ISSN: 2710-0588

Automated Caption-Guided Image Retrieval: A Multi-Positive Contrastive Learning Paradigm

Danyang Cao*, Liang Cheng, Hongbo Zhou

School of Artificial Intelligence and Computer Science, North China University of Technology, Beijing 100144, China

*Corresponding author

Abstract

[Objective] This study aims to improve textual similarity measurement for image retrieval while mitigating the anisotropy of sentence embeddings. [Methods] We propose an image-caption semantic encoder trained with a multi-positive contrastive loss. The conventional contrastive objective is extended to accommodate multiple positive samples, and image captioning is used to generate training data automatically. On this basis, we construct a caption-based image-to-image retrieval framework. [Results] Experiments show that the proposed model outperforms baseline methods on semantic textual similarity (STS) benchmarks and improves the agreement between retrieval results and human semantic judgments. [Limitations] Short captions cannot fully represent the complex semantics, ambiguity, and fine-grained details of an image. [Conclusion] The proposed Multi-Positive Example Contrastive Learning (MPC) model provides more discriminative sentence embeddings and improves semantic similarity measurement in image retrieval.

Keywords

Image-to-Image Retrieval; Semantic Representation; Multi-Positive Contrastive Loss; Contrastive Learning.

1. Introduction

With the rapid development of information technology, the volume of online image data has increased dramatically. Images have become a major medium for information transmission and storage in social media, e-commerce, medical imaging, and security surveillance. Efficiently and accurately retrieving images whose content is similar to that of a query image remains a challenging problem and has made image retrieval a major research topic in computer vision. Image-to-image retrieval plays an important role in e-commerce, medicine, security, and many other fields and is commonly referred to as content-based image retrieval (CBIR). Early CBIR methods converted images into low-level visual representations and performed retrieval using color, texture, and related features [1]. Such methods are limited because they do not capture the high-level semantic information contained in images. Metadata-based CBIR methods instead represent images with textual or other metadata to support semantic retrieval [2]. Text-description-based image retrieval is both simple and practical. With advances in deep learning, researchers have used mature models to generate image metadata automatically and measure the semantic similarity of that metadata [3]. In this study, image captions serve as metadata. Image-captioning technology converts visual semantics from pixel space into textual space, after which the similarity between textual metadata is measured [4]. Extracting robust semantic representations and computing their similarity are therefore central to this task.

A common approach embeds text into a vector space with a deep neural network, producing a high-dimensional distribution of semantic information, as in SBERT [5]. However, previous

ISSN: 2710-0588

studies have shown that SBERT embeddings are anisotropic: their nonuniform distribution causes them to cluster within a narrow region of the vector space [6]. Consequently, cosine similarities are densely concentrated and poorly discriminative. This property makes the embeddings less suitable for semantic similarity measurement and degrades performance on related tasks.

To address this problem and improve sentence embeddings for semantic similarity tasks, we exploit the structure of synonymous caption sets and extend the conventional contrastive loss to a supervised formulation with multiple positive samples. The resulting Multiple Positive Example Contrastive Learning model is denoted as MPC. Compared with SBERT, MPC reduces embedding anisotropy and produces a more uniform distribution that is better suited to semantic similarity measurement. To alleviate the limited scale and high annotation cost of conventional supervised data, we also exploit cross-modal datasets and an image-captioning model to annotate data automatically and construct synonym sets for contrastive training. Experiments on STS tasks demonstrate that MPC reduces anisotropy, while human-consistency experiments show that integrating MPC into an image retrieval system yields results that better agree with human semantic judgments.

The main contributions of this study are as follows:

1) We propose the MPC model based on a multi-positive contrastive loss. Image-captioning technology is used to generate synonymous caption sets, and the contrastive objective is extended to multiple positive samples. The resulting model supports caption-similarity assessment and mitigates sentence-embedding anisotropy.

2) We construct an image retrieval framework that uses automatically generated captions as metadata and combines EKA-BLIP with MPC to perform image-to-image retrieval.

3) We evaluate the model from the perspectives of semantic textual similarity and human consistency in image retrieval, demonstrating the effectiveness of both the model and the retrieval framework.

2. Related Work

2.1. Image Retrieval

Image retrieval is a core research direction in computer vision and information retrieval. Its objective is to retrieve, efficiently and accurately, images related to a query image or query condition from a large-scale database. Traditional methods rely mainly on handcrafted feature descriptors, such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) [7]. Their generalization ability is limited in complex scenes. With the rapid development of deep learning, substantial progress has been made in feature representation, metric learning, cross-modal retrieval, and large-scale retrieval efficiency [8].

Image retrieval systems can be constructed in several ways. In addition to directly using feature vectors as retrieval evidence, metadata-based retrieval has attracted considerable attention. Early methods used manually annotated labels or image descriptions as metadata. Manual annotation, however, is time-consuming, expensive, and difficult to scale [8]. Many existing systems therefore use contextual information as retrieval labels. For example, user tags, comments, and geolocation can support multimodal retrieval in social-media collections [9]. Such auxiliary information can improve performance in specific domains, but it also has important limitations. Many images have no strongly associated user labels, while context scraped automatically from web pages is often inaccurate, weakly related, or even unrelated to the image and therefore cannot reliably reflect image semantics.

Researchers have attempted to construct semantic metadata automatically. Li et al. [10] used an image-captioning model to generate descriptions for surveillance data and retrieved people

ISSN: 2710-0588

in monitored scenes through keyword search. Wei et al. [11] generated dense captions for database images and converted those captions into structured scene graphs. A graph-matching algorithm then measured semantic distances among objects, relations, and attributes to estimate image similarity. Yoon et al. [12] combined scene-graph generation with an image retrieval framework called IRSGS. Their method represents images as compact, structured, symbolic scene graphs and uses graph neural networks to measure graph similarity, producing results that better reflect human perceptions of image similarity.

Although more complex metadata can contain more information, it can also introduce interference. We argue that a concise caption is sufficient to express the principal semantic information required for retrieval. Accordingly, we generate captions automatically as metadata and implement metadata-based image retrieval by comparing caption embeddings.

