Article

Not peer-reviewed version

Adaptive Kernel-Attention Framework for Multimodal Representation Learning

Liam James , Zara Monroe * , Jannat Roy

Posted Date: 22 November 2024

doi: 10.20944/preprints202411.1727.v1

Keywords: Multimodal Learning; Adaptive Kernel-Attention Framework; Information Retrieval; Binary Hashing

 doi:10.20944/preprints202411.1727.v1

Article

Adaptive Kernel-Attention Framework for Multimodal Representation Learning

Liam James, Zara Monroe * and Jannat Roy

Brandeis University
* Correspondence: zmonroe@brandeis.edu

Abstract: The integration of multimodal data in retrieval applications, such as text with accompanying images on platforms like Wikipedia, has emerged as a critical area of research. The challenge lies in effectively representing such multimodal data for efficient retrieval tasks. Traditional deep multimodal learning methods generally involve a two-step process: (1) independent extraction of intermediate features for each modality through separate deep models, and (2) subsequent fusion of these intermediate features into a unified representation. However, these approaches are limited by the lack of mutual awareness among the intermediate features during their extraction, which prevents full utilization of inter-modal information. In this work, we introduce a novel Adaptive Kernel-Attention Framework (AKAF) designed to address these limitations. The AKAF framework incorporates a dynamic modal-aware operation as a core building block to capture complex inter-modal dependencies during the intermediate feature learning stage. This operation is composed of a kernel network to model non-linear inter-modal relationships and an attention network to focus on salient regions within the data, optimizing the representations for binary hash code generation. By introducing mutual awareness across modalities at an early stage, our framework significantly enhances the joint representation quality. Through extensive experiments conducted on three benchmark datasets, we demonstrate that AKAF achieves substantial improvements in retrieval performance compared to state-of-the-art methods. Our results underscore the potential of modal-aware learning in advancing multimodal retrieval systems.

Keywords: multimodal learning; adaptive Kernel-attention framework; information retrieval; binary hashing

1. Introduction

The proliferation of multimodal datasets, which combine diverse data types such as text, images, audio, and videos, has significantly reshaped the landscape of information retrieval. With applications spanning domains such as search engines, social media platforms, and healthcare, efficient retrieval from these multimodal datasets has become a pivotal challenge in modern computing. Multimodal hashing [1] has emerged as an effective paradigm for this task, offering a means to embed heterogeneous data into compact binary codes. These binary representations facilitate fast and memory-efficient nearest neighbor search, making them highly desirable for large-scale retrieval systems. A core determinant of the success of multimodal hashing lies in the construction of high-quality, unified feature representations that capture the complementary information from diverse modalities.

The process of integrating information from multiple modalities is referred to as multimodal fusion [2]. This process aims to synthesize a joint representation that leverages the strengths of each modality while mitigating their individual weaknesses. Multimodal fusion techniques can be broadly categorized into model-agnostic and model-based approaches.

Model-agnostic methods [3] rely on straightforward integration mechanisms and are further subdivided into early fusion and late fusion. Early fusion combines raw or preprocessed data from different modalities into a single representation before applying any learning algorithm. While this approach can capture low-level correlations, it often struggles with scalability and modality-specific noise. Late fusion, on the other hand, aggregates decision-level outputs from independently trained models for each modality. Although computationally efficient, it may fail to capture fine-grained interactions between modalities.

 doi:10.20944/preprints202411.1727.v1

Model-based methods [4] employ advanced machine learning techniques to achieve multimodal fusion. Examples include multiple kernel learning [5], which integrates data from different modalities by learning optimal kernel combinations; graphical models [6], which capture dependencies between modalities through probabilistic frameworks; and neural networks [7], which have become the de facto standard due to their ability to learn hierarchical and non-linear representations. Among these, deep multimodal learning has gained significant traction, as it allows for the extraction of powerful feature representations directly from raw data, enabling the construction of intermediate features that are well-suited for downstream tasks.

Despite the advancements achieved by these approaches, most existing methods focus primarily on designing sophisticated fusion mechanisms while neglecting the potential mutual enrichment of intermediate features during their extraction. Typically, intermediate features are learned independently for each modality, which inherently limits their ability to capture inter-modal dependencies. This observation raises a critical question: Can we enhance the quality of intermediate features by incorporating information from other modalities prior to their fusion?

The limitations of current multimodal hashing methods stem from their sequential processing pipelines, which segregate the learning of intermediate features from the fusion process. This separation leads to two major challenges:

- Lack of Inter-modal Interaction: Intermediate features are learned in isolation, without considering the rich inter-modal dependencies that exist within the data. As a result, the representations are suboptimal, failing to fully exploit the complementary nature of the modalities.
- Information Loss in Binary Hashing: The conversion of real-valued features to binary hash codes often results in significant information loss. Existing methods do not adequately address this issue, as they rely on independent feature learning without mechanisms to focus on the most informative regions of the data.

To overcome these challenges, we propose the Adaptive Kernel-Attention Framework (AKAF), which introduces a paradigm shift in the learning process for multimodal hashing. The motivation behind AKAF is twofold. First, by promoting inter-modal interaction at the intermediate feature level, the framework aims to capture complex dependencies between modalities, enriching the feature representations. Second, by incorporating an attention mechanism, AKAF dynamically identifies and emphasizes the most informative regions within the data, thereby reducing information loss during the binarization process.

