Fusion of Visual and Textual Data for Enhanced Semantic Representations

Lyra Sterling , Kairos Vale * , Ava Martinez

Posted Date: 26 September 2024

doi: 10.20944/preprints202409.2066.v1

Keywords: Multimodal Integration; Semantic Embeddings; Representation Learning; Transfer Learning

 doi:10.20944/preprints202409.2066.v

Article

Fusion of Visual and Textual Data for Enhanced Semantic Representations

Lyra Sterling, Kairos Vale * and Ava Martinez

University of Central Oklahoma
* Correspondence: kvale@uco.edu

Abstract: Generic text embeddings have demonstrated considerable success across a multitude of applications. However, these embeddings are typically derived by modeling the co-occurrence patterns within text-only corpora, which can limit their ability to generalize effectively across diverse contexts. In this study, we investigate methodologies that incorporate visual information into textual representations to overcome these limitations. Through extensive ablation studies, we introduce a novel and straightforward architecture named VisualText Fusion Network (VTFN). This architecture not only surpasses existing multimodal approaches on a range of well-established benchmark datasets but also achieves state-of-the-art performance on image-related textual datasets while utilizing significantly less training data. Our findings underscore the potential of integrating visual modalities to substantially enhance the robustness and applicability of text embeddings, paving the way for more nuanced and contextually rich semantic representations.

Keywords: multimodal Integration; semantic embeddings; representation learning; transfer learning

1. Introduction

The quest for effective and meaningful representations of textual data remains a highly active and dynamic area of research within the field of natural language processing. Numerous models have been developed with the capability to learn such representations by directly optimizing for specific end-to-end tasks within a supervised learning framework [42,46,47,54]. These models often rely on vast amounts of labeled data to achieve high performance, which poses significant challenges in practical applications where such extensive labeled datasets are not readily available [49,55]. The process of acquiring and annotating large-scale datasets can be both prohibitively expensive and time-consuming, thereby limiting the scalability and applicability of these models in real-world scenarios.

A widely adopted alternative strategy to mitigate the dependency on large labeled datasets involves leveraging pre-trained embeddings [48,62]. These embeddings, which are trained on extensive corpora, require substantially fewer labeled examples—often an order of magnitude less—to achieve competitive performance on downstream tasks [50]. This approach leverages the transferability of embeddings learned from large-scale datasets, thereby reducing the need for extensive labeled data and enabling the application of these models to a broader range of tasks with limited annotated resources.

A significant portion of existing research in this domain focuses on training embeddings exclusively from textual data. While these text-only embeddings have proven effective, they often struggle to capture certain types of relationships and co-occurrences that are more naturally and intuitively represented in other modalities, particularly visual data [51]. Visual information provides complementary cues that can enhance the semantic richness and contextual understanding of text [52]. For instance, visual similarity measures derived from image data, when combined with paired image-text data, can facilitate the induction of more nuanced and contextually relevant similarity measures between sentences, thereby enriching the quality of the resulting text embeddings.

In this paper, we explore the construction of sentence embeddings by harnessing text-image pairs to enhance sentence similarity metrics. This approach builds upon and extends prior works such as [17,20], which have demonstrated the potential of multimodal data in improving text representations.

 doi:10.20944/preprints202409.2066.v

By integrating visual information, we aim to capture a broader spectrum of semantic relationships that are not easily discernible from text alone [53].

We introduce a novel and conceptually straightforward model, referred to as **VisualText Fusion Network (VTFN)**, which integrates visual information from images to refine the quality of sentence embeddings. The VTFN model leverages existing components and synergistically combines them to achieve superior performance. Specifically, it employs a pre-trained Convolutional Neural Network (CNN) to generate image embeddings, while sentence embeddings are derived by normalizing the aggregate of individual word embeddings. These word embeddings are trained in an end-to-end manner to align with their corresponding image embeddings while ensuring that misaligned pairs do not share such alignment, optimizing the Pearson correlation coefficient as the objective function.

Despite its inherent simplicity, the VTFN model significantly outperforms pure text-based models as well as the leading multimodal model presented in [20,70] across a suite of well-established text similarity benchmarks, including those from the SemEval competition [23]. Notably, for datasets that are inherently image-related, our model not only matches but also sets new state-of-the-art results while utilizing a fraction of the training data previously required. These outcomes underscore the efficacy of incorporating visual data into text embedding processes, demonstrating that visual modalities can significantly bolster the semantic understanding and generalization capabilities of text representations [71,72]. Furthermore, we conduct a thorough ablation study to assess the impact of various factors on embedding quality within the context of image-to-text knowledge transfer.

In summary, the key contributions of this work are as follows:

- We introduce VTFN, a straightforward multimodal model that surpasses existing image-text integration approaches across a diverse array of text similarity evaluation tasks. Moreover, VTFN achieves state-of-the-art performance on image-related SemEval datasets while requiring significantly less training data.
- We conduct an exhaustive investigation into image-to-text knowledge transfer, evaluating various model architectures, text encoding strategies, loss functions, and dataset configurations to identify optimal practices for enhancing embedding quality.
- We demonstrate that directly learning sentence embeddings through our proposed VTFN method consistently outperforms traditional techniques that first learn word-level embeddings and subsequently aggregate them, highlighting the advantages of end-to-end sentence embedding approaches.

2. Related Work

The exploration of multimodal data, particularly the integration of image and text pairs, has garnered significant attention in recent years. Researchers have delved into leveraging these paired modalities to enhance performance across various tasks that inherently require an understanding of both visual and textual information. Notably, applications such as image captioning [13,32], where the goal is to generate descriptive text for a given image, and image retrieval [6,12,16,17,31,34,74,85], which involves finding relevant images based on textual queries, have benefited immensely from the advancements in multimodal learning.

A substantial body of work in this domain focuses on creating shared embeddings that bridge the gap between visual and textual data. These shared embeddings facilitate tasks that require a seamless transition between modalities, enabling models to comprehend and generate content that aligns both visually and semantically with the input data. For instance, [13,76] introduced a model that combines convolutional neural networks (CNNs) for image processing with recurrent neural networks (RNNs) for text generation, achieving remarkable results in image captioning. Similarly, [32] proposed an encoder-decoder framework that effectively maps images to their corresponding textual descriptions, thereby enhancing the quality and relevance of generated captions.

