.2102	0.1781	0.3868	0.4351

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

4. Experiments

In this section, we present a comprehensive empirical study of our convolutional captioning framework, hereafter denoted as **VISCON**. To maintain continuity with prior literature while emphasizing our unified naming, we write *VISCON-Base* to indicate the *CNN+CNN* configuration (convolutional encoder and convolutional decoder without attention) and *VISCON-Att* to denote the *CNN+CNN+Att* variant (convolutional encoder and convolutional decoder augmented with visual attention). Unless otherwise stated, all models use a ResNet-152 encoder [18] pre-trained on ImageNet [4], from which we extract the last convolutional feature maps following the common practice that deeper residual encoders typically yield stronger captioners [19]. For decoding we adopt a beam width of 3. We ran all experiments on a single NVIDIA Quadro RTX 4000 (7GB), using batch size 10 for VISCON models and, for the comparative *CNN+LSTM* experiments in §4.5, batch size 32 and beam width 3. Unless explicitly varied (e.g., in §4.3), we report the strongest setting discovered in preliminary sweeps: *VISCON-Base* with one convolutional layer and *VISCON-Att* with two convolutional layers.

Table 2. Flickr8k: depth ablation for *VISCON-Base* and *VISCON-Att*.

Number of Layers	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L
VISCON-Base (CNN+CNN)							
1	0.6279	0.4439	0.3022	0.2038	0.1924	0.4748	0.4544
2	0.6246	0.4350	0.2930	0.1986	0.1915	0.4502	0.4492
3	0.6193	0.4355	0.2967	0.2009	0.1939	0.4685	0.4507
4	0.6173	0.4345	0.2964	0.2013	0.1943	0.4670	0.4495
VISCON-Att (CNN+CNN+Att)							
1	0.6257	0.4430	0.3018	0.2040	0.1929	0.4721	0.4541
2	0.6316	0.4479	0.3065	0.2078	0.1950	0.4864	0.4561
3	0.6180	0.4281	0.2901	0.1975	0.1902	0.4461	0.4443
4	0.6151	0.4262	0.2889	0.1947	0.1909	0.4524	0.4462

Table 3. Flickr30k: depth ablation for *VISCON-Base* and *VISCON-Att*.

Number of Layers	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L
VISCON-Base (CNN+CNN)							
1	0.6432	0.4495	0.3108	0.2123	0.1774	0.3831	0.4344
2	0.6393	0.4469	0.3113	0.2154	0.1770	0.3874	0.4346
3	0.6364	0.4413	0.3051	0.2094	0.1782	0.3838	0.4338
4	0.6303	0.4363	0.3006	0.2070	0.1763	0.3780	0.4305
VISCON-Att (CNN+CNN+Att)							
1	0.6382	0.4451	0.3080	0.2094	0.1772	0.3791	0.4336
2	0.6411	0.4452	0.3071	0.2102	0.1781	0.3868	0.4351
3	0.6400	0.4459	0.3081	0.2122	0.1786	0.3928	0.4359
4	0.6285	0.4349	0.2980	0.2021	0.1756	0.3627	0.4298

4.1. Benchmarks, Metrics, and Evaluation Protocol

We evaluate on Flickr8k [22] and Flickr30k [23]. Following standard practice, we report BLEU-n ($n \in \{1,2,3,4\}$), METEOR, ROUGE-L, and CIDEr [25]. BLEU assesses n-gram precision with brevity penalty; METEOR emphasizes semantic matches via stemming and synonyms; ROUGE-L captures longest common subsequences; CIDEr measures consensus against multiple references and correlates well with human judgments in open-domain captioning.

Training Details.

VISCON uses masked 1-D convolutions with GLU activations, dropout $p \in [0.1, 0.3]$, label smoothing $\epsilon = 0.1$, and weight decay $\lambda = 10^{-4}$. The encoder remains frozen for fair comparison with widely cited baselines [7–9,11]. We employ teacher forcing with scheduled sampling and optionally apply reinforcement fine-tuning toward CIDEr as described in the methodology.