2.2. Textual Similarity Measurement

Traditional methods for semantic textual similarity rely on literal matching or statistical analysis. Literal-matching approaches such as N-grams focus on character- or word-level overlap and cannot capture semantic meaning. Statistical methods based on vector-space models construct text vectors from term frequencies but likewise neglect semantic representation.

With the development of deep learning, word-embedding and sentence-embedding algorithms have shown strong performance. Mikolov et al. [13] proposed Word2Vec, which learns word representations using the Skip-gram and continuous bag-of-words architectures. Pennington et al. [14] proposed GloVe, which learns word vectors from a global word co-occurrence matrix. These methods established the foundation for subsequent research.

BERT [15] stacks multiple Transformer encoder layers and is pretrained using masked language modeling (MLM) and next-sentence prediction (NSP). These objectives enable the model to capture contextual information, learn relationships among words, and generate high-quality textual representations. BERT has achieved excellent results on many downstream tasks. Liu et al. [16] improved and optimized the BERT pretraining procedure to obtain RoBERTa. Compared with BERT, RoBERTa is trained on a larger unsupervised corpus and uses an optimized MLM objective. Although these changes require more computational resources, they also improve model performance.

Building on this work, Reimers and Gurevych [5] proposed Sentence-BERT (SBERT), which uses a Siamese architecture to improve the efficiency of BERT-based textual similarity computation. Semantic similarity can then be calculated using Euclidean distance or cosine similarity. To address the anisotropy of BERT embeddings, Li et al. proposed BERT-flow, which transforms sentence vectors into a smooth and isotropic Gaussian distribution through an invertible flow. Su et al. [17] attributed anisotropy to the non-orthogonality of the coordinate basis and introduced a simpler whitening transformation that maps vectors onto a standard orthogonal basis, thereby improving representation quality.

Metric learning is one of the key techniques for narrowing the gap between low-level features and high-level semantics. Classical approaches include triplet loss and contrastive loss, which optimize the embedding space so that similar images are closer and dissimilar images are farther apart [18]. More recently, self-supervised and contrastive methods such as MoCo [19] have used large-scale unlabeled data to improve the robustness and discriminative ability of learned representations.

Despite substantial progress in textual similarity measurement, important challenges remain. Further research is needed to exploit data more effectively, reduce anisotropy, and construct semantic embeddings that are better suited to similarity computation. Focusing on the characteristics of image captions, we propose MPC, which improves both the training data and

ISSN: 2710-0588

the loss function to extract caption semantics more effectively and produce readily comparable feature vectors.

3. Methodology

3.1. Contrastive-Learning Data Generation with an Image-Captioning Model

Training data have a decisive effect on deep neural networks, while the cost and limited availability of manually labeled data remain major challenges in supervised learning. Automatic annotation has therefore become a promising solution. We use an image-captioning model to generate sets of synonymous captions automatically, providing the data foundation for model training.

Our objective is to construct a synonymous-text dataset for contrastive learning. Contrastive learning trains a model by bringing semantically related texts closer in the embedding space. Defining and acquiring semantically related text is therefore a central issue. In the standard setting, given two texts A and B with similar meanings, A can be treated as the anchor and B as its positive sample. The definition of positive samples is crucial to the effectiveness of contrastive learning.

Image-captioning datasets are large-scale multimodal resources that contain many images and associated descriptions in a one-to-many relationship. Captions associated with the same image therefore form natural sets of synonymous texts. Specifically, an image I is independently described by M human annotators, producing a set of short captions. Any two captions from the same set can be defined as a semantically similar anchor-positive pair, allowing the dataset to be reorganized into synonym sets for contrastive training.

To reduce the substantial cost of manual annotation, we further explore automatic dataset construction. An image-captioning model generates multiple descriptions for each image, thereby producing synonymous text pairs automatically for subsequent training.

We use COCO [20], Flickr30K [21], and NLVR2 [22] as image sources. MS COCO is a large-scale computer-vision dataset containing more than 120,000 images of everyday scenes across 80 common object categories, including people, animals, vehicles, and household objects. Its scenes are complex, dense, and diverse. Flickr30K contains approximately 31,783 everyday images collected from Flickr and covers natural landscapes, human activities, urban streets, and other scenarios. NLVR2 is designed for visual-reasoning tasks and evaluates the ability to understand complex relationships between images and text; it contains more than 100,000 examples.

We use EKA-BLIP [23] to generate semantically equivalent but linguistically diverse captions. Conventional beam search maintains a fixed-size set of candidate sequences during decoding and repeatedly expands and prunes them to obtain the highest-scoring output. This strategy can become trapped in local optima, produce repetitive or monotonous sequences, and generate candidates that are too similar to one another. Because MPC training requires diverse expressions of the same meaning, conventional beam search is not suitable for constructing synonym sets.

We therefore adopt diverse beam search to generate varied semantic descriptions. This improved decoding strategy combines diversity-promoting constraints, multiple parallel beam groups, and post-decoding reranking. It alleviates repetition and lack of diversity while maintaining high generation quality.

3.2. Multi-Positive Contrastive Loss for Image Captions

Given the synonym-set structure described above, we extend the self-supervised contrastive loss to a supervised formulation with multiple positive samples. Unlike the conventional

ISSN: 2710-0588

objective, the proposed loss jointly minimizes the angular distance among multiple synonymous captions in the embedding space, as illustrated in Figure 1.