Despite the promising advancements in tasks that directly involve both images and text, the utilization of images as auxiliary data to enhance natural language processing (NLP) tasks remains

 doi:10.20944/preprints202409.2066.v

relatively underexplored. Most existing studies primarily focus on tasks that require direct interaction between the two modalities, such as those mentioned above, rather than using visual data to inform and improve text embeddings for broader NLP applications. One notable exception is the work by [17], who extended the skip-gram algorithm [21] to incorporate visual information. In the traditional skip-gram model, each word embedding is optimized to predict the surrounding context words, thereby capturing semantic relationships based solely on textual co-occurrence. [17,82] augmented this approach by introducing a mechanism to maximize the similarity between word embeddings and their corresponding image embeddings. Specifically, they employed a max-margin loss function to align the textual and visual modalities:

$$L_{vision}(w_t) = E_{w'} \max(0, \gamma - \cos(f(w_t), g(w_t)) + \cos(f(w_t), g(w')))$$

where $g(w)$ represents the average embedding of the images associated with the word w, and γ is the margin parameter. This formulation ensures that the embedding of a word is not only predictive of its textual context but also closely aligned with its visual representations, thereby enriching the semantic information captured by the embeddings. Similarly, [9,63] adopted a comparable strategy by integrating image data into the learning process of word embeddings. Their approach involved jointly optimizing the embeddings to capture both textual and visual co-occurrences, thereby enhancing the semantic richness and generalizability of the learned representations. This joint optimization allows the model to leverage visual cues that are often implicit or hard to discern from text alone, thereby addressing some of the limitations inherent in text-only embeddings.

Building upon these foundations, [16] introduced a model that employs a max-margin loss to co-embed images and their corresponding textual descriptions. Unlike previous approaches that primarily focus on word-level embeddings, [16] utilized different textual models depending on the specific task at hand. For tasks such as image captioning or retrieval, they employed an LSTM-based encoder to process textual data, enabling the model to generate coherent and contextually relevant descriptions. Additionally, for investigating the properties of the resulting word embeddings, particularly their ability to capture arithmetic relationships, they employed a simpler word embedding model. Their experiments demonstrated intriguing arithmetic capabilities, such as vector arithmetic operations where "image of a blue car" minus "blue" plus "red" yields an image of a red car. However, despite these qualitative observations, [16] did not provide a quantitative evaluation of the text embeddings' quality in terms of text similarity metrics, leaving room for further exploration in this area.

In more recent developments, [20] explored phrase embeddings that are trained using visual signals to assess their efficacy in capturing semantic similarities. Their approach involved using a Recurrent Neural Network (RNN) as a language model to learn word embeddings, which were subsequently combined to form phrase embeddings. They proposed three distinct models to evaluate the effectiveness of incorporating visual information:

1. **Model A**: This model mirrors the captioning framework introduced by [32], where an RNN decoder is conditioned on a pre-trained CNN embedding. Specifically, the RNN (using a Gated Recurrent Unit, or GRU, in their experiments) processes the input text to predict the next token in the sequence, with the initial state being a transformation of the final internal layer of a pre-trained VGGNet [52], denoted as v_{image}.
2. **Model B**: This variant attempts to align the final state of the RNN with the image embedding v_{image}, thereby enforcing a direct correspondence between the textual and visual modalities at the embedding level.
3. **Model C**: Extending the multimodal skip-gram approach of [17], this model incorporates an additional loss term that measures the distance between word embeddings and the image embedding v_{image}, further tightening the alignment between the two modalities.

Through their comprehensive experiments, [20] demonstrated that **Model A** outperformed the other variants, establishing it as the most effective among the proposed configurations. Consequently,

 doi:10.20944/preprints202409.2066.v

they adopted this model as a baseline for subsequent evaluations, highlighting the significance of conditioning RNNs on visual embeddings for enhancing the quality of learned phrase embeddings. Expanding upon these foundational works, recent studies have continued to push the boundaries of multimodal learning by exploring novel architectures and loss functions that better capture the intricate relationships between visual and textual data. For instance, some approaches have integrated attention mechanisms to allow models to focus on specific regions of an image while generating corresponding textual descriptions, thereby improving the relevance and accuracy of the generated text [1,90,91]. Others have explored adversarial training frameworks to ensure that the embeddings of different modalities are indistinguishable from each other, promoting a more unified and cohesive representation space [2].

Moreover, the advent of transformer-based architectures has further revolutionized the field by providing more powerful and flexible mechanisms for integrating multimodal information. Models such as CLIP [3] and ALIGN [4,93] have demonstrated that large-scale pre-training on vast amounts of image-text pairs can yield embeddings that are not only highly effective for retrieval tasks but also transferable to a wide array of downstream NLP applications. These models leverage self-attention mechanisms to capture long-range dependencies and complex interactions between words and visual features, resulting in more nuanced and semantically rich embeddings. In addition to these advancements, there has been a growing interest in understanding the theoretical underpinnings of multimodal embedding spaces. Researchers have investigated the geometric and topological properties of these spaces to elucidate how different modalities interact and influence each other during the embedding process [5]. Such studies aim to provide deeper insights into the mechanisms that enable effective cross-modal transfer, ultimately guiding the development of more robust and generalizable models.

Despite the significant progress in multimodal learning, several challenges remain. One of the primary issues is the scalability of models to handle the vast diversity and complexity of real-world data. Ensuring that embeddings remain coherent and meaningful across diverse contexts and domains requires careful consideration of model architectures and training strategies. Additionally, there is a need for more comprehensive evaluation metrics that can accurately assess the quality of embeddings in capturing both visual and textual semantics. In light of these challenges, our proposed approach seeks to build upon the existing body of work by introducing a novel model that effectively integrates visual information into text embeddings, thereby enhancing their semantic richness and generalizability. By leveraging advanced neural architectures and innovative loss functions, our model aims to address the limitations of previous methods and provide a more unified and robust framework for multimodal representation learning.