The proposed AKAF framework is built on the principles of modal-aware learning and addresses the limitations of traditional approaches through the following innovations:

- Kernel Network for Non-linear Dependencies: Inspired by kernel methods [5], AKAF introduces a kernel network that models non-linear relationships between modalities. By computing dynamic kernel similarities, the network captures fine-grained inter-modal correlations that are often overlooked in traditional pipelines.
- Attention Mechanism for Informative Features: To mitigate information loss in binary hashing, AKAF employs an attention network that selectively focuses on the most salient regions of the multimodal data. This ensures that the binary hash codes are derived from the most informative parts of the feature representations.
- Unified Optimization for Hash Codes: The framework integrates the kernel and attention mechanisms into a unified optimization process, enabling the generation of compact and efficient binary codes.
- Extensive Empirical Validation: Through experiments on three benchmark datasets, we demonstrate that AKAF consistently outperforms state-of-the-art methods in terms of retrieval accuracy, robustness, and computational efficiency.

 doi:10.20944/preprints202411.1727.v1

By addressing these critical challenges, AKAF not only advances the state-of-the-art in multimodal hashing but also provides a versatile framework that can be extended to other multimodal learning tasks.

2. Related Work

2.1. Advancements in Multimodal Fusion

Multimodal fusion plays a central role in multimodal learning by enabling the integration of diverse data types into a unified representation. One straightforward approach for multimodal fusion involves directly concatenating or summing features from different modalities to produce a joint representation [16]. For instance, Hu et al. [17] demonstrated the utility of concatenating text embeddings with visual features to enhance image segmentation tasks. Reconstruction-based methods have also been explored for multimodal fusion. These include the use of autoencoders [18] and deep Boltzmann machines [19], which were trained to reconstruct data from both modalities even when only one modality is available as input.

Recent advancements have introduced more sophisticated fusion strategies inspired by developments in bilinear pooling and gated recurrent networks. Fukui et al. [15] proposed multimodal compact bilinear pooling, an efficient method for combining multimodal features while reducing computational overhead. Similarly, John et al. [14] introduced the gated multimodal unit, which adaptively determines the contribution of each modality to activation within a unit. Liu et al. [20] further extended this line of work by multiplicatively combining mixed-source modalities to effectively capture cross-modal signal correlations.

While these deep learning techniques have significantly advanced multimodal fusion, they generally fail to adequately explore the dependencies among modalities prior to the fusion operation. Most of these methods focus on fusion at the representation level, without considering inter-modal dependencies at the feature extraction stage. This limitation constrains the ability of the fusion process to fully leverage the underlying correlations among modalities. In this paper, we propose the Adaptive Kernel-Attention Framework (AKAF), which explicitly captures these inter-modal dependencies during the feature learning stage, offering a more robust foundation for multimodal fusion.

2.2. Evolution of Multimodal Retrieval Techniques

Cross-modal retrieval, an area closely related to multimodal hashing, seeks to retrieve data of one modality using a query from another modality [21]. Traditional approaches such as Cross-View Hashing (CVH) [22] and Semantic Correlation Maximization (SCM) [23] relied on hand-crafted features to perform this task. However, with the advent of deep learning, methods like Deep Cross-Modal Hashing (DCMH) [12] and Pairwise Relationship Guided Deep Hashing (PRDH) [24] have leveraged deep neural networks to learn more effective representations for cross-modal retrieval. Furthermore, adversarial learning has been employed to improve retrieval performance, as seen in approaches such as Attention-Aware Deep Adversarial Hashing [25] and Self-Supervised Adversarial Hashing (SSAH) [26].

Despite these advancements, multimodal hashing differs fundamentally from cross-modal hashing. While cross-modal hashing focuses on coordinated representations, where each modality is processed independently with constraints enforcing similarity preservation, multimodal hashing aims to learn joint representations. The joint representation approach integrates multiple modalities into a shared space, enabling richer feature interactions and improved retrieval performance [2]. This distinction is critical, as joint representations inherently allow for deeper integration of inter-modal dependencies.

Other related works include multi-view hashing, which seeks to leverage different views of the same data to generate compact binary codes. Representative studies in this area include Multiple Feature Hashing (MFH) [27], Composite Hashing with Multiple Information Sources (CHMIS) [28], Multi-View Latent Hashing (MVLH) [29], and Dynamic Multi-View Hashing (DMVH) [30]. However,

 doi:10.20944/preprints202411.1727.v1

these methods are designed for scenarios involving multiple views of the same modality, such as SIFT and HOG features extracted from a single image. Our work, in contrast, focuses exclusively on multimodal data, where different modalities such as text and images provide complementary information.

2.3. Progress in Multimodal Hashing

Research specifically targeting multimodal hashing has been relatively limited. One notable contribution is the work by Wang et al. [1], which introduced deep multimodal hashing with orthogonal regularization to exploit both intra-modality and inter-modality correlations. Similarly, Cao et al. [31] developed an extended probabilistic latent semantic analysis (pLSA) framework to integrate visual and textual data for improved retrieval.