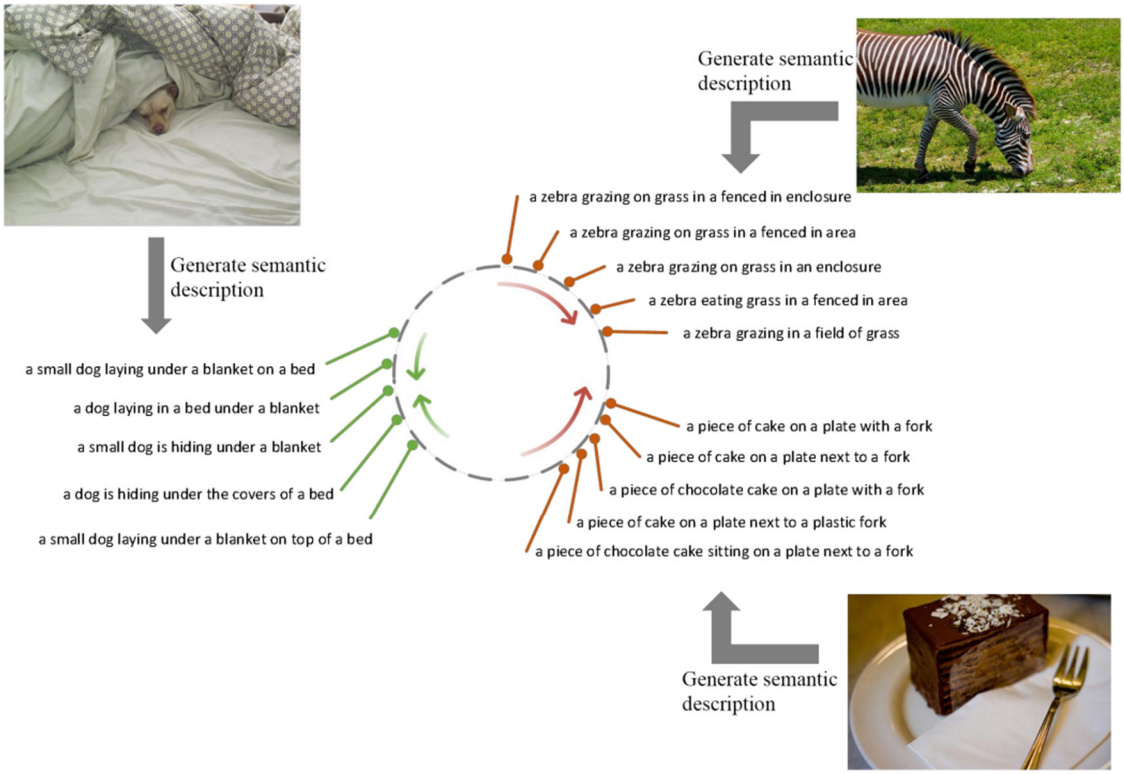

Figure 1. Illustration of Multi-Positive Instance Contrastive Learning Mechanism.

In self-supervised contrastive learning, K randomly sampled instances are used to generate K corresponding positive samples according to a predefined semantic-preservation rule. The training batch therefore contains K anchor-positive pairs. For any anchor, the remaining samples in the batch serve as negatives, and the contrastive loss is defined in Equation (1).

$$\mathcal{L}_{contrastive} = -\sum_{k=1}^{K} log \frac{e^{sim(h_i,h_i^+)/\tau}}{\sum_{k=1}^{K} e^{sim(h_i,h_k)/\tau} - e^{sim(h_i,h_i)/\tau}} \tag{1}$$

A synonym set contains several mutually positive samples and can therefore expand the number of positives available to each anchor. Conventional supervised contrastive learning usually optimizes one anchor-positive pair at a time and cannot fully exploit this structure. We address this limitation by extending the objective to supervised contrastive learning with multiple positives.

We first extend the self-supervised objective to a supervised anchor-positive setting. The supervised dataset consists of anchor-positive pairs. A randomly sampled batch contains K pairs, or 2K samples in total. For any sample, negatives are defined as all batch elements other than the sample itself and the other member of its pair. The resulting supervised contrastive loss is expressed in the following form.

$$\mathcal{L}_{sup\ contrastive} = -\sum_{k \in B} log \frac{e^{sim(h_k,h_{l(k)})/\tau}}{\sum_{j \in \mathbb{N}} e^{sim(h_k,h_j)/\tau}} \tag{2}$$

ISSN: 2710-0588

We then extend the loss to the multi-positive setting based on the synonym-set dataset introduced in Section 3.1. Each synonym set contains M samples. If a randomly sampled batch contains K synonym sets, the batch can be represented as the union of these sets.

Because vector magnitude is strongly affected by linguistic form, the dot product is not an appropriate measure of semantic similarity. We therefore use cosine similarity, as defined below.

$$sim(x, y) = cos(x, y) = \frac{x \cdot y}{|x| \cdot |y|} \tag{3}$$

Each training item is a synonym set containing T elements. Consequently, a randomly sampled batch of K sets contains K multiplied by T caption samples.

For any caption, the positive-index set contains the indices of all other captions from the same synonym set. The negative set contains all captions drawn from the other synonym sets in the batch.

The proposed multi-positive contrastive loss is therefore defined as follows.

$$\mathcal{L}_{mpc} = -\sum_{k \in B} log \frac{\frac{1}{T}\sum_{i \in L_k} e^{sim(h_k, h_i)/\tau}}{\sum_{j \in H_k} e^{sim(h_k, h_j)/\tau}} \tag{4}$$

$$= -\sum_{k \in B} \left\{ log \left[\frac{1}{T} \sum_{i \in L_k} e^{sim(h_k, h_i)/\tau} \right] - log \left[\sum_{j \in H_k} e^{sim(h_k, h_j)/\tau} \right] \right\} \tag{5}$$

To smooth optimization, we also construct entailment positives as auxiliary data and process them with the self-supervised contrastive objective. For each anchor caption, 20% of its words are deleted at random, creating a shortened caption that is treated as an entailment-positive sample.

$$h' = DEL(h) \tag{6}$$

$$\mathcal{L}_e = -\sum_{k \in B} log \frac{e^{sim(h_k, h_k')/\tau}}{\sum_{j \in H_k} e^{sim(h_k, h_j)/\tau}} \tag{7}$$

Here, the random-deletion operator generates an entailment-positive variant of the original caption.

The final objective combines the multi-positive loss and the auxiliary self-supervised loss, with their relative contribution controlled by a hyperparameter.

$$\mathcal{L}_{MPC} = \mathcal{L}_m + \lambda \mathcal{L}_e \tag{8}$$

ISSN: 2710-0588

3.3. Semantic Encoder Architecture

We build the Multiple Positive Example Contrastive Learning model (MPC) on SBERT. MPC uses a Siamese architecture with a 12-layer RoBERTa encoder as the feature extractor; the encoders share parameters. Each caption is encoded independently, and its [CLS] token represents the sentence-level semantic feature. Unlike pairwise cross-encoding with BERT, this architecture allows captions to be encoded in advance and supports efficient semantic-similarity computation.

