Identification of Necessary Semantic Undertakers in the Causal View for Image-Text Matching

Huatian Zhang, Lei Zhang, Kun Zhang, Zhendong Mao*
University of Science and Technology of China, Hefei, China
{huatianzhang, kkzhang}@mail.ustc.edu.cn, {leizh23, zdmao}@ustc.edu.cn

Abstract
Image-text matching bridges vision and language, which is a fundamental task in multimodal intelligence. Its key challenge lies in how to capture visual-semantic relevance. Fine-grained semantic interactions come from fragment alignments between image regions and text words. However, not all fragments contribute to image-text relevance, and many existing methods are devoted to mining the vital ones to measure the relevance accurately. How well image and text relate depends on the degree of semantic sharing between them. Treating the degree as an effect and fragments as its possible causes, we define those indispensable causes for the generation of the degree as necessary undertakers, i.e., if any of them did not occur, the relevance would be no longer valid. In this paper, we revisit image-text matching in the causal view and uncover inherent causal properties of relevance generation. Then we propose a novel theoretical prototype for estimating the probability-of-necessity of fragments, $\text{PN}_f$, for the degree of semantic sharing by means of causal inference, and further design a Necessary Undertaker Identification Framework (NUIF) for image-text matching, which explicitly formalizes the fragment’s contribution to image-text relevance by modeling $\text{PN}_f$ in two ways. Extensive experiments show that our method achieves state-of-the-art on benchmarks Flickr30K and MSCOCO.

1 Introduction
Image-text matching aims to search for semantically relevant images given text or retrieve descriptive texts given image, which is a fundamental task in multimodal intelligence facilitating many applications, such as information database search and e-commerce recommendation. Despite considerable development in recent years, image-text matching remains the challenge in capturing visual-semantic relevance.

Extensive research has been done to study the semantic relevance from interactions of cross-modal contents. A common framework is aligning constituent fragments (image regions or text words) semantically and aggregating the resulted alignments accordingly. (Lee et al. 2018) proposed a cross-attention mechanism to capture all latent alignments by attending to regions and words with each other as context, and inspires a bunch of studies. (Zhang et al. 2020b; Wehrmann, Kolling, and Barros 2020; Chen and Luo 2020; Liu et al. 2020) constructed thoughtful aligning rules to capture fine interactions. (Diao et al. 2021) explored self-attention reasoning as an aggregation mechanism to enhance meaningful alignments. (Zhang et al. 2022a) assigned high confidence to image regions consistent with the global semantics in aggregating. (Pan, Wu, and Zhang 2023) proposed to eliminate redundant or irrelevant fragment alignments from the perspective of information coding. In general, not all fragments contribute to image-text relevance, and a large branch of existing methods is devoted to mining the vital ones to measure the relevance accurately.

Normally, how well image and text relate depends on the degree to which they overlap into shared semantics. Fragments that contribute to image-text relevance are those that, if any of them did not occur, the relevance would be no longer valid, i.e., they are necessary to image-text relevance. In other words, these fragments are necessary undertakers of the degree of semantic sharing between image and text. Although the unnecessary ones may also locally correspond to the other modality, identifying necessity can filter out such spurious correspondences from image-text relevance measure by their low necessity, to reduce the impact on match-
ing. As shown in Fig. 1(a), even if the unnecessary man in blue is locally related to the text, the necessity suppresses its contribution to overall relevance. Meanwhile, as the redundancy of unnecessary fragments, excluding them from image or text can help to establish alignments more discriminatively, without altering the inherently shared semantics. Particularly for hard negatives, identifying necessity helps to focus on the evidentiary conflicts. As shown in Fig. 1(b), necessary basin region points out the semantic contradiction.

Treating the degree of semantic sharing as an effect and fragments as its possible causes, the necessary undertakers refer to those indispensable causes for the generation of the degree. Thereby, identifying necessary undertakers is equivalent to determining the probability that the fragment is the degree’s cause. For this purpose, we aim to answer “What would happen to image-text relevance if a fragment did not occur?”. From the perspective of causal inference (Pearl 2009; Glymour, Pearl, and Jewell 2016), the question hypothesizes an absence of fragment, and introduces a comparison on the degree of semantic sharing between actuality and the imaginary scenario where the fragment absents. To capture how the degree varies, we express semantic change of image or text caused by the absence of fragment by the fragment’s semantic dependency, which collects regions or words that have direct semantic causalities with it. Then we structurally model the generation of image-text relevance to specify the functional relationships that connect semantic changes and the relevance, and further uncover two causal properties of relevance generation in matching, the exogeneity of semantic dependency and the matching monotonicity.