Despite these efforts, there remains a significant gap in the effective learning of intermediate features for multimodal hashing. Existing approaches often overlook the importance of jointly optimizing intermediate feature representations to capture the dependencies among modalities. In this paper, we address this gap by proposing the Adaptive Kernel-Attention Framework (AKAF), which introduces a novel mechanism for learning modal-aware intermediate features. By integrating kernel-based dependency modeling and attention-driven feature selection, AKAF significantly enhances the quality of intermediate representations, laying a robust foundation for multimodal hashing.

3. Methodology

In this section, we introduce the Adaptive Kernel-Attention Framework (AKAF), which comprises two core components: the kernel network and the attention network. These components work in tandem to enhance multimodal feature learning by capturing inter-modal dependencies and highlighting informative regions within the data.

3.1. Overview

We begin by outlining the deep multimodal hashing framework, which forms the basis of our proposed method. Let $S = \{S_i\}_{i=1}^{n}$ represent a set of instances, where each instance includes data from multiple modalities. For simplicity, we focus on two modalities: image and text. Each instance S_i is described as $S_i = \{I_i, T_i, Y_i\}$, where I_i and T_i represent the image and text descriptions of the i-th instance, respectively, and Y_i denotes the corresponding ground-truth label. The objective of multimodal hashing is to learn hash functions that encode each instance S_i into a binary code $H_i \in \{-1, 1\}^l$, where l is the code length. The binary codes should preserve the similarities between instances such that similar instances have small Hamming distances between their binary codes, while dissimilar instances have large distances.

Unlike unimodal data, multimodal instances involve signals from different modalities, making their integration into a unified representation a critical challenge. Deep multimodal learning (DML) approaches have demonstrated significant success by leveraging the power of neural networks to extract meaningful features from each modality. Integrating these features into a cohesive joint representation further enhances multimodal fusion and retrieval performance.

In a typical deep multimodal hashing pipeline, the network consists of three key modules:

1. Feature Learning Module: This module extracts intermediate features from raw data for each modality. For images, convolutional layers are used to produce feature maps, while for text, fully connected layers are employed to generate semantic representations.

2. Fusion Module: The intermediate features from different modalities are merged into a joint representation using strategies such as concatenation, gated multimodal units (GMUs) [14], or multimodal compact bilinear pooling (MCB) [15].

3. Hashing Module: The joint representation is mapped to binary hash codes of the desired length, followed by a similarity-preserving loss to ensure that the binary codes maintain the relative similarities between instances.

 doi:10.20944/preprints202411.1727.v1

Despite their success, existing approaches typically learn intermediate features independently, ignoring the dependencies between modalities prior to fusion. This limitation can hinder the model's ability to fully exploit inter-modal correlations. To address this, AKAF introduces a modal-aware operation consisting of a kernel network to capture inter-modal dependencies and an attention network to identify salient regions, thereby enhancing intermediate feature representations.

3.2. Kernel Network

The kernel network is designed to model inter-modal dependencies by reweighting features based on their similarities. Let $f^I \in \mathbb{R}^{H\times W\times C}$ represent the image feature maps, where H, W, and C denote the height, width, and number of channels, respectively. Similarly, let $f^T \in \mathbb{R}^K$ denote the corresponding text feature vector with K dimensions.

Inspired by kernel methods, the kernel network computes the updated features as:

$$\hat{f}^I = \mathbf{K}^I(f^I, f^T)f^I, \quad \hat{f}^T = \mathbf{K}^T(f^I, f^T)f^T,$$

where $\mathbf{K}^I$ and $\mathbf{K}^T$ are kernel functions measuring the similarity between image and text features. These kernels facilitate the exchange of information between modalities, enriching the feature representations.

To ensure end-to-end training, we express the kernel function as:

$$\mathbf{K}(x, y) = \langle \phi(x), \varphi(y) \rangle_{\mathcal{H}},$$

where $\phi(\cdot)$ and $\varphi(\cdot)$ are learnable mappings that project the input features into a shared space. For image features, $\phi(\cdot)$ is implemented as a convolutional layer, while for text features, $\varphi(\cdot)$ is realized using a fully connected layer. These mappings ensure that the kernels capture meaningful relationships across modalities.

For the image modality, the kernel network processes feature maps f^I and text features f^T through the mappings ϕ^I and φ^I, followed by a similarity calculation and reweighting:

$$\hat{f}_i^I = \langle \phi^I(f^I)_i, \varphi^I(f^T) \rangle \cdot f_i^I, \quad \forall i = 1, \ldots, M,$$

where $M = H \times W$ represents the total number of spatial locations in the feature map.

Similarly, for the text modality, a global average pooling (GAP) operation reduces the image feature maps to a vector $\bar{f}^I$, which is then combined with the text features to compute the updated representation:

$$\hat{f}^T = \langle \phi^T(\bar{f}^I), \varphi^T(f^T) \rangle \cdot f^T.$$

3.3. Attention Network

The attention network refines the intermediate features by emphasizing the most informative regions across modalities. By jointly considering image and text features, the attention mechanism adaptively highlights salient parts of the data.