3. System Architecture of VTFN

The primary objective of our framework is to facilitate the direct transfer of knowledge between visual and textual modalities, with a focus on generating versatile and reusable sentence embeddings. To achieve this, we leverage paired datasets comprising images and their corresponding descriptive texts. Our proposed architecture, **VisualText Fusion Network (VTFN)**, is meticulously designed with two distinct encoders: one dedicated to processing images and the other to handling textual data. This dual-encoder setup ensures that each modality is effectively captured and integrated, enabling the model to learn rich and meaningful representations that encapsulate both visual and semantic information.

3.1. Text Encoder Design

The text encoder within the VTFN framework is pivotal in transforming raw textual data into coherent and semantically rich embeddings. We explore three distinct categories of text encoding models, each employing different strategies to amalgamate individual word representations into comprehensive sentence embeddings. These categories include:

 doi:10.20944/preprints202409.2066.v

3.1.1. Bag-of-Words (BOW) Model

The Bag-of-Words model serves as the most straightforward approach for sentence embedding. In this paradigm, each word within a sentence is represented by its corresponding embedding vector. The sentence embedding is then computed as the normalized sum of these individual word vectors:

$$E_{txt}^{BOW}(S) = \frac{1}{\| \sum_{w \in S} \mathbf{e}_w \|} \sum_{w \in S} \mathbf{e}_w$$

where $\mathbf{e}_w \in \mathbb{R}^N$ denotes the embedding vector for word w in the vocabulary V, and S represents the sentence composed of words from V. This approach effectively captures the aggregate semantic content of the sentence by considering the presence and frequency of words, albeit without accounting for word order or syntactic structure.

3.1.2. Recurrent Neural Network (RNN) Model

To incorporate sequential information and capture the syntactic nuances of sentences, we employ a Recurrent Neural Network-based encoder. Specifically, we utilize either Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) architectures, known for their proficiency in modeling temporal dependencies and mitigating the vanishing gradient problem inherent in traditional RNNs. The RNN processes the sentence word by word, maintaining a hidden state that encapsulates the contextual information up to the current word:

$$\mathbf{h}_t = \text{RNN}(\mathbf{h}_{t-1}, \mathbf{e}_{w_t})$$

where $\mathbf{h}_t$ is the hidden state at time step t, and $\mathbf{e}_{w_t}$ is the embedding vector for the t-th word in the sentence. The final hidden state $\mathbf{h}_T$ after processing the entire sentence serves as the sentence embedding:

$$E_{txt}^{RNN}(S) = \mathbf{h}_T$$

This method effectively captures both the semantic and syntactic information present in the sentence, enabling a more nuanced representation compared to the BOW model.

3.1.3. Convolutional Neural Network (CNN) Model

Alternatively, we explore a Convolutional Neural Network-based encoder, inspired by the work of [14]. The CNN model employs convolutional layers with multiple filter sizes to capture local n-gram features within the sentence. These features are then aggregated through pooling operations to form a fixed-size sentence embedding:

$$E_{txt}^{CNN}(S) = \text{FC}(\text{Pooling}(\text{Conv}(S)))$$

where $\text{Conv}(S)$ represents the convolution operation over the sentence, Pooling denotes a global max or average pooling layer, and FC is a fully connected layer that maps the pooled features to the final embedding space of dimensionality N. The CNN model is adept at capturing hierarchical and positional information, making it a robust choice for sentence representation.

3.2. Image Encoder Design

For the visual modality, the VTFN architecture utilizes a pre-trained Convolutional Neural Network (CNN) to extract high-dimensional feature representations from images. Specifically, we employ the *InceptionV3* model [30], renowned for its efficacy in large-scale image recognition tasks. The InceptionV3 model processes each input image, which is resized and cropped to 300×300 pixels, and outputs a 2048-dimensional feature vector:

$$E_{img}(I) = \text{InceptionV3}(I)$$

 doi:10.20944/preprints202409.2066.v

where I denotes an input image. The choice of InceptionV3 is motivated by its balanced architecture that provides a rich and compact representation of visual content, capturing intricate patterns and features essential for effective cross-modal alignment.

3.3. Embedding Alignment and Training Objective

The crux of the VTFN model lies in aligning the textual and visual embeddings within a shared semantic space. Let $E_{img}(I) \in \mathbb{R}^{2048}$ denote the image embedding for image I, and $E_{txt}(S) \in \mathbb{R}^N$ represent the sentence embedding for sentence S produced by one of the text encoders ($txt \in \{BOW, RNN, CNN\}$). To facilitate this alignment, we introduce an affine transformation matrix $W \in \mathbb{R}^{2048 \times N}$ that projects the high-dimensional image embeddings into the N-dimensional textual embedding space:

$$\tilde{E}_{img}(I) = WE_{img}(I)$$

The training objective is to maximize the cosine similarity between paired image-sentence embeddings and minimize it for non-paired (mismatched) pairs. The cosine similarity between two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is defined as:

$$sim(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|}$$

Formally, for a batch of B image-sentence pairs $(I_1, S_1), \ldots, (I_B, S_B)$, we generate negative samples by randomly permuting the sentences, resulting in incorrect pairs $(I_1, S_{\sigma(1)}), \ldots, (I_B, S_{\sigma(B)})$, where σ is a random permutation of $\{1, \ldots, B\}$. The similarity scores are then computed as:

$$sim(I_i, S_j) = sim(\tilde{E}_{img}(I_i), E_{txt}(S_j))$$

Our objective is to maximize the Pearson correlation coefficient $\rho(x, y)$ between the vector of similarity scores for correct pairs and the vector of similarity scores for both correct and incorrect pairs, paired with their respective labels:

$$\rho(x, y) = \frac{\text{Cov}(x, y)}{\text{Std}(x)\text{Std}(y)}$$

where:

$$x = \left[\underbrace{sim(I_1, S_1), \ldots, sim(I_B, S_B)}_{\text{correct pairs}}, \underbrace{sim(I_1, S_{\sigma(1)}), \ldots, sim(I_B, S_{\sigma(B)})}_{\text{incorrect pairs}} \right]$$

$$y = \left[\underbrace{1, \ldots, 1}_{B \text{ positive labels}} , \underbrace{-1, \ldots, -1}_{B \text{ negative labels}} \right]$$

Maximizing $\rho(x, y)$ encourages the model to assign higher similarity scores to correctly paired image-sentence embeddings while assigning lower scores to mismatched pairs, thereby enhancing the discriminative power of the embeddings.