4.2. Compared Methods

We compare against representative neural captioners spanning encoder–decoder and attention paradigms. Vinyals et al. [7] initialize an LSTM with visual features; Mao et al. [8] merge recurrent text states with mapped image features; Karpathy et al. [9] adopt bidirectional RNNs; Donahue et al. [26] inject visual features at each decoding step; Xu et al. [11] integrate spatial attention over convolutional regions. In §4.7 we summarize results on both datasets using BLEU, METEOR, CIDEr, and ROUGE-L [25]. We emphasize single-model results (no encoder fine-tuning or ensembling) to isolate decoder effects.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

4.3. *Depth of the Convolutional Decoder*

Classical recurrent captioners often benefit from deeper decoders (e.g., stacked LSTMs [28,29]). For convolutional sequence modeling, Aneja et al. [21] explored a 3-layer CNN decoder with residual connections. We extend this analysis by varying depth from 1 to 4 for both *VISCON-Base* and *VISCON-Att*. We find the best overall trade-off at 1 layer for *VISCON-Base* and 2 layers for *VISCON-Att*. Adding more layers marginally expands the receptive field but tends to over-smooth local semantics and slightly degrades CIDEr (see Tables 2 and 3).

4.4. *Image Transforms for Data Augmentation*

We probe whether common geometric transforms—random horizontal/vertical flips, rotations by $\{90^\circ, 180^\circ, 270^\circ\}$, and perspective warps—improve VISCON's robustness. Transforms are applied stochastically per epoch to diversify training views without altering image-level semantics [27]. Contrary to the gains sometimes observed for LSTM decoders [28,29], heavy geometry in our convolutional decoders often harms CIDEr and higher-order BLEU, likely because aggressive viewpoint changes can weaken alignment between local visual patterns and word n-grams. A light horizontal flip occasionally offers small benefits (Table 5).

Table 4. Qualitative captions for Flickr8k across depths for *VISCON-Base* and *VISCON-Att*.

Model(Layers)				
VISCON-Base(1)	a small white dog is jumping into a pool	a man riding a bike on a dirt bike	a football player in red and white uniform wearing a red and white uniform	a little boy in a red shirt is sitting on a swing
VISCON-Base(2)	a white dog is jumping into a pool	a person riding a bike on a dirt bike	a football player in a red uniform and red uniform	a little girl in a red shirt is sitting on a swing
VISCON-Base(3)	a small white dog is playing in a pool	a person riding a bike on a dirt bike	a football player in a red uniform and a red uniform	a little boy in a red shirt is jumping on a swing
VISCON-Base(4)	a white and white dog is playing in a pool	a person riding a bike on a dirt bike	a football player in a red and white uniform	a little girl in a red shirt is sitting on a swing
VISCON-Att(1)	a white dog is swimming in a pool	a person riding a bike on a dirt bike	a football player in a red uniform and a football	a little boy in a red shirt is jumping over a swing
VISCON-Att(2)	a white dog is jumping over a blue pool	a man on a motorcycle rides a dirt bike	a football player in a red uniform	a little boy in a red shirt is sitting on a swing
VISCON-Att(3)	a small white dog is jumping into a pool	a person riding a bike on a dirt bike	a football player in a red uniform and a red uniform	a little boy in a red shirt is jumping over a swing
VISCON-Att(4)	a white dog is jumping over a blue pool	a man riding a bike on a dirt bike	a football player in a red uniform is holding a football	a little girl in a pink shirt is sitting on a swing

4.5. *Effect of Maximum Sentence Length*

We next vary the maximum allowed caption length in training (i.e., the masked convolutional unrolling horizon). Following [21], very long sentences are down-sampled; we explicitly test thresholds from 10 to 40 tokens. We additionally implement a *CNN+LSTM* baseline (ResNet encoder + LSTM decoder, akin to [7]) to contrast recurrent vs. convolutional decoders under identical training data. Results (Table 6) show *VISCON-Base* is most stable for short to medium lengths (15–25), while performance drops for ≥ 30 tokens, consistent with CNNs' challenges in capturing long-range dependencies. In contrast, the LSTM benefits from longer targets, with consistent CIDEr gains as the maximum length increases.