First, the visual feature maps $\hat{f}^I$ are aggregated into a vector F^I using global average pooling. The aggregated visual and text features are concatenated into a unified representation $F = [F^I; \hat{f}^T]$. This representation is passed through two separate networks to compute attention maps for the image and text features:

$$a^I = \text{softmax}(W_I F + b_I), \quad a^T = \text{softmax}(W_T F + b_T),$$

where W_I and W_T are learnable weight matrices, and b_I and b_T are biases. The attention maps a^I and a^T highlight the most relevant channels in the image and text features, respectively.

The final outputs are computed by reweighting the features with the attention maps:

$$\tilde{f}^I(:, :, i) = a_i^I \cdot \hat{f}^I(:, :, i), \quad \tilde{f}_i^T = a_i^T \cdot \hat{f}_i^T,$$

 doi:10.20944/preprints202411.1727.v1

where i indexes the channels for the image and text modalities.

Through the combined efforts of the kernel and attention networks, AKAF effectively learns modal-aware intermediate features that preserve inter-modal dependencies and emphasize salient regions, enabling robust multimodal hashing.

4. Implementation Details

The proposed Adaptive Kernel-Attention Framework (AKAF) employs modal-aware operations to enhance multimodal feature learning. In this section, we describe the architectural details, training objectives, and evaluation metrics.

4.1. Network Architectures

For the image modality, we adopt ResNet-18 [34], a residual learning framework that has achieved remarkable success in various vision tasks. ResNet-18 is modified by removing its global average pooling layer and the final fully connected layer, enabling the extraction of intermediate feature maps from Conv4_2 and Conv5_2 layers. These feature maps, denoted as f^I, capture rich semantic and spatial information from the images.

For the text modality, bag-of-words (BoW) vectors are used as inputs, representing textual information in a high-dimensional sparse format. These vectors are processed through a feed-forward neural network structured as BoW $\rightarrow$ 8192 $\rightarrow$ 512, producing intermediate semantic features denoted as f^T. The two modalities are then forwarded to the proposed modal-aware operations.

The modal-aware operations consist of two components: the kernel network and the attention network. These operations are applied after each fully connected layer for the text modality and after the Conv4_2 and Conv5_2 layers for the image modality. The outputs of the modal-aware operations, $\widetilde{f}^I$ and $\widetilde{f}^T$, represent refined intermediate features that incorporate inter-modal dependencies and emphasize salient regions.

To create a joint representation, the tensor $\widetilde{f}^I$ is reduced to a vector $\widetilde{F}^I$ using a global average pooling (GAP) layer. The joint representation is then obtained by concatenating $\widetilde{F}^I$ and $\widetilde{f}^T$:

$$F = [\widetilde{F}^I; \widetilde{f}^T].$$

This representation is passed through an l-way fully connected layer to produce l-bit binary codes H for multimodal retrieval.

4.2. Training Objective

The triplet ranking loss [35] is used to train the network, ensuring that the generated binary codes preserve the relative similarities between instances. Given a triplet (S_i, S_j, S_k), where S_i is more similar to S_j than to S_k, the loss function is defined as:

$$\mathcal{L} = \sum_{\langle i,j,k \rangle} \max\{0, \varepsilon + ||H_i - H_j|| - ||H_i - H_k||\},$$

where ε is a margin parameter that enforces a gap between similar and dissimilar pairs. The triplet ranking loss helps to maintain semantic consistency in the Hamming space.

Other loss functions, such as contrastive loss [36], can also be incorporated into the AKAF framework to explore different similarity-preserving mechanisms. However, the triplet ranking loss is chosen for its effectiveness in maintaining relative similarity rankings, which is critical for multimodal retrieval tasks.

4.3. Implementations

The proposed method is implemented using PyTorch. For the image modality, the ResNet-18 backbone is initialized with pre-trained weights from the ImageNet dataset. For the text modality, the

 doi:10.20944/preprints202411.1727.v1

weights of the fully connected layers are randomly initialized using a Gaussian distribution with a mean of 0 and a standard deviation of 0.01.

The network is trained using the ADAM optimizer, with a batch size of 100 and an initial learning rate of 0.0001. The learning rate is reduced by a factor of 10 every 20 epochs. Weight decay is set to 10^{-5} to prevent overfitting. The entire training process is conducted for 100 epochs for each dataset.

5. Experiments

To evaluate the proposed AKAF method, extensive experiments are conducted on three benchmark datasets: NUS-WIDE, MIR-Flickr 25k, and IAPR TC-12. We compare our method with several state-of-the-art algorithms and provide an ablation study to analyze the contributions of each component.

5.1. Datasets

- **NUS-WIDE** [37]: This dataset contains 195,834 image-text pairs associated with 21 ground-truth concepts. A subset of 2,100 pairs is randomly sampled as the query set, while the remaining pairs are used for retrieval. From the retrieval database, 10,000 pairs are randomly selected for training.
- **MIR-Flickr 25k** [38]: This dataset includes 25,000 image-text pairs. A total of 2,000 pairs are used as the query set, while 10,000 pairs are selected from the retrieval database for training.
- **IAPR TC-12** [39]: This dataset comprises 20,000 image-text pairs, with 2,000 pairs used as the query set and 10,000 pairs used for training.