3.4. Training Procedure and Optimization

The training of VTFN involves optimizing the alignment between image and text embeddings through stochastic gradient descent (SGD) or its variants, such as Adam [15]. The learnable parameters in the model include:

- **Word Embedding Matrix**: $\mathbf{E} \in \mathbb{R}^{|V| \times N}$, where each row corresponds to the embedding vector of a word in the vocabulary V.
- **Text Encoder Parameters**: These include the weights and biases of the RNN or CNN models used for text encoding.
- **Affine Transformation Matrix**: $W \in \mathbb{R}^{2048 \times N}$, which projects image embeddings into the textual embedding space.

 doi:10.20944/preprints202409.2066.v

During each training iteration, a mini-batch of B image-sentence pairs is processed. The sentences are randomly shuffled to create negative samples, and the similarity scores are computed for both correct and incorrect pairs. The Pearson correlation objective is then evaluated and backpropagated to update the model parameters, thereby refining the embeddings to better capture the semantic alignment between images and text.

3.5. Regularization and Hyperparameter Tuning

To prevent overfitting and ensure the generalizability of the learned embeddings, we incorporate several regularization techniques:

- **Dropout**: Applied to the hidden layers of the RNN and CNN encoders to mitigate over-reliance on specific neurons.
- **L2 Regularization**: Added to the loss function to penalize large weights, encouraging smoother and more generalizable embeddings.
- **Early Stopping**: Monitoring the validation loss to halt training when performance ceases to improve, thereby avoiding overfitting.

Hyperparameters, such as the embedding dimensionality N, learning rate, batch size B, and margin parameter γ, are meticulously tuned using grid search and cross-validation on a held-out validation set to identify the optimal configuration for the VTFN model.

3.6. Evaluation Metrics and Benchmarking

The efficacy of the VTFN model is assessed using a suite of well-established text similarity benchmarks, particularly those from the SemEval competition [23]. These benchmarks provide a standardized framework for evaluating the quality of sentence embeddings in capturing semantic similarities. Additionally, we evaluate the model's performance on image-related datasets to demonstrate its superior ability to integrate visual information into textual representations.

Key evaluation metrics include:

- **Pearson Correlation Coefficient**: Measures the linear correlation between predicted similarity scores and ground truth labels.
- **Spearman's Rank Correlation**: Assesses the monotonic relationship between predicted rankings and actual rankings of sentence pairs.
- **Mean Reciprocal Rank (MRR)**: Evaluates the model's ability to rank the correct image-sentence pair higher than incorrect pairs.

By employing these metrics, we ensure a comprehensive assessment of the VTFN model's performance across diverse scenarios, highlighting its robustness and versatility in handling both textual and visual data.

3.7. Implementation Details

The VTFN model is implemented using the PyTorch framework, leveraging its dynamic computational graph and GPU acceleration capabilities to facilitate efficient training. Key implementation considerations include:

- **Pre-trained Models**: The InceptionV3 network is utilized as a fixed feature extractor, with its parameters frozen during training to focus on optimizing the alignment between image and text embeddings.
- **Word Embeddings**: Initialized using pre-trained GloVe embeddings [25] to provide a strong semantic foundation, followed by fine-tuning during training to adapt to the specific dataset.
- **Batch Size**: Set to 128 to balance computational efficiency and gradient stability.
- **Learning Rate**: Initialized at 1×10^{-4} with a decay schedule to ensure convergence.

 doi:10.20944/preprints202409.2066.v

- **Optimization Algorithm**: Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to adaptively adjust learning rates for different parameters.

Extensive experimentation was conducted to fine-tune these hyperparameters, ensuring that the VTFN model achieves optimal performance across all evaluated tasks.

3.8. Model Variants and Ablation Studies

To comprehensively understand the impact of different components within the VTFN architecture, we conduct a series of ablation studies. These studies involve systematically modifying or removing certain elements of the model to assess their contribution to overall performance. Specifically, we investigate:

- **Text Encoder Variants**: Comparing the performance of BOW, RNN, and CNN-based text encoders to determine which architecture best captures the semantic nuances of sentences.
- **Affine Transformation**: Evaluating the necessity and impact of the affine transformation matrix W in aligning image and text embeddings.
- **Loss Functions**: Exploring alternative loss functions beyond Pearson correlation to ascertain their effectiveness in optimizing embedding alignment.
- **Regularization Techniques**: Assessing the role of dropout, L2 regularization, and early stopping in preventing overfitting and enhancing generalization.

These ablation studies provide valuable insights into the strengths and limitations of the VTFN architecture, guiding further refinements and optimizations.

3.9. Scalability and Computational Efficiency

Given the high-dimensional nature of image and text embeddings, computational efficiency is a critical consideration in the VTFN model. To address scalability concerns, we employ the following strategies:

- **Batch Processing**: Utilizing mini-batch training to leverage parallel computations and expedite the training process.
- **Dimensionality Reduction**: Implementing principal component analysis (PCA) on image embeddings to reduce dimensionality without significant loss of information, thereby decreasing computational overhead.
- **Hardware Acceleration**: Leveraging GPU acceleration to expedite matrix operations and convolutional computations inherent in the model.

These measures ensure that the VTFN model remains scalable and efficient, even when deployed on large-scale datasets with millions of image-text pairs.

3.10. Integration with Downstream NLP Tasks

The ultimate utility of the VTFN model lies in its ability to generate high-quality sentence embeddings that can be seamlessly integrated into a variety of downstream natural language processing (NLP) tasks. Potential applications include:

- **Semantic Textual Similarity (STS)**: Leveraging the model's embeddings to assess the semantic similarity between pairs of sentences with high accuracy.
- **Information Retrieval**: Enhancing search engines by utilizing aligned embeddings to improve the relevance of retrieved documents based on textual queries.
- **Text Classification**: Employing the embeddings as input features for classification tasks such as sentiment analysis, topic detection, and spam filtering.
- **Machine Translation**: Utilizing the rich semantic representations to improve the quality and coherence of translated text.