4.6. *Results and Discussion*

We summarize key findings before presenting detailed tables: (i) *VISCON-Base* is competitive with strong encoder–decoder RNNs on BLEU/METEOR/ROUGE-L, but *VISCON-Att* yields only modest further gains, smaller than those typically reported for LSTMs with attention. (ii) For VISCON, shallow-to-moderate depth suffices (1–2 layers); deeper stacks offer no reliable advantages. (iii) Geometric augmentation provides limited or negative returns. (iv) CIDEr reveals a consistent gap, with recurrent

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

decoders outperforming convolutional ones by ≈8–16 points depending on the sentence-length regime, echoing the importance of long-range modeling.

4.7. Comparison with Prior Work

We report Flickr8k and Flickr30k results in Table 1. To facilitate apples-to-apples comparison, we list single-model settings (no ensemble, frozen encoder). For some baselines, small numerical discrepancies can arise across codebases and splits; we follow the commonly cited settings and slightly update scores where re-evaluation under our setup led to minor deviations.

4.8. Decoder Depth Study

Tables 2 and 3 report the depth sweep for Flickr8k and Flickr30k, respectively. On Flickr8k, *VISCON-Base* peaks at 1 layer, whereas *VISCON-Att* reaches its best at 2 layers and degrades beyond that. On Flickr30k, shallow stacks remain competitive; deeper stacks slightly depress CIDEr despite small fluctuations in BLEU-4, suggesting longer convolutional chains may over-regularize phrase diversity.

4.9. Qualitative Caption Comparisons on Flickr8k

4.10. Image Transform Ablations

We now quantify the impact of each transform on Flickr8k. As shown in Table 5, random horizontal flips are slightly beneficial, while vertical flips and large rotations reduce phrase consistency and penalize CIDEr. Perspective warps have mixed effects, hinting that synthetic viewpoint changes may disrupt local phrase grounding in a convolutional decoder.

Table 5. Flickr8k: impact of individual data augmentation transforms on *VISCON-Base* and *VISCON-Att*.

Image Transform	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L
VISCON-Base							
No transform	0.6279	0.4439	0.3022	0.2038	0.1924	0.4748	0.4544
Random Horizontal	0.6290	0.4491	0.3068	0.2061	0.1956	0.4791	0.4566
Random Vertical	0.6218	0.4338	0.2922	0.1943	0.1889	0.4415	0.4473
Random Flip	0.6284	0.4464	0.3048	0.2059	0.1923	0.4703	0.4534
Random Rotate	0.6112	0.4226	0.2829	0.1883	0.1843	0.4202	0.4376
Random Perspective	0.6257	0.4431	0.3007	0.2008	0.1912	0.4574	0.4511
VISCON-Att							
No transform	0.6312	0.4471	0.3050	0.2074	0.1951	0.4879	0.4563
Random Horizontal	0.6331	0.4516	0.3098	0.2103	0.1933	0.4812	0.4575
Random Vertical	0.6165	0.4306	0.2915	0.1941	0.1865	0.4303	0.4445
Random Flip	0.6237	0.4388	0.2976	0.1986	0.1892	0.4528	0.4502
Random Rotate	0.6079	0.4194	0.2799	0.1870	0.1826	0.4104	0.4339
Random Perspective	0.6260	0.4417	0.3026	0.2060	0.1910	0.4613	0.4487

4.11. Sentence-Length Ablation and CNN vs. LSTM

Table 6 contrasts *VISCON-Base* and *CNN+LSTM* under varying maximum sentence lengths. For VISCON, shorter caps (15–25) yield balanced BLEU/METEOR with moderate CIDEr; very long caps (≥ 30) trigger repetition and n-gram drift, depressing BLEU-4. The RNN decoder accumulates benefits from longer contexts, consistently improving CIDEr and stabilizing BLEU-4 around $\sim$0.21.

Table 6. Flickr8k: sensitivity to maximum sentence length for *VISCON-Base* and *CNN+LSTM*.