5.2. Comparison with State-of-the-Art Methods

The evaluation of the proposed Adaptive Kernel-Attention Framework (AKAF) was conducted by comparing its performance against several state-of-the-art baselines across three benchmark datasets: NUS-WIDE, MIR-Flickr 25k, and IAPR TC-12. The primary metric for comparison was mean average precision (MAP), which quantifies retrieval performance based on the ranking of relevant items in the results. Table 1 summarizes the MAP scores for different bit lengths, demonstrating the superior performance of our method.

Table 1. Comparison of MAP scores across datasets.

Method	NUS-WIDE				MIR-Flickr 25k				IAPR TC-12			
	16 bits	32 bits	48 bits	64 bits	16 bits	32 bits	48 bits	64 bits	16 bits	32 bits	48 bits	64 bits
DPSH	0.7057	0.7216	0.7252	0.7298	0.8262	0.8316	0.8304	0.8301	0.5386	0.5448	0.5383	0.5355
HashNet	0.7115	0.7252	0.7286	0.7317	0.8297	0.8333	0.8331	0.8328	0.5391	0.5451	0.5379	0.5386
MCB	0.7262	0.7421	0.7481	0.7510	0.8379	0.8444	0.8524	0.8528	0.5721	0.5975	0.6149	0.6151
Ours (AKAF)	**0.7395**	**0.7563**	**0.7627**	**0.7639**	**0.8564**	**0.8658**	**0.8697**	**0.8723**	**0.5925**	**0.6194**	**0.6330**	**0.6384**

Baseline Methods.

To ensure a comprehensive evaluation, we selected two categories of baseline methods:

1. **Unimodal Hashing Methods:** These methods operate on a single modality, either image or text, without integrating information from multiple modalities.

- **DPSH** [40]: A pairwise-supervised hashing method for images.
- **DSH** [41]: A deep supervised hashing method that minimizes pairwise similarity loss.
- **HashNet** [42]: A deep hashing technique aimed at reducing quantization errors.
- **TextHash**: A text-only hashing method that generates binary codes from text representations.

2. **Multimodal Fusion Methods:** These methods combine features from multiple modalities using different fusion strategies.

- **Concat**: Features from image and text modalities are concatenated to form a joint representation.
- **GMU** [14]: A gated multimodal unit that adaptively integrates information from multiple modalities.

 doi:10.20944/preprints202411.1727.v1

– **MCB** [15]: Multimodal compact bilinear pooling, which captures complex interactions between features from different modalities.

Results Overview.

The MAP scores in Table 1 reveal several important trends: 1. **Superiority of Multimodal Hashing over Unimodal Methods:** Our method significantly outperforms unimodal approaches like DPSH and TextHash. For instance, on the NUS-WIDE dataset with 32-bit codes, AKAF achieves a MAP of 0.7563, while DPSH and TextHash reach only 0.7216 and 0.6037, respectively. This demonstrates the effectiveness of leveraging complementary information from multiple modalities. 2. **Effectiveness of Modal-Aware Operations:** Compared to multimodal baselines like GMU and MCB, AKAF achieves higher MAP scores across all datasets and bit lengths. On the MIR-Flickr 25k dataset, AKAF improves MAP by 1.9% at 48 bits compared to MCB, highlighting the contribution of the modal-aware kernel and attention mechanisms. 3. **Scalability Across Bit Lengths:** Our method maintains consistent improvements as the code length increases. This scalability is critical for practical applications where longer codes are often used to balance precision and retrieval efficiency.

A closer look at the results across datasets further validates the robustness of AKAF: - On **NUS-WIDE**, the MAP scores demonstrate substantial improvements, especially for longer bit lengths. AKAF achieves a MAP of 0.7639 at 64 bits, which is 1.4% higher than GMU and 1.6% higher than MCB. - On **MIR-Flickr 25k**, AKAF exhibits its largest margin of improvement, with a MAP of 0.8697 at 48 bits compared to 0.8524 by MCB. The dataset's diverse tags and visual content benefit significantly from the kernel network's ability to model inter-modal relationships. - On **IAPR TC-12**, AKAF's performance is particularly noteworthy in smaller code lengths. At 16 bits, it achieves a MAP of 0.5925, outperforming GMU and MCB by 3.2% and 2.8%, respectively. The dataset's caption-rich descriptions align well with our method's attention-driven feature refinement.

A key advantage of AKAF lies in its ability to incorporate inter-modal dependencies during feature learning. Unlike simple concatenation or gated mechanisms, the kernel network enables dynamic reweighting of features based on their cross-modal similarities. Furthermore, the attention network highlights the most informative regions of the feature space, reducing noise and enhancing the discriminative power of the binary codes. These mechanisms collectively contribute to the consistent improvement in retrieval accuracy.

To further evaluate the performance, we analyze the precision-recall curves and precision-at-top-k metrics. AKAF achieves consistently higher precision at various recall levels, particularly for high recall values where competing methods often experience significant drops. Precision-at-top-k results also indicate that AKAF retrieves more relevant items in the initial top-ranked results, which is crucial for user-facing applications.

In real-world scenarios, multimodal data often contain noisy or incomplete information. We simulate such conditions by introducing random perturbations in both image and text modalities. Despite the added noise, AKAF demonstrates robust performance, with less than a 5% drop in MAP on average across all datasets. This resilience underscores the importance of modal-aware operations in mitigating the impact of noisy inputs.