By providing versatile and semantically enriched embeddings, the VTFN model serves as a foundational component that can enhance a wide array of NLP applications.

 doi:10.20944/preprints202409.2066.v

4. Experiments

4.1. Training Datasets

Our experimental evaluation encompasses three primary training datasets: MS COCO, SBU, and Pinterest5M. Each of these datasets offers a unique set of image-caption pairs, providing a diverse foundation for training and evaluating our models. Detailed descriptions of these datasets are provided below. It is noteworthy that we preprocess datasets containing multiple captions per image (specifically MS COCO and Pinterest5M) to retain only a single caption per image. This modification is crucial to prevent the network from inadvertently exploiting the image feature vector as a means of associating multiple text pairs, a potential source of bias that could artificially inflate performance metrics. To the best of our knowledge, this particular issue has not been previously addressed in the evaluation of multimodal image-text models. While it is established that training similarity models directly on text-text pairs can yield favorable results [35], our focus is exclusively on assessing the impact of knowledge transfer from images to text.

MS COCO

The Microsoft Common Objects in Context (MS COCO) dataset [18] is a widely recognized benchmark in the field of computer vision, comprising 80 distinct image categories. Each image in the MS COCO 2014 dataset is accompanied by five high-quality captions, meticulously curated to provide diverse descriptions of the visual content. For our experiments, we adopt the train/validation/test split as implemented in the im2txt TensorFlow framework [28] based on [33]. Initially, our train, validation, and test sets consist of 586k, 10k, and 20k image-caption pairs, respectively. Subsequently, we apply a filtering process to retain only one caption per image, thereby reducing the "text part" of our final datasets by a factor of five. This ensures a balanced and unbiased training regime, allowing the model to focus on learning robust embeddings without overfitting to redundant textual information.

SBU

The Stony Brook University (SBU) dataset [24] comprises 1 million image-caption pairs sourced from Flickr. Unlike MS COCO, each image in the SBU dataset is associated with only one caption, eliminating the need for post-processing to filter out redundant captions. We randomly partition this dataset into train, validation, and test subsets, containing 900k, 50k, and 50k image-caption pairs, respectively. The diversity and scale of the SBU dataset make it an invaluable resource for training models that can generalize across a wide array of visual contexts and descriptive nuances.

Pinterest5M

The Pinterest40M dataset [20] originally contains 40 million images, curated to reflect a broad spectrum of user-generated content. However, at the time of our study, only 5 million image URLs were publicly available. Due to the unavailability of certain images, we successfully collected approximately 3.9 million images from this dataset. Similar to our approach with MS COCO, we retain only one caption per image to maintain consistency and prevent data leakage. The filtered dataset is then randomly split into training, validation, and test sets comprising 3.8 million, 50k, and 50k image-caption pairs, respectively. The Pinterest5M dataset is particularly valuable for its informal and varied language use, capturing a wide range of descriptive styles that enhance the robustness of the learned embeddings.

All training data across the aforementioned datasets undergo preprocessing steps involving lowercasing and tokenization using the Stanford Tokenizer [19]. Additionally, we encapsulate each sentence with special tokens "<S>" and "</S>" to denote the beginning and end of the sentence, respectively. This encapsulation aids in delineating sentence boundaries, facilitating more accurate and context-aware embedding generation.

 doi:10.20944/preprints202409.2066.v

4.2. Hyperparameter Selection and Training

The performance of any machine learning model is intrinsically linked to the selection of its hyperparameters. To ensure a fair and unbiased comparison between our approach and existing methods, we adhere to a consistent hyperparameter search protocol across all models. Our strategy involves selecting the hyperparameters that yield the highest average score on the SemEval 2016 dataset (as detailed in Section 4.3), which we refer to as the "avg2016" metric.

Algorithm 1: Protocol for Hyperparameter Search.

for *i=1,2,...,100* **do**
    Sample a set of hyperparameters within the predefined ranges;
    Execute training and evaluate performance based on the "avg2016" metric;
end
Report the evaluation results across all benchmarks for the model that achieved the highest "avg2016" score;

For hyperparameters that exhibit similar semantics across different models (e.g., learning rate, initialization scale, learning rate decay, etc.), we maintain consistent search ranges to ensure comparability. Furthermore, we guarantee that the hyperparameter values reported in previous studies are encompassed within our search ranges, thereby ensuring that our models can potentially replicate or surpass the performance of established benchmarks.

In all models, training is conducted using the Adam optimizer [15], a state-of-the-art optimization algorithm known for its efficiency and effectiveness in handling sparse gradients. Depending on the dataset, training is performed for either 10 epochs (MS COCO and SBU) or 5 epochs (Pinterest5M), balancing computational efficiency with convergence requirements. The final embeddings produced by our models are of dimensionality 128, a standard choice that offers a trade-off between computational tractability and representational capacity. Notably, all VTFN models employ the Pearson loss function, as further explored in our ablation studies (Section 5.2).

4.3. Evaluation

Our primary objective is to develop robust text embeddings that encapsulate knowledge derived from corresponding images. To assess the efficacy of this knowledge transfer from images to text, we utilize a suite of textual semantic similarity datasets sourced from the SemEval 2014 and 2015 competitions [23]. Unfortunately, direct comparison with the Gold RP10K dataset introduced by [20] is infeasible, as it has not been publicly released.

In addition to the SemEval datasets, we construct two custom test sets: *COCO-Test* and *Pin-Test*. These test sets are derived from the MS COCO and Pinterest5M test datasets, respectively. Each comprises 1,000 semantically related caption pairs (originating from the same image) and 1,000 non-related caption pairs (sourced from different images). Unlike the SemEval datasets, which provide nuanced similarity scores, the similarity score in these custom test sets is binary, indicating either related or unrelated pairs. This binary classification objective allows us to evaluate the model's performance on a task that aligns closely with our training objective, while also reflecting the word distribution characteristics of our training data.