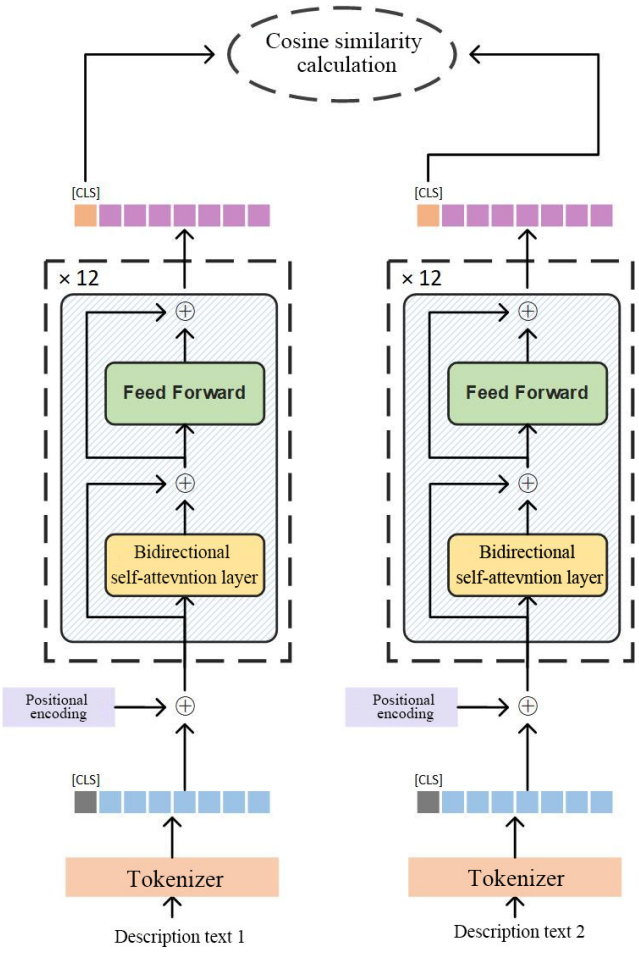

Figure 2. Architecture of the MPC Model.

The MPC architecture is shown in Figure 2. For an input caption, the model first performs tokenization and positional encoding to obtain a token sequence. The sequence is then embedded into the semantic vector space by the multilayer encoder. The resulting sentence representation is used to compute cosine similarity.

3.4. Image-to-Image Retrieval

Semantic image retrieval requires the database images and the query image to be represented semantically before their similarities can be computed. Processing every image online with a deep model is prohibitively expensive for large databases. By using automatically generated captions and their embeddings as intermediate metadata, we separate offline indexing from online querying and construct a two-stage image retrieval framework that explicitly represents image semantics.

ISSN: 2710-0588

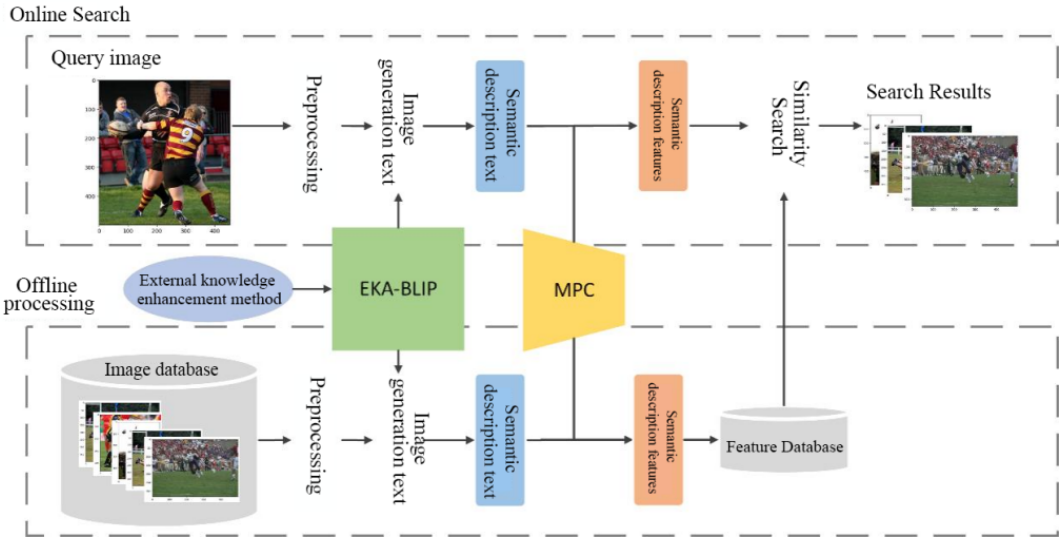

Figure 3. Semantic Description-Based Image Retrieval Framework.

As shown in Figure 3, the framework follows an offline-processing/online-query paradigm. During offline processing, every image in the database is indexed in advance. EKA-BLIP first generates a semantic caption for each uploaded image. MPC then embeds the caption into a semantic feature vector, which is stored in retrieval database E. The construction of the feature index can be represented as follows.

$$t_i = EKABLIP(x_i) \tag{9}$$

$$v_i = NPOS(t_i) \tag{10}$$

$$\mathbb{E} = \{(x_i, v_i) | i = 1, \dots, L\} \tag{11}$$

Here, the variable denotes the total number of images in the database.

During online querying, a query image is processed only once to generate its caption and extract its semantic feature. The resulting vector is compared with all candidate vectors in the precomputed index using cosine similarity. The N most similar vectors are retrieved and reranked by similarity score.

Compared with a one-stage framework that directly embeds images into vectors, the proposed framework uses semantic captions as explicit metadata. It therefore captures image meaning more clearly and accurately and performs better in complex scenes.

4. Experiments and Discussion

4.1. Experimental Settings

All experiments and evaluations were conducted on an NVIDIA GeForce RTX 3090 GPU. Models were implemented and trained using PyTorch and the Transformers library. Parameters were initialized from the SRoBERTa-NLI-base checkpoint. Unless otherwise specified, the initial learning rate was 3e-5 and the batch size was 64. Training lasted four epochs; the model was evaluated on the STS-B development set every 500 steps, and the checkpoint with the highest score was retained.