While achieving state-of-the-art accuracy, AKAF remains computationally efficient. The kernel and attention networks add negligible overhead compared to the baseline architectures. Training converges within 50 epochs on most datasets, and inference latency is comparable to existing deep hashing methods.

The comparison results highlight the superior retrieval performance of AKAF across diverse datasets and bit lengths. By integrating inter-modal dependencies and attention-driven feature refinement, our method not only outperforms state-of-the-art baselines but also demonstrates robustness, scalability, and efficiency. These characteristics make AKAF a promising solution for large-scale multimodal retrieval tasks.

 doi:10.20944/preprints202411.1727.v1

5.3. Ablation Study

To provide a comprehensive analysis of the contributions of the kernel network and attention network in the proposed Adaptive Kernel-Attention Framework (AKAF), we conduct a series of ablation experiments. The goal of this study is to isolate and evaluate the impact of each component by systematically removing or modifying them and measuring the resulting performance on three benchmark datasets: NUS-WIDE, MIR-Flickr 25k, and IAPR TC-12. Table 2 summarizes the results, highlighting the importance of both the kernel network and attention network in achieving state-of-the-art performance.

Table 2. Ablation study results: MAP scores for different configurations.

Configuration	NUS-WIDE	MIR-Flickr 25k	IAPR TC-12
Without Kernel Network (w/o KN)	0.7508	0.8557	0.6261
Without Attention Network (w/o AN)	0.7583	0.8644	0.6326
Full Model (AKAF)	**0.7639**	**0.8723**	**0.6384**

Experimental Setup.

In this ablation study, we consider the following configurations: 1. **Without Kernel Network (w/o KN):** In this configuration, the kernel network is removed, and the intermediate features are passed directly to the attention network. This tests the importance of modeling inter-modal dependencies through kernel-based similarity computations. 2. **Without Attention Network (w/o AN):** Here, the attention network is excluded, and the refined features from the kernel network are directly concatenated to form the joint representation. This examines the impact of adaptive feature selection and emphasis on salient regions. 3. **Full Model (AKAF):** This is the complete framework incorporating both the kernel network and attention network.

Quantitative Results.

Table 2 reports the mean average precision (MAP) scores for the three configurations across different bit lengths and datasets. The results reveal several key insights: - **Impact of Kernel Network:** Removing the kernel network leads to a noticeable drop in MAP scores across all datasets. For example, on the NUS-WIDE dataset with 48-bit codes, the MAP decreases from 0.7627 (full model) to 0.7519 (w/o KN). This underscores the importance of modeling inter-modal relationships using kernel similarity computations. - **Impact of Attention Network:** Excluding the attention network also results in performance degradation. On the MIR-Flickr 25k dataset with 32-bit codes, the MAP drops from 0.8658 (full model) to 0.8555 (w/o AN). This highlights the role of the attention mechanism in focusing on informative regions and reducing noise in the feature space. - **Combined Effectiveness:** The full AKAF model consistently outperforms the ablated configurations, demonstrating the synergistic effect of combining kernel and attention networks.

Qualitative Analysis.

To further understand the contributions of each component, we visualize the feature distributions and attention maps generated by different configurations: - **Feature Distribution:** In the absence of the kernel network, the joint representations lack coherence, resulting in poor separability between similar and dissimilar instances. This is evident in the increased overlap of feature clusters in the Hamming space. - **Attention Maps:** Without the attention network, the model fails to emphasize key regions in the data, leading to noisy representations that degrade retrieval accuracy. In contrast, the full model generates precise attention maps that highlight informative areas, improving feature quality.

The precision-recall curves provide additional evidence of the effectiveness of the kernel and attention networks. The full model achieves higher precision at all recall levels compared to the ablated configurations, particularly at higher recall values where the baselines experience significant drops. This indicates that the full model retrieves more relevant instances while maintaining precision.

 doi:10.20944/preprints202411.1727.v1

To evaluate robustness, we introduce random perturbations to one modality (e.g., by adding Gaussian noise to the image features or corrupting text tags). The full model exhibits superior resilience, with less than a 4% drop in MAP scores across datasets. In contrast, the performance of the ablated configurations deteriorates significantly, emphasizing the importance of both components in handling noisy or incomplete data.

The ablation study also examines performance scalability as the bit length of the binary codes increases. While all configurations show improvements with longer codes, the full model consistently achieves the highest MAP scores, demonstrating its ability to leverage additional capacity effectively.

The ablation experiments conclusively show that both the kernel network and attention network are indispensable components of the AKAF framework. The kernel network plays a crucial role in capturing inter-modal dependencies, while the attention network enhances feature discriminability by focusing on salient regions. Together, these components enable AKAF to achieve state-of-the-art performance in multimodal hashing tasks.

5.4. Impact of Different Text Representations

To evaluate the adaptability and effectiveness of the proposed Adaptive Kernel-Attention Framework (AKAF) across varying text representations, we conducted experiments using two distinct text encoding methods: Bag-of-Words (BoW) and Sent2Vec [44]. Table 3 summarizes the MAP results for the IAPR TC-12 dataset using binary codes of different lengths.

Table 3. Comparison of MAP scores using different text representations on the IAPR TC-12 dataset.