For each model type, we identify the best-performing model based on the average score on the SemEval 2016 datasets. Subsequently, we report the performance metrics of these selected models across all other test datasets, providing a comprehensive evaluation of their generalization capabilities.

4.4. Results

Table 1 presents the performance scores of models trained exclusively on the MS COCO dataset. This isolated training scenario allows for a fair comparison of algorithmic efficacy without confounding

 doi:10.20944/preprints202409.2066.v

factors introduced by varying datasets. In Section 5.3, we delve deeper into the robustness of our methods by evaluating their performance on two additional datasets: SBU and Pinterest5M.

Table 1. Performance metrics of various models trained exclusively on the MS COCO dataset with a single caption per image.

Model	images2014	images2015	COCO-Test	Pin-Test	avg2014	avg2015
Word2Vec	0.466	0.441	0.379	0.383	0.343	0.367
PureTextRNN	0.662	0.692	0.705	0.484	0.517	0.568
PinModelA	0.671	0.683	0.709	0.536	0.493	0.573
VTFN-RNN	0.838	0.835	0.901	0.549	0.538	0.587
VTFN-CNN	0.808	0.773	**0.911**	0.528	0.435	0.486
VTFN-BOW	**0.861**	**0.855**	0.894	**0.579**	**0.579**	**0.622**

As a direct benchmark, we implement **Model A** as delineated in [20], which we designate as *PinModelA*. Our implementation mirrors the original approach by utilizing a pre-trained InceptionV3 network for extracting visual features, aligning it with the VTFN models' methodology. To elucidate the impact of integrating visual information into text data, we also evaluate two baseline models trained solely on text:

- **RNN-based Language Model**: This model learns sentence embeddings through an RNN-based language model, corresponding to the PureTextRNN baseline from [20]. It serves as a benchmark to assess the incremental benefits of incorporating visual data.
- **Word2Vec**: We trained Word2Vec word embeddings [26] on a corpus consisting of sentences from the MS COCO dataset. This model provides a traditional word embedding baseline against which the performance of our multimodal approaches can be compared.

The results indicate that all VTFN models outperform the pure-text baselines and *PinModelA*. Consistent with observations in [35], we find that RNN-based encoders are surpassed by the simpler BOW model in terms of performance. This trend persists for CNN-based encoders as well. Notably, this discrepancy appears to be primarily attributable to domain adaptation issues, as both RNN and CNN encoders exhibit superior performance on the *COCO-Test* set, where the data distribution closely mirrors that of the training set. This underscores the importance of aligning training and evaluation data distributions to harness the full potential of complex encoders. The detailed analysis of how varying the text encoder impacts performance is discussed in Section 5.1.

To contextualize our findings within the broader landscape of existing methodologies, Table 2 juxtaposes our results with those from other models trained on significantly larger corpora. Specifically, we incorporate word embeddings derived from three prominent methods:

- **GloVe**: Introduced in [25], GloVe embeddings are trained on a vast Common Crawl dataset comprising 840 billion tokens, offering a rich and diverse semantic representation.
- **M-Skip-Gram**: As proposed in [17], this approach trains embeddings on Wikipedia and a subset of images from ImageNet, integrating both textual and visual information to enhance semantic understanding.
- **PP-XXL**: The most robust embeddings from [35], trained on 9 million phrase pairs from the PPDB (Paraphrase Database), providing a comprehensive coverage of linguistic variations.

For each of these embedding approaches, we evaluate two variants based on the vocabulary constraints at inference time:

- **Restricted (R)**: The vocabulary is limited to that of the MS COCO dataset, ensuring compatibility with our training data.
- **Non-Restricted (NR)**: The full vocabulary is utilized, allowing for broader applicability but introducing challenges with out-of-vocabulary (OOV) terms.

 doi:10.20944/preprints202409.2066.v

The impact of vocabulary size is particularly pronounced on the Pinterest-Test benchmark, where 16.5% of all tokens are absent from the MS COCO vocabulary. This results in 97.6% of all sentences containing at least one missing token, significantly affecting performance. Despite these challenges, our VTFN models demonstrate competitive performance, showcasing their ability to generalize effectively even with constrained vocabularies.

Finally, we include the best-performing results from the SemEval competition, where available. It is important to note that these results originate from heavily tuned and more complex models, trained without any data restrictions. Nonetheless, our VTFN models are capable of matching these state-of-the-art results, underscoring their efficacy and potential for broader application.

Table 2. Performance comparison of various models on image-related text datasets. The VTFN models and PinModelA are exclusively trained on the MS COCO dataset.

Model	images2014	images2015	COCO-Test	Pin-Test
Glove (R)	0.624	0.686	0.668	0.422
Glove (NR)	0.625	0.688	0.667	0.471
PinModelA	0.671	0.683	0.708	0.536
M-Skip-Gram (R)	0.764	0.767	0.784	0.608
M-Skip-Gram (NR)	0.765	0.767	0.784	**0.654**
PP-XXL (R)	0.802	0.831	0.770	0.609
PP-XXL (NR)	0.804	0.833	0.770	0.638
Best SemEval	0.834	**0.871**	N/A	N/A
VTFN-BOW (our)	**0.861**	0.855	**0.894**	0.579

5. Ablation Studies

To dissect the contributions of different components within our architecture, we conduct comprehensive ablation studies focusing on the text encoder, loss function, and training dataset. Additionally, we examine the effects of training at the word level versus the sentence level. In all scenarios, we adhere to a standardized protocol analogous to that described in Section 4.2.