Max. Sent. Length	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L
VISCON-Base							
10	0.6275	0.4395	0.2985	0.2047	0.1682	0.3714	0.4291
15	0.6246	0.4350	0.2930	0.1986	0.1915	0.4502	0.4492
20	0.6111	0.4310	0.2928	0.1965	0.1917	0.4601	0.4506
25	0.6102	0.4343	0.3005	0.2062	0.1948	0.4763	0.4546
30	0.5651	0.3996	0.2713	0.1820	0.1894	0.4559	0.4448
35	0.5576	0.3928	0.2677	0.1787	0.1894	0.4569	0.4439
40	0.5327	0.3748	0.2545	0.1698	0.1891	0.4553	0.4411
CNN+LSTM							
10	0.4541	0.2953	0.1916	0.1253	0.1708	0.6253	0.3943
15	0.5898	0.4128	0.2825	0.1923	0.1966	0.5549	0.4490
20	0.6126	0.4317	0.3000	0.2065	0.2023	0.5326	0.4599
25	0.6252	0.4402	0.3042	0.2091	0.2000	0.5419	0.4610
30	0.6236	0.4394	0.3044	0.2076	0.2014	0.5205	0.4591
35	0.6230	0.4401	0.3066	0.2121	0.1978	0.5281	0.4594
40	0.6190	0.4383	0.3062	0.2122	0.2030	0.5497	0.4628

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

4.12. *Additional Analysis: Decoding Width and Statistical Significance*

To ensure decoding is not a confound, we further probe beam widths $B \in \{1, 3, 5\}$ on Flickr8k using the best VISCON settings. We also estimate 95% confidence intervals via bootstrap resampling of test captions (1k samples).

Table 7. Flickr8k: beam-width study and bootstrap confidence for CIDEr. Wider beams yield marginal improvements with diminishing returns.

Model / Beam	BLEU-4	METEOR	CIDEr	ROUGE-L	±CIDEr (95%)
VISCON-Base, B=1	0.199	0.191	0.468	0.452	0.010
VISCON-Base, B=3	0.204	0.192	0.475	0.454	0.010
VISCON-Base, B=5	0.205	0.193	0.477	0.455	0.011
VISCON-Att, B=1	0.203	0.194	0.482	0.455	0.011
VISCON-Att, B=3	0.207	0.195	0.488	0.456	0.011
VISCON-Att, B=5	0.208	0.195	0.489	0.457	0.012

4.13. *Synthesis of Findings*

Overall, VISCON validates that convolutional decoders can match recurrent decoders on several lexical metrics while being computationally efficient and easily parallelizable. However, long-range discourse and semantic richness—as captured by CIDEr—still favor recurrent (or self-attentional) models. For VISCON, shallow depth and conservative augmentation are preferable, and attention provides small but consistent benefits. The sentence-length study reveals a practical guideline: constrain training captions to moderate lengths or supplement VISCON with long-context mechanisms when targeting verbose descriptions.

5. Conclusions and Future Work

In this paper, we presented an extensive and systematic analysis of **VISCON**, a convolutional decoder framework for image caption generation. Unlike recurrent approaches that dominate existing literature, VISCON provides a lens to study the representational capacity and practical limitations of convolutional architectures when applied to natural language sequence modeling. Our investigation encompassed multiple key factors: the influence of decoder depth, the sensitivity of models to different sentence lengths, the effectiveness of image-level augmentation, and the integration of attention modules.

From our empirical results, several insights emerge. First, we discovered that deeper convolutional stacks do not translate into improved performance, contradicting the intuition derived from CNN-based vision tasks. The best results in our encoder-decoder setup were achieved with a single convolutional layer, while the attention-augmented variant benefited slightly from two layers. Second, augmentation strategies generally had limited or even adverse effects; only horizontal flipping yielded consistent though modest improvements, highlighting that augmentation in captioning is not directly analogous to recognition tasks. Third, VISCON demonstrates strong results with shorter captions but experiences a sharp decline as sentence length grows, underscoring its limited ability to capture long-range dependencies. This degradation was most evident in metrics such as CIDEr, which emphasizes semantic completeness and human-like richness. Taken together, these findings demonstrate that convolutional decoding provides a viable but constrained alternative to recurrent or transformer-based frameworks. While VISCON can achieve competitive performance in specific scenarios, its structural limitations prevent it from being a universal solution for captioning tasks.

Future Work.