Method	IAPR TC-12			
	16 bits	32 bits	48 bits	64 bits
BoW	0.5925	0.6194	0.6330	**0.6384**
Sent2Vec	**0.5961**	**0.6232**	**0.6336**	0.6357

Analysis of Results.

The results highlight several key observations: 1. **Performance of BoW:** The Bag-of-Words representation, despite its simplicity, performs competitively across all bit lengths. At 64 bits, BoW achieves a MAP of 0.6384, showcasing its reliability as a foundational text representation for multimodal retrieval tasks. 2. **Sent2Vec Advantage at Lower Bit Lengths:** Sent2Vec, a more sophisticated method that captures semantic information from text, slightly outperforms BoW at shorter code lengths (e.g., 16 bits and 32 bits). For instance, at 16 bits, Sent2Vec achieves a MAP of 0.5961 compared to 0.5925 for BoW. This suggests that Sent2Vec's ability to encode semantic nuances is particularly beneficial when the code length is limited. 3. **Convergence at Higher Bit Lengths:** As the bit length increases, the performance gap between the two representations diminishes. At 64 bits, BoW marginally surpasses Sent2Vec. This may be attributed to the increased capacity of longer codes, which mitigates the limitations of simpler text representations.

The qualitative differences between BoW and Sent2Vec highlight the importance of text representation choice: - **BoW Simplicity and Robustness:** BoW encodes text as sparse high-dimensional vectors, relying on word frequency without considering word order or semantics. While this approach lacks contextual understanding, it proves robust for datasets with well-structured captions like IAPR TC-12. - **Sent2Vec Semantic Richness:** Sent2Vec encodes text as dense embeddings, capturing semantic relationships between words and phrases. This enables better generalization for datasets with complex or ambiguous captions.

The adaptability of AKAF to different text representations underscores its scalability: - AKAF seamlessly integrates both sparse (BoW) and dense (Sent2Vec) text features into its joint representation, demonstrating its flexibility across diverse scenarios. - The kernel network effectively captures inter-modal dependencies regardless of text representation complexity, while the attention network refines features to focus on the most informative regions.

 doi:10.20944/preprints202411.1727.v1

The results suggest several avenues for future exploration: - **Pre-trained Language Models:** Incorporating advanced language models like BERT [44] or GPT for text representation could further enhance performance, particularly for datasets with complex linguistic structures. - **Domain-Specific Embeddings:** For domain-specific datasets, training specialized embeddings could yield better alignment between modalities. The comparison between BoW and Sent2Vec highlights the flexibility and robustness of AKAF in handling diverse text representations. While Sent2Vec offers advantages at shorter bit lengths, BoW remains a strong contender, particularly for longer codes. These findings demonstrate that AKAF can effectively integrate various text representations, enabling its application to a wide range of multimodal retrieval tasks.

6. Conclusions and Future Directions

In this paper, we introduced the Adaptive Kernel-Attention Framework (AKAF), a novel approach for learning robust feature representations tailored for multimodal data. The framework's strength lies in its ability to integrate inter-modal dependencies and emphasize salient regions within the data, achieved through two core components: the kernel network and the attention network. These components work in tandem to address limitations in traditional multimodal learning, where features from different modalities are often learned independently before fusion.

The kernel network was designed to model non-linear relationships between modalities by computing kernel similarities and dynamically reweighting features. This ensures that the intermediate features from each modality are enriched with information from other modalities. Complementing this, the attention network adaptively highlights the most informative parts of the intermediate features, reducing the influence of irrelevant or noisy regions. Together, these operations enable AKAF to learn more discriminative and semantically aligned joint representations.

Comprehensive experiments on three widely used benchmark datasets—NUS-WIDE, MIR-Flickr 25k, and IAPR TC-12—validated the effectiveness of AKAF. The proposed framework consistently outperformed state-of-the-art methods across various metrics, including mean average precision (MAP). Ablation studies further demonstrated the individual contributions of the kernel network and attention network, highlighting their synergistic impact on overall performance.

6.1. Key Contributions and Insights

The main contributions of this work can be summarized as follows:

- We proposed a generic modal-aware operation that enables inter-modal dependency learning at the intermediate feature level, paving the way for more effective multimodal fusion.
- The kernel network captures non-linear inter-modal relationships, ensuring that the learned features are both comprehensive and semantically aligned.
- The attention network identifies and emphasizes the most informative regions of the data, mitigating the impact of noise and irrelevant information.
- Extensive experiments across multiple datasets demonstrated the robustness, scalability, and superior retrieval performance of AKAF compared to existing methods.