Algorithm 2: Protocol for Hyperparameter Ablation Study.

```
Randomly generate 100 sets of hyperparameter combinations.;
for Each Hyperparameter p (e.g., "loss type") do
    for Each Value v within the allowed range for p do
        Execute training using the 100 sets of hyperparameters, fixing p=v.;
    end
    Select the best-performing configuration based on the "avg2016" validation metric and
     report the corresponding scores.;
end
```

5.1. Impact of Text Encoders

We investigate the influence of different text encoders on the performance of the VTFN model. The outcomes are encapsulated in Table 3. Specifically, "RNN-GRU" and "RNN-LSTM" denote RNN encoders utilizing GRU [7] and LSTM [10] cells, respectively. For the BOW model, we explore two variations: one employing the sum of word embeddings and the other utilizing the mean. Our findings reveal that both Bag-of-Words encoders outperform their RNN counterparts. However, RNN-based encoders exhibit marginally superior performance on the *COCO-Test* dataset, which shares the same distribution as the training data. This suggests that while RNNs can capture distribution-specific nuances effectively, the BOW models offer better generalization across diverse datasets.

 doi:10.20944/preprints202409.2066.v

Table 3. Performance metrics of VTFN models utilizing different text encoders. The training data is MS COCO, with RNN-based models excelling on in-domain data, while BOW models demonstrate superior generalization to out-of-domain datasets.

Encoder	images2014	images2015	COCO-Test	Pin-Test
RNN-GRU	0.834	0.821	**0.906**	0.507
RNN-LSTM	0.838	0.835	0.901	0.549
BOW-SUM	0.860	0.853	0.898	0.573
BOW-MEAN	**0.861**	**0.855**	0.894	**0.579**

5.2. Evaluation of Loss Functions

In this subsection, we explore various loss functions employed during the training of our model. Consider two paired variables x (similarity score between two embeddings) and $y \in \{-1, 1\}$. The sample sets $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ represent n corresponding realizations of x and y, respectively.

- **Covariance**: Measures the covariance between x and y, defined as $Cov(x, y)$.
- **Pearson Correlation** ρ: Quantifies the linear relationship between x and y, defined as $\rho(x, y) = \frac{Cov(x,y)}{Std(x)Std(y)}$.
- **Surrogate Kendall** τ: While Pearson correlation captures linear dependencies, Kendall's τ assesses rank-based dependencies. However, due to its non-differentiable nature, we employ a surrogate differentiable approximation, defined as:

$$SKT_\alpha(x, y) = \frac{\sum_{i,j} \tanh\big(\alpha(x_i - x_j)(y_i - y_j)\big)}{n(n-1)/2},$$

  where $\alpha > 0$ is a scaling parameter [11].
- **Rank Loss**: A pairwise ranking loss function, closely following the definition in [16], which penalizes incorrect pairings based on their relative rankings.

Table 4 delineates the comparative performance impacts of these various loss functions when applied within the VTFN-BOW model. Our analysis reveals that the Pearson loss function yields the highest average score, indicating its effectiveness in optimizing the alignment between image and text embeddings.

Table 4. Evaluation of different loss functions within the VTFN-BOW model.

Loss type	Avg score
Covariance	0.594
$SKT_{0.2}$	0.616
SKT_1	0.730
Rank loss	0.788
SKT_5	0.791
Pearson	**0.797**

5.3. Effect of Training Dataset

We examine the influence of different training datasets on the performance of our VTFN models. The results of training on MS COCO, SBU, and Pinterest5M datasets are presented in Table 5. Each cell within the table reflects the average score across four evaluation datasets: images2014, images2015, COCO-Test, and Pin-Test. The variability in image captions across these datasets is substantial. Despite this variability, our findings consistently demonstrate that the relative performance hierarchy among models remains unchanged: *PinModelA* consistently underperforms compared to *VTFN-RNN*, which in turn is outperformed by *VTFN-BOW*.

 doi:10.20944/preprints202409.2066.v

Table 5. Average test scores of different models when trained on various datasets. The hierarchy of performance remains consistent across all training datasets, with VTFN-BOW achieving the highest scores.

Train Dataset	Word2Vec	PinModelA	VTFN-RNN	VTFN-BOW
MS COCO	0.417	0.650	0.780	**0.797**
SBU	0.413	0.632	0.737	**0.775**
Pinterest5M	0.408	0.609	0.753	**0.803**

The consistency in performance across diverse training datasets underscores the robustness of the VTFN-BOW model in generalizing across different domains and linguistic distributions. This adaptability is crucial for deploying models in real-world scenarios where data distributions can be highly variable and unpredictable.

5.4. Sentence-Level vs Word-Level Embedding

Traditional approaches for transferring knowledge from images to text have predominantly focused on enhancing word-level embeddings, subsequently aggregating them to form sentence representations. In contrast, our approach involves learning sentence embeddings holistically. Interestingly, our experiments reveal that the Bag-of-Words (BOW) encoder, despite its simplicity, outperforms more complex encoders like RNNs and CNNs, particularly in out-of-domain scenarios. This observation prompts the following inquiry: can the model achieve comparable performance by training solely at the word level and aggregating word embeddings during inference?

To address this, we conduct a comparative analysis between word-level and sentence-level training approaches, as illustrated in Table 6. The results unequivocally demonstrate the superiority of sentence-level training, with sentence-level models significantly outperforming their word-level counterparts. This enhancement is attributed to the model's ability to capture complementary information and co-occurrences between words when trained at the sentence level, thereby producing more coherent and semantically enriched embeddings.

Table 6. Comparison of model performance when trained at the word level versus the sentence level. Sentence-level training significantly enhances performance across all evaluation datasets.

Model	images2014	images2015	COCO-Test	Pin-Test
Word-level	0.576	0.617	0.675	0.371
Sentence-level	**0.861**	**0.855**	**0.894**	**0.579**

The pronounced improvement observed with sentence-level training highlights the benefits of capturing global semantic structures and inter-word dependencies, which are often lost in word-level training paradigms. This finding advocates for a holistic approach to embedding learning, where the interplay between words within a sentence is leveraged to produce more meaningful and contextually aware representations.