ISSN: 2710-0588

4.2. Evaluation Datasets and Metrics

We first evaluate the model on standard Semantic Textual Similarity (STS) tasks. Following prior work [4], we use STS 2012-2016 (STS12-STS16), the STS Benchmark (STS-B), and SICK Relatedness (SICKr). Each dataset assigns a score from 0 to 5 to every sentence pair. Unless otherwise stated, STS-B and SICKr refer to their respective test sets. Model performance is measured using Spearman's rank correlation coefficient.

We then integrate MPC into the image retrieval framework and evaluate it on image-to-image retrieval. Caption-based retrieval targets semantic relationships in complex scenes rather than superficial visual similarity or instance-level category matching. Most popular image retrieval datasets are designed for instance-level retrieval and are therefore unsuitable for this setting. We consequently evaluate the framework primarily through human-consistency scores, which measure agreement between model decisions and human semantic judgments.

Evaluation is performed on the Image Semantic Similarity Triplet benchmark [3], which contains 10,712 annotations collected from 21 human annotators through a web-based interface. Each item comprises a query image Q and two candidate images A and B. Annotators choose one of four responses: (a) A is semantically more related to Q; (b) B is semantically more related to Q; (c) A and B are equally related to Q; or (d) neither A nor B is related to Q.

The human-consistency score quantifies agreement between a model decision and human annotations. For an image triplet T, let c1, c2, c3, and c4 denote the numbers of annotators selecting responses (a), (b), (c), and (d), respectively. Once the model selects an answer, the consistency score for that triplet is defined as follows.

$$TScore = \frac{c_i + 0.5c_3}{c_1 + c_2 + c_3 + c_4} \times 100\% \tag{12}$$

Because the decision threshold between similarity scores is ambiguous, the model is restricted to responses (a) and (b). Thus, i belongs to {1, 2}; i = 1 when the model selects (a), and i = 2 when it selects (b).

The overall human-consistency score of the model is then computed as follows.

$$Score = \frac{1}{N}\sum_{j=1}^{N} TScore_j \tag{13}$$

Here, N is the total number of triplets in the dataset.

4.3. Results on STS Tasks

To evaluate MPC on textual similarity measurement, we conduct experiments on STS12-STS16, STS-B, and SICKr.

Table 1. Experimental Results of Semantic Textual Similarity Tasks (STS12 - STS16).

Method	STS12	STS13	STS14	STS15	STS16
InferSent-GloVe	52.86	66.75	62.15	72.77	66.87
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18
SBERT-flow	69.78	77.27	74.35	82.01	77.46
SRoBERTa-whitening	70.46	77.07	74.46	81.64	76.43
SBERT	70.97	76.53	73.19	79.09	74.30
SRoBERTa	71.54	72.49	70.80	78.74	73.69
MPC (ours)	73.76	79.04	74.37	81.98	78.51

ISSN: 2710-0588

Table 2. Experimental Results of Semantic Similarity Tasks on STS-B and SICKr.

Method	STS-B	SICKr	Avg
InferSent-GloVe	68.03	65.65	65.01
Universal Sentence Encoder	74.92	76.69	71.22
SBERT-flow	79.12	76.21	76.60
SRoBERTa-whitening	79.49	76.65	76.60
SBERT	77.03	72.91	74.89
SRoBERTa	77.77	74.46	74.21
MPC (ours)	80.60	77.72	78.00

Tables 1 and 2 report Spearman correlation scores, with the best result in each column shown in bold. Avg denotes the mean across STS12-STS16, STS-B, and SICKr. MPC performs strongly on most tasks compared with the SBERT and SRoBERTa baselines. Relative to SRoBERTa, MPC improves the score by 2.83 points on STS-B, 3.26 points on SICKr, and 3.79 points on Avg. It also outperforms other SBERT-based improvements, including SBERT-flow and SRoBERTa-whitening. These results indicate that MPC maps text into a vector space whose geometry is more suitable for similarity computation.

4.4. Discussion of Similarity Distributions

Anisotropy causes sentence embeddings to be distributed nonuniformly and degrades performance on STS tasks. Vectors representing both semantically similar and dissimilar sentences become concentrated in the same narrow cone and consequently obtain similar cosine-similarity values. To assess whether MPC alleviates this problem, we compute cosine similarities for the STS-B test samples and visualize their distributions.

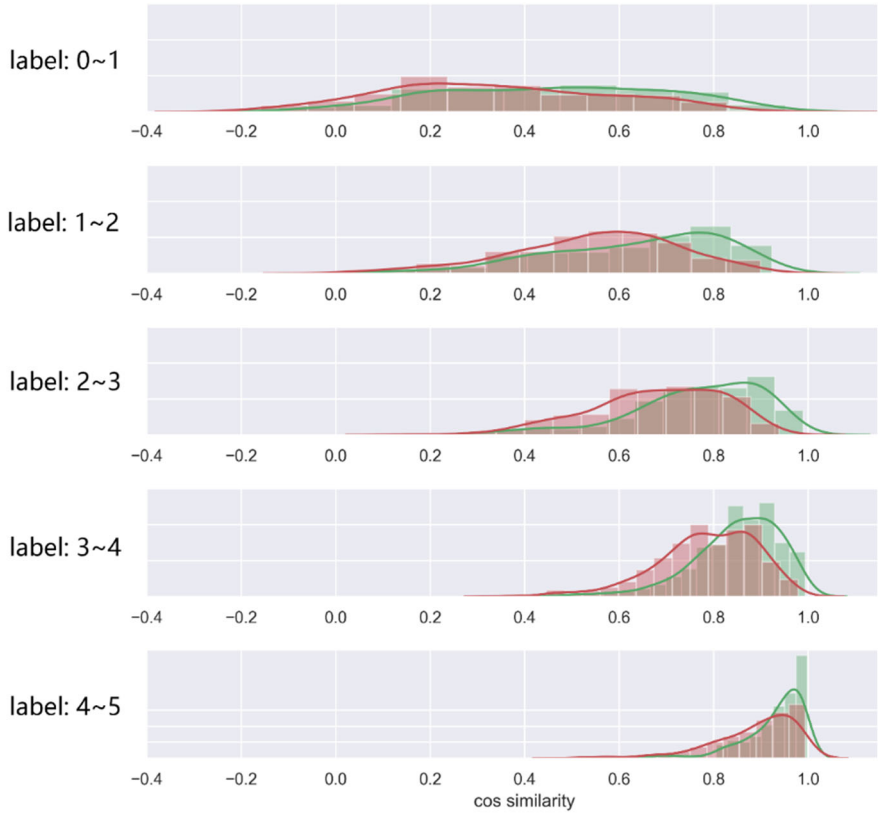

Figure 4. Cosine Similarity Distribution between SBERT and MPC Models.