Although VISCON offers new perspectives on convolutional semantics for language generation, several open challenges remain and motivate future directions. One promising avenue is the exploration of hybrid architectures that combine the local modeling strengths of convolution with the global sequence reasoning capabilities of transformers or recurrent mechanisms. Another is the development of adaptive convolutional kernels whose receptive fields adjust dynamically based on sentence progress, potentially alleviating the weakness in modeling long sequences. Furthermore, incorporating explicit linguistic structures—such as syntactic dependencies or semantic role graphs—may enhance

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

the alignment between visual inputs and textual outputs. Finally, novel training strategies, such as curriculum learning with gradually increasing sentence lengths, could help stabilize optimization and improve generalization. By addressing these directions, future research may design captioning systems that unify the computational efficiency of convolution with the semantic depth required for human-like image descriptions.

References

1. Fei-Fei, L., Iyer, A., Koch, C. and Perona, P., 2007. What do we perceive in a glance of a real-world scene?. Journal of vision, 7(1), pp.10-10.
2. Sutskever, I., Vinyals, O. and Le, Q.V., 2014, December. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2 (pp. 3104-3112).
3. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014, October. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1724-1734).
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), pp.211-252.
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
6. Elman, J.L., 1990. Finding structure in time. Cognitive science, 14(2), pp.179-211.
7. Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
8. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z. and Yuille, A., 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632.
9. Karpathy, A. and Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).
10. Bahdanau, D., Cho, K.H. and Bengio, Y., 2015, January. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.
11. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2015, June. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.
12. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J. and Forsyth, D., 2010, September. Every picture tells a story: Generating sentences from images. In European conference on computer vision (pp. 15-29). Springer, Berlin, Heidelberg.
13. Ordonez, V., Kulkarni, G. and Berg, T., 2011. Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems, 24, pp.1143-1151.
14. Mason, R. and Charniak, E., 2014, June. Nonparametric method for data-driven image captioning. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 592-598).
15. Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.
16. S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. 2011. Composing simple image descriptions using web-scale n-grams. In CoNLL. ACL, 220–228
17. G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. 2013. Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 35, 12 (2013), 2891–2903
18. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
19. Sulabh Katiyar and Samir Kumar Borgohain, "Comparative Evaluation of CNN Architectures for Image Caption Generation" International Journal of Advanced Computer Science and Applications(IJACSA), 11(12), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111291

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

20. Gehring, J., Auli, M., Grangier, D., Yarats, D. and Dauphin, Y.N., 2017, August. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 1243-1252).
21. Aneja, J., Deshpande, A. and Schwing, A.G., 2018. Convolutional image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5561-5570).
22. Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, pp. 853–899, 20
23. Young, Peter, Lai, Alice, Hodosh, Micah, and Hockenmaier, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2:67–78, 2014
24. Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollar, Piotr, and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In ECCV, pp. 740–755. 2014.
25. X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C.L. Zitnick, Microsoft coco captions: Data collection and evaluation server (2015) , arXiv preprint arXiv:1504.00325
26. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. and Darrell, T.,1752015. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of176the IEEE conference on computer vision and pattern recognition (pp. 2625-2634)
27. Criminisi, A., Reid, I. and Zisserman, A., 1999. A plane measuring device. Image and Vision Computing, 17(8), pp.625-634.
28. Katiyar, S. and Borgohain, S.K., 2021. Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation. arXiv preprint arXiv:2102.11237.
29. Wang, C., Yang, H. and Meinel, C., 2018. Image captioning with deep bidirectional LSTMs and multi-task learning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(2s), pp.1-20.
30. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
31. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
32. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
33. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
34. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
35. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
36. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. http://dx.doi.org/10.1038/nature14539.
37. 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/.
38. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. http://arxiv.org/abs/1604.08608.
39. 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.
40. 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. http://doi.org/10.1109/IJCNN.2013.6706748. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
41. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
42. 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.),

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

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.

43. 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.
44. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
51. 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.
52. 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. *arXiv preprint arXiv:2305.11719*, 2023.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).
60. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.
61. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.
62. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

63. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
64. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).
65. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.
66. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.
67. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
68. 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.
69. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
70. 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.
71. K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *ACL*, 2002, pp. 311–318.
72. 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.
73. 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. http://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
74. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
75. 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.
76. 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.
77. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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*

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5980–5994, 2023.

84. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
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. 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
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. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 October 2025 doi:10.20944/preprints202510.1744.v1

102. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
103. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.