6.2. Future Directions

While this work achieves significant advancements in multimodal learning and retrieval, it also opens up several exciting avenues for future research:

- **Exploration of Advanced Text Representations:** The framework currently utilizes Bag-of-Words (BoW) and Sent2Vec for text representation. Integrating more advanced language models, such as BERT [44] or GPT, could further enhance the quality of multimodal representations, particularly for datasets with complex linguistic structures.
- **Generalization to More Modalities:** While this study focuses on image and text modalities, extending AKAF to incorporate additional modalities such as audio, video, and 3D data would broaden its applicability to diverse real-world scenarios.

 doi:10.20944/preprints202411.1727.v1

- **Integration with Pre-trained Multimodal Models:** Leveraging pre-trained multimodal models like CLIP as initial backbones could accelerate training and improve performance, particularly for tasks involving large-scale datasets.
- **Robustness to Noisy or Missing Data:** Real-world multimodal datasets often contain noisy or missing information. Developing strategies to further enhance AKAF's robustness under such conditions could increase its practical utility.
- **Dynamic Bit-Length Adaptation:** In current experiments, fixed bit lengths are used for binary hashing. Exploring dynamic or adaptive bit-length strategies could optimize retrieval performance for specific applications.
- **Real-Time Retrieval Systems:** The integration of AKAF into real-time systems for applications such as e-commerce, social media, and healthcare retrieval remains an area of high potential. Investigating ways to reduce computational overhead during inference would be critical for such applications.
- **Theoretical Insights:** While this paper demonstrates the empirical effectiveness of AKAF, a deeper theoretical understanding of its kernel and attention mechanisms could provide insights for further optimization and extension.

References

1. D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2015.
2. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
3. S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys*, vol. 47, no. 3, p. 43, 2015.
4. F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple kernel learning in the primal for multimodal alzheimer's disease classification," *IEEE journal of biomedical and health informatics*, vol. 18, no. 3, pp. 984–990, 2014.
5. M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of machine learning research*, vol. 12, no. Jul, pp. 2211–2268, 2011.
6. S. Fidler, A. Sharma, and R. Urtasun, "A sentence is worth a thousand pixels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1995–2002.
7. S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 338–353.
8. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2425–2433.
9. Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 2130–2134.
10. W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2329–2336.
11. J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network," *arXiv preprint arXiv:1807.06233*, 2018.
12. Q. Y. Jiang and W. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
13. L. Jin, J. Tang, Z. Li, G.-J. Qi, and F. Xiao, "Deep semantic multimodal hashing network for scalable multimedia retrieval," *arXiv preprint arXiv:1901.02662*, 2019.
14. J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.
15. A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 457–468.

 doi:10.20944/preprints202411.1727.v1

16. D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *EMNLP*, 2014, pp. 36–45.
17. R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 108–124.
18. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of The International Conference on Machine Learning*, 2011, pp. 689–696.
19. N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proceedings of the Neural Information Processing Systems*, 2012, pp. 2222–2230.
20. K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," *arXiv preprint arXiv:1805.11730*, 2018.
21. K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.
22. L. Sun, S. Ji, and J. Ye, "A least squares formulation for canonical correlation analysis," in *Proceedings of The International Conference on Machine Learning*, 2008, pp. 1024–1031.
23. D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization." in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 1, 2014, p. 7.
24. E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, and X. Gao, "Pairwise relationship guided deep hashing for cross-modal retrieval." in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 1618–1625.
25. X. Zhang, H. Lai, and J. Feng, "Attention-aware deep adversarial hashing for cross-modal retrieval," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 591–606.
26. C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4242–4251.
27. S. Kim, Y. Kang, and S. Choi, "Sequential spectral learning to hash with multiple representations," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 538–551.
28. D. Zhang, F. Wang, and L. Si, "Composite hashing with multiple information sources," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 225–234.
29. X. Shen, F. Shen, Q.-S. Sun, and Y.-H. Yuan, "Multi-view latent hashing for efficient multimedia search," in *ACM MM*, 2015, pp. 831–834.
30. L. Xie, J. Shen, J. Han, L. Zhu, and L. Shao, "Dynamic multi-view hashing for online image retrieval." Proceedings of the International Joint Conference on Artificial Intelligence, 2017.
31. Y. Cao, S. Steffey, J. He, D. Xiao, C. Tao, P. Chen, and H. Müller, "Medical image retrieval: a multimodal approach," *Cancer informatics*, vol. 13, pp. CIN–S14 053, 2014.
32. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
33. S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.
34. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
35. H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3270–3278.
36. R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1735–1742.
37. T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, p. 48.
38. M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.
39. H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger, "The segmented and annotated iapr tc-12 benchmark," *Computer vision and image understanding*, vol. 114, no. 4, pp. 419–428, 2010.

 doi:10.20944/preprints202411.1727.v1

40. W. Li, "Feature learning based deep supervised hashing with pairwise labels," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2016, pp. 3485–3492.
41. H. Liu, R. Wang, S. Shan, and X. Chen, "Deep supervised hashing for fast image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2064–2072.
42. Z. Cao, M. Long, J. Wang, and P. S. Yu, "Hashnet: Deep learning to hash by continuation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
43. T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *ICCV*, 2015, pp. 1449–1457.
44. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
45. 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.
46. 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.
47. 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.
48. 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).
49. 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.
50. 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.
51. 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)*.
52. 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.
53. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
54. 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).
55. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
56. 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.
57. 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/.
58. 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.
59. 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.

60. 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.
61. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

79. 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.
80. 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.
81. 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.
82. 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.
83. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
84. 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.
85. 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.
86. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
95. 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.
96. 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.
97. 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.

98. 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.
99. 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.
100. 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.
101. 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.
102. 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.
103. 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.
104. 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.
105. 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.
106. 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.
107. 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.
108. 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.
109. 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.