ISSN: 2710-0588

Figure 4 presents cosine-similarity distributions for STS label intervals [0, 1], (1, 2], (2, 3], (3, 4], and (4, 5]. The x-axis denotes cosine similarity, and the y-axis denotes sample count or proportion. Compared with SBERT, the MPC distributions shift to the left and become more uniform and discriminative. The MPC embeddings therefore alleviate anisotropy to a meaningful extent and are better suited to semantic similarity measurement.

4.5. Human-Consistency Experiment

For an image retrieval system, agreement with human semantic understanding is essential. A query image may contain many irrelevant elements, and the content of interest is not necessarily the largest or most visually salient object. Users instead expect results that capture the central meaning of the image. Aligning retrieval decisions with human semantic judgments is therefore particularly important.

We compare the proposed method with several strong baselines. IRSGS is a scene-graph-based image-to-image retrieval approach. It first converts each image into a semantic scene graph and then measures similarity between graphs. IRSGS-GCN uses graph convolutional networks, whereas IRSGS-GIN uses graph isomorphism networks. CLIP is a multimodal model pretrained on large-scale image-text data and has strong visual-semantic feature extraction capabilities. In a one-stage retrieval setting, CLIP directly extracts image embeddings and searches by vector similarity.

Table 3. Evaluation of Human Consistency Results.

Method	Image Semantic Extraction	Similarity Computation	Human Consistency Score
Human	Human	Human	73.0 ± 5
IRSGS-GCN	GCN	IRSGS	61.1
IRSGS-GIN	GIN	IRSGS	61.2
CLIP	CLIP	-	62.3
SBERT	EKA-BLIP	SBERT	64.9
SRoBERTa	EKA-BLIP	SRoBERTa	65.6
Ours	EKA-BLIP	MPC	66.7

For a more comprehensive comparison, we separate the image-semantic extraction method from the similarity-computation method. Image-semantic extraction converts an image into metadata such as a scene graph or caption. In the Human condition, one of the five manually annotated MS COCO captions is sampled at random. Similarity computation then encodes metadata and calculates similarity; Human denotes direct human judgment. CLIP is treated as a one-stage method because its image embedding is searched directly and requires no additional metadata encoder.

The first row of Table 3 is the mean human-consistency score among all annotators and reflects disagreement in human interpretations of image semantics. The proposed method achieves a higher score than the other automatic approaches, indicating closer agreement with human perceptions of image similarity. Compared with the widely used one-stage CLIP baseline, the two-stage framework improves the score by 4.7 points, demonstrating the advantage of explicit caption metadata and MPC-based similarity computation.

ISSN: 2710-0588

4.6. Visualization of Retrieval Results

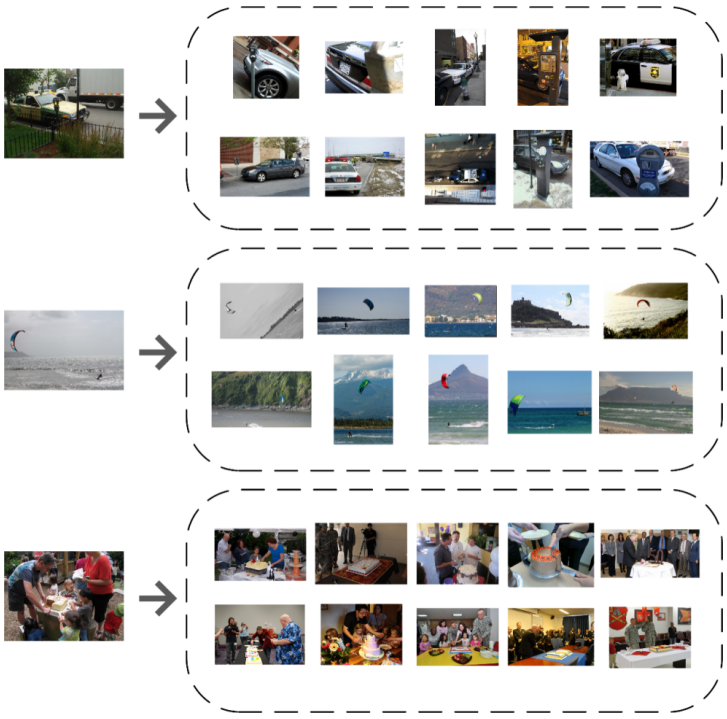

Figure 5. Query Image and Retrieval Results Ordered by Similarity.

As shown in Figure 5, the proposed algorithm retrieves images that are semantically similar rather than merely similar in appearance. The retrieved images differ from the query in dominant color and viewpoint while describing similar content. For example, the third row retrieves the action of cutting a cake in different scenes rather than the same person. This result indicates that the algorithm captures high-level image semantics and produces meaningful retrieval results.

5. Ablation Studies

5.1. Effect of Learning Rate

We investigate the effect of the learning rate by training MPC with several values and recording Spearman correlation on the STS-B development set (STS-B-dev) and SICKr development set (SICKr-dev). The results are also visualized for comparison.