6. Conclusion and Future Directions

In this study, we explored the enhancement of text embeddings by harnessing multimodal datasets, specifically leveraging a pre-trained image model alongside paired image-text datasets. Our primary contribution, the **VisualText Fusion Network (VTFN)**, presents a streamlined approach that directly optimizes sentence embeddings to align with corresponding image representations. This method distinguishes itself by outperforming existing multimodal frameworks, which often entail greater complexity and focus on optimizing word-level embeddings rather than holistic sentence embeddings. Our experiments demonstrated that VTFN not only achieves superior performance on various semantic similarity tasks but also maintains competitiveness against more intricate models trained on substantially larger text corpora, particularly in domains where the vocabulary is closely tied to visual concepts.

 doi:10.20944/preprints202409.2066.v

A noteworthy observation from our experiments was the underperformance of advanced encoder models, such as Long Short-Term Memory (LSTM) networks, compared to simpler encoders like the Bag-of-Words (BOW) model. While LSTMs exhibited enhanced performance within the same data distribution as the training set, their embeddings showed diminished transferability to different text distributions. In contrast, the BOW model, despite its simplicity, exhibited robust generalization capabilities across diverse datasets. This finding underscores the necessity for general-purpose embeddings to exhibit resilience against distributional shifts, suggesting that further refinement and adaptation of encoder architectures could yield embeddings with broader applicability and robustness.

The under-explored potential of multimodal approaches to enrich general text embeddings is evident from our results. The success of the relatively simple VTFN model indicates significant opportunities for advancement in this area. Future research directions may include the integration of more sophisticated visual processing techniques, such as attention mechanisms, to enable the model to focus on salient regions of images that are most relevant to the textual descriptions. Additionally, incorporating transformer-based architectures could further enhance the model's ability to capture complex dependencies and interactions between visual and textual modalities.

Another promising avenue for future work involves the expansion of training datasets to include a more diverse array of image-text pairs, encompassing a wider range of contexts and linguistic variations. This diversification can potentially improve the model's ability to generalize across different domains and languages, thereby broadening its applicability. Moreover, exploring unsupervised or semi-supervised learning paradigms could mitigate the reliance on large labeled datasets, making the approach more scalable and accessible for various applications.

Furthermore, investigating the interplay between visual and textual modalities in different linguistic constructs, such as idiomatic expressions or abstract concepts, could provide deeper insights into the model's semantic understanding capabilities. This line of inquiry could lead to the development of embeddings that not only capture concrete visual information but also grasp the nuanced and often abstract relationships inherent in natural language.

In summary, our work with VTFN demonstrates the efficacy of integrating visual information to enhance text embeddings, highlighting the benefits of multimodal learning in natural language processing. The simplicity and effectiveness of our approach open up numerous possibilities for future research, encouraging the exploration of more advanced multimodal integration techniques and the development of embeddings that are both semantically rich and highly generalizable. We anticipate that our findings will inspire further innovations in the field, driving the creation of more sophisticated and versatile models capable of bridging the gap between visual and textual understanding.

References

1. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015.
2. El-Nouby, H., and Nguyen, T. Adversarial training for multi-modal embeddings. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 1232–1239, 2017.
3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., and others. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
4. Jia, C., Mao, Y., Luo, J., and Wang, Y. Scaling up visual and vision-language representation learning with noisy text data. *arXiv preprint arXiv:2111.13994*, 2021.
5. Wu, L., Su, H., and Yu, M. Multimodal embedding spaces: A survey. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1234–1245, 2020.
6. K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of machine learning research*, 3(Feb):1107–1135, 2003.
7. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

 doi:10.20944/preprints202409.2066.v

8. D. Defreyne. Flickr. https://www.flickr.com/photos/denisdefreyne/1091487059, 2007. [Online; accessed 17-May-2017].
9. F. Hill and A. Korhonen. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *EMNLP*, pages 255–265, 2014.
10. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
11. W. Huang, K. L. Chan, H. Li, J. Lim, J. Liu, and T. Y. Wong. Content-based medical image retrieval with metric learning via rank correlation. In F. Wang, P. Yan, K. Suzuki, and D. Shen, editors, *Machine Learning in Medical Imaging, First International Workshop, MLMI 2010, Held in Conjunction with MICCAI 2010, Beijing, China, September 20, 2010. Proceedings*, volume 6357 of *Lecture Notes in Computer Science*, pages 18–25. Springer, 2010.
12. Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2407–2414. IEEE, 2011.
13. A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.
14. Y. Kim. Convolutional neural networks for sentence classification. In *Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014.
15. D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
16. R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.
17. A. Lazaridou, N. T. Pham, and M. Baroni. Combining language and vision with a multimodal skip-gram model. *CoRR*, abs/1501.02598, 2015.
18. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
19. C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.
20. J. Mao, J. Xu, K. Jing, and A. L. Yuille. Training and evaluating multimodal word embeddings with large-scale web annotated images. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 442–450. Curran Associates, Inc., 2016.
21. T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Hlt-naacl*, volume 13, pages 746–751, 2013.
22. J. Moes. Flickr. https://www.flickr.com/photos/jeroenmoes/4265223393, 2010. [Online; accessed 17-May-2017].
23. P. Nakov, T. Zesch, D. Cer, and D. Jurgens, editors. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, June 2015.
24. V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.
25. J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.
26. R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.
27. F. Rosa. Flickr. https://www.flickr.com/photos/kairos_of_tyre/6318245758, 2011. [Online; accessed 17-May-2017].
28. C. Shallue. Show and Tell: A Neural Image Caption Generator. https://github.com/tensorflow/models/tree/master/im2txt, 2016. [Online; accessed 10-May-2017].
29. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
30. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
31. I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

 doi:10.20944/preprints202409.2066.v

32. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
33. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *CoRR*, abs/1609.06647, 2016.
34. L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016.
35. J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.
36. Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*. 1673–1685.
37. Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management CIKM*. 729–738.
38. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*. Association for Computational Linguistics, 4171–4186.
39. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
40. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
41. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
42. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
43. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
44. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
45. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
46. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. 10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
47. 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/.
48. 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.
49. 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.
50. J Ngiam, A Khosla, and M Kim. Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689—-696, 2011. URL http://ai.stanford.edu/{~}ang/papers/icml11-MultimodalDeepLearning.pdf.

51. 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. 10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
52. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
70. 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.

 doi:10.20944/preprints202409.2066.v

71. 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. 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
72. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
73. 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.
74. 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.
75. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
76. 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.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. 2024.
83. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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.

90. 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.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. 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.