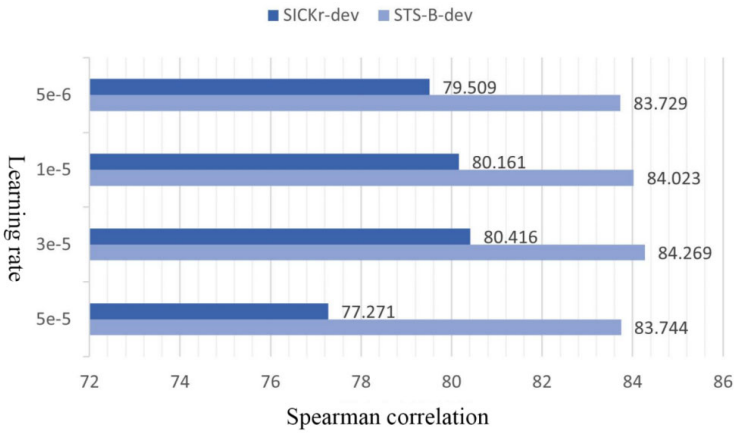

Figure 6. The Influence of Learning Rate on Training Outcomes.

ISSN: 2710-0588

Figure 6 and Table 4 show that a learning rate of 3e-5 produces the best overall training performance on STS-B-dev and SICKr-dev. A smaller learning rate updates model parameters too slowly, whereas an excessively large learning rate can impair convergence and generalization. Learning-rate selection is therefore important, and 3e-5 is used as the default value in the remaining experiments.

Table 4. Effects of Learning Rate on Model Performance.

Learning Rate	STS12	STS13	STS14	STS15	STS16	STS-B	SICKr	Avg
5e-5	73.327	77.901	72.675	81.228	77.924	79.624	75.562	76.891
3e-5	73.759	79.035	74.368	81.975	78.513	80.603	77.716	77.996
1e-5	73.487	79.699	75.026	81.389	78.004	79.836	78.01	77.922
5e-6	73.121	79.708	75.248	80.71	77.469	79.338	77.587	77.597

5.2. Effect of Batch Size

In the multi-positive contrastive objective, samples from other synonym sets in the same batch serve as negatives. Batch size should therefore have a substantial influence on training. We evaluate batch sizes from 16 to 64 to examine this effect.

Table 5. Effect of Batch Size on Model Performance.

Batch Size	STS-B-dev	SICKr-dev
16	83.626	76.854
32	84.273	79.665
40	83.811	78.742
48	84.239	79.88
56	84.187	79.572
64	84.269	80.416

Ablation experiments are conducted on STS-B-dev and SICKr-dev. As shown in Table 5, performance generally improves as batch size increases. Results are relatively stable for batch sizes from 32 to 64, whereas a batch size of 16 causes a marked decline. The relationship is therefore not linear: beyond a certain threshold, the marginal benefit of additional in-batch negatives gradually diminishes.

To examine this hyperparameter more comprehensively, we also evaluate different batch sizes on the remaining STS tasks. Figure 7 shows that scores generally increase with batch size, consistent with the preceding results. Because other samples in the batch serve as negatives, a larger batch provides more negative examples and strengthens contrastive training. The trend is not strictly monotonic: local decreases occur on STS13 and SICKr-dev, followed by recovery. Overall, increasing batch size is beneficial within the evaluated range.

ISSN: 2710-0588

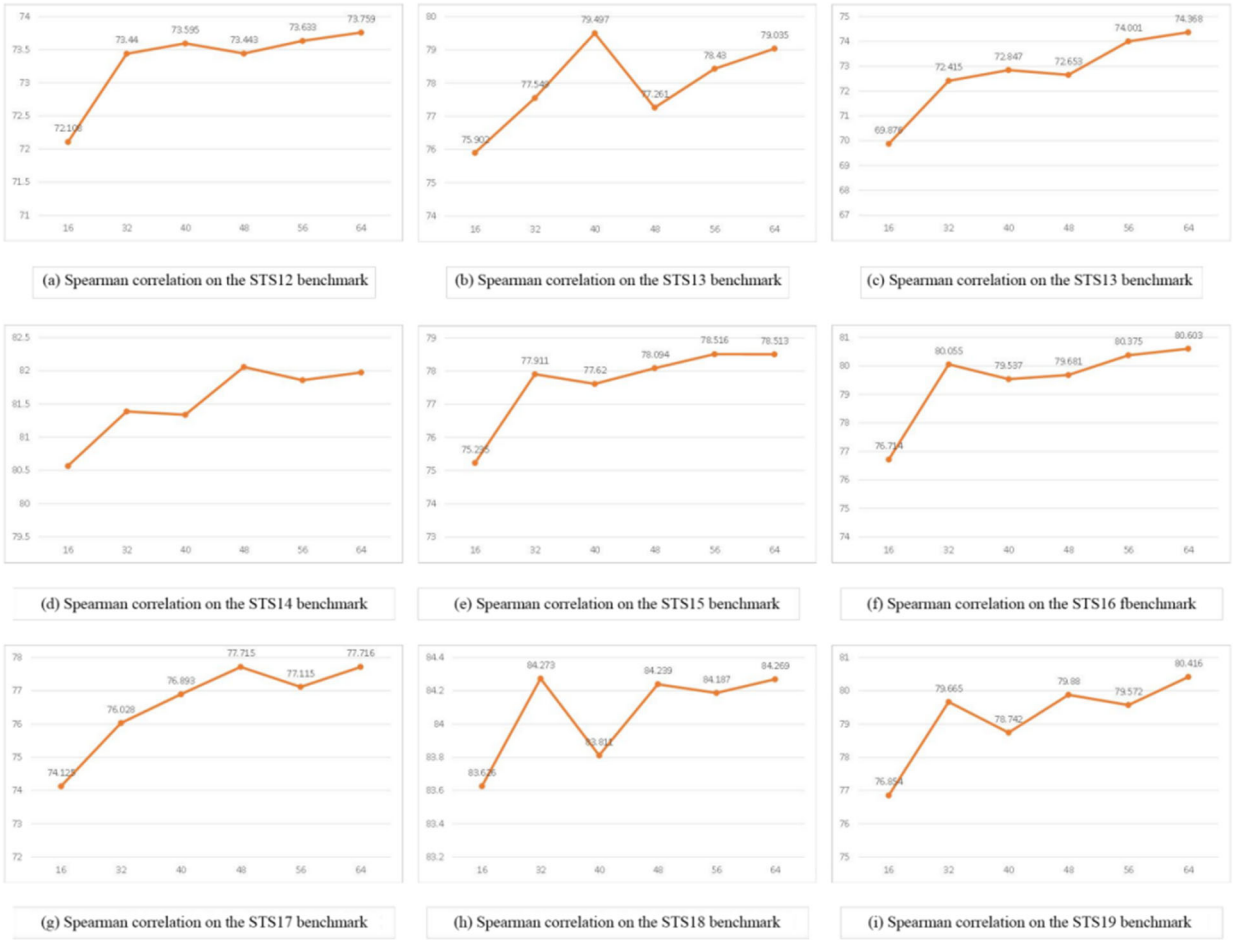

Figure 7. Analysis Chart of Spearman Correlation Results on STS Benchmarks.

5.3. Effect of the Loss-Weight Hyperparameter

The hyperparameter lambda controls the composition of the final loss. We evaluate lambda values of 0.01, 0.1, 0.5, and 1.0 and compare model performance.

Table 6. Ablation Study on Hyperparameters.

lambda	STS12	STS13	STS14	STS15	STS16	STS-B	SICKr	Avg
1	72.993	78.645	73.701	82.183	78.512	80.029	77.434	77.642
0.5	73.805	79.135	74.257	82.246	78.302	80.26	77.443	77.921
0.1	73.759	79.035	74.368	81.975	78.513	80.603	77.716	77.996
0.01	73.911	78.842	73.993	81.098	77.934	79.717	77.587	77.583

Table 6 reports the effect of lambda on the STS benchmarks. Adjusting this parameter changes model performance to a meaningful extent. The best STS-B score and the strongest overall result are obtained at lambda = 0.1. Because lambda controls the relative contribution of the loss terms, the result indicates that the multi-positive contrastive component is important to training and provides additional evidence for the effectiveness of the proposed method.

6. Conclusion

This study addresses textual metadata similarity measurement in image retrieval by proposing MPC, a semantic encoder trained with a multi-positive contrastive loss. First, EKA-BLIP automatically generates synonymous caption sets, substantially reducing the cost of manual annotation. Second, the conventional contrastive objective is extended to a supervised multi-

ISSN: 2710-0588

positive formulation that exploits the many-to-one relationship between captions and images and alleviates embedding anisotropy. Third, MPC is integrated into a two-stage image-to-image retrieval framework. Experiments on STS benchmarks and a human-consistency task demonstrate the effectiveness of the model and retrieval framework. The method nevertheless has limitations: a short caption is not necessarily the optimal metadata representation and may omit ambiguity or fine-grained details. Future work will investigate richer metadata forms and more comprehensive evaluation settings.

Funding

This work was supported by the Humanities and Social Sciences Research Project of the Ministry of Education of China (Grant No. 22YJA87003), as part of the project Research on Deep-Learning-Based Image Retrieval Methods for Digital Libraries.

Author Contributions

Cao Danyang: conceptualization, research design, and methodology.

Cheng Liang: writing, review and editing, data collection, and data cleaning.

Zhou Hongbo: project administration, testing and analysis, and data processing.

References

[1] Dubey, S. R. (2021). A decade survey of content based image retrieval using deep learning. IEEE Transactions on Circuits and Systems for Video Technology, 32(5), 2687–2704. https://doi.org/10.1109/tcsvt.2021.3080920.

[2] Huang, K. (2018). Content-based image retrieval using generated textual meta-data. In Proceedings of the 2nd International Conference on Advances in Artificial Intelligence (pp. 16–19).

[3] Yoon, S., Kang, W. Y., Jeon, S., Lee, S., Han, C., Park, J., & Kim, E.-S. (2021). Image-to-image retrieval by learning similarity between scene graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 10718–10726. https://doi.org/10.1609/aaai.v35i12.17281AAAI Publi...

[4] Li, J., Li, D., Xiong, C., & Hoi, S. (2023). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International Journal of Computer Vision, 131, 1795–1814.

[5] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992). https://doi.org/10.18653/v1/D19-1410

[6] Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020). On the sentence embeddings from pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9119–9130).

[7] Ismail, F., Ramzan, N., Sajid, M., & Khan, H. (2022). Deep learning for image retrieval: Recent progress and challenges. Information Fusion, 82, 59–83.

[8] Yang, H., & Shi, S. (2023). A review of content-based image retrieval technology. Software Guide, 22(4), 229–244.

[9] Wang, Y., Zhang, J., Li, X., & Wang, S. (2023). Multimodal image retrieval: A survey of methods and applications. Information Fusion, 93, 24–45.

[10] Li, Y., Mo, H., Wang, S., & Liu, Q. (2018). A person retrieval method based on image descriptions. Journal of System Simulation, 7, 2794–2800.

[11] Wei, X., Qi, Y., Liu, J., & Chen, J. (2017). Image retrieval by dense caption reasoning. In 2017 IEEE Visual Communications and Image Processing (VCIP) (pp. 1–4).

ISSN: 2710-0588

[12] Yoon, S., Kang, W. Y., Jeon, S., Lee, S., Han, C., Park, J., & Kim, E.-S. (2021). Image-to-image retrieval by learning similarity between scene graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 10718–10726. https://doi.org/10.1609/aaai.v35i12.17281AAAI Publi...

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26.

[14] Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).

[15] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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) (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

[16] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[17] Su, J., Cao, J., Liu, W., & Li, Y. (2021). Whitening sentence representations for better semantics and faster retrieval. arXiv preprint arXiv:2103.15316.

[18] Musgrave, K., Belongie, S., & Lim, S.-N. (2023). A metric learning reality check. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3272–3289.

[19] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9729–9738).

[20] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740–755). Springer International Publishing.

[21] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67–78.

[22] Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., & Artzi, Y. (2019). A corpus for reasoning about natural language grounded in photographs. In Proceedings of the Annual Meeting of the Association for Computational Linguistics.

[23] Cao, D., Zhou, H., & Wang, Y. (2025). Improve the image caption generation on out-domain dataset by external knowledge augmented. Multimedia Systems, 31(1), 1–18.