Based on the insights above, we propose a novel theoretical prototype for estimating the probability-of-necessity of fragments by means of causal inference, and further design a Necessary Undertaker Identification Framework (NUIF) for image-text matching, which quantitatively identifies necessary undertakers of the degree of semantic sharing between image and text in measuring overall relevance. Specifically, we first relate fragments between modalities to obtain vision-language alignments. Then we represent image and text adaptively by highlighting regions or words that have direct semantic causalities with it. Then we structurally model the generation of image-text relevance to specify the functional relationships that connect semantic changes and the relevance, and further uncover two causal properties of relevance generation in matching, the exogeneity of semantic dependency and the matching monotonicity.

To capture image-text relevance for matching critical fragments in representing image and text, and quantifies the probability-of-necessity of fragments counterfactually by relative variation in how image and text overlap after removing fragments’ semantic dependencies. (3) The experimental results validate the effectiveness of our proposed method, and demonstrate that NUIF achieves state-of-the-art on benchmarks Flickr30K and MSCOCO.

2 Related Work

**Image-Text Matching.** To capture image-text relevance for matching from fine-grained interactions on fragments, extensive works have been proposed. Different from the research line that focuses on representing the holistic image or text to perform coarse cross-modal interaction (Chen et al. 2021; Yan, Yu, and Xie 2021; Li et al. 2022b; Fu et al. 2023), the research line examining fine-grained interactions attracts a lot of attention. One of the representative (Lee et al. 2018) proposed the cross-attention mechanism that aims to discover all region-word fragmental alignments and inspires a series of works (Wehrmann, Kolling, and Barros 2020; Chen and Luo 2020; Liu et al. 2020; Ji, Chen, and Wang 2021; Zhang et al. 2023a). Some works focused on exploiting more information, such as scene graph (Wang et al. 2020b), consensus knowledge (Wang et al. 2020a), and external pre-trained knowledge (Wei et al. 2020; Qu et al. 2021; Yao et al. 2021), etc., to enhance cross-modal alignments. Another line of methods focused on constructing thoughtful aggregating rules to capture vital fragmental interactions. (Liu et al. 2020) and (Diao et al. 2021) explored the structure aligning between regions and words via graph neural network. (Zhang et al. 2022a) assigned confidence to regions to emphasize alignments queried by reliable ones in semantic relevance aggregation. (Zhang et al. 2022b) proposed the negative-aware attention to use the misaligned fragments explicitly. (Kim, Kim, and Kwak 2023) coded samples into a set of different embeddings that captures diverse semantics to handle ambiguity. (Pan, Wu, and Zhang 2023) proposed eliminating irrelevant alignments through cross-modal hard aligning based on coding theory.

**Causality in Computer Vision.** Causal inference (Pearl 2009; Glymour, Pearl, and Jewell 2016) has been widely applied to computer vision to gain insight into the intrinsic causal mechanism of tasks, including visual recognition (Wang et al. 2020c; Liu et al. 2022; Mao et al. 2022), semantic segmentation (Zhang et al. 2020a), scene graph generation (Tang et al. 2020), video analysis (Li et al. 2021; Liu et al. 2021), domain generalization (Lv et al. 2022; Chen et al. 2023a), object navigation (Zhang et al. 2023b), etc. In multimodal machine learning, (Yang et al. 2021) alleviated the dataset bias in image captioning based on the backdoor and frontdoor adjustment principles. (Wei et al. 2022) synthesized counterfactual samples to augment training data for image-text matching. (Chen et al. 2023b) proposed a counterfactual samples synthesizing and training strategy to improve visual-explainable and question-sensitive abilities of visual question answering. (Zang et al. 2023) captured video features causally related to question to restrain redundant language semantics on question answering. In this paper, we examine image-text matching in the causal view, to estimate
the probability-of-necessity of fragments to undertake the
degree of semantic sharing between image and text, in order
to identify the necessary undertakers quantitatively in mea-
suring image-text relevance.

3 Image-Text Matching in the Causal View

We start by structurally modeling the generation of image-
text relevance from the perspective of causality, to under-
stand the semantic change in image or text when a fragment
absents, and extract the causal properties inherent in the rel-
ance generation. Then we derive a theoretical prototype
for estimating the probability-of-necessity of fragments to
the degree of semantic sharing by counterfactual means.

3.1 Structural Modeling

Given an image or text, image-text matching is to rank can-
didates (texts or images) based on semantic relevance be-
tween modalities. Generally, a limited number of visual con-
cepts that have semantic dependencies with a region can host
almost losslessly visual context related to this region in the
image. The rest have no direct cause-and-effect on whether
the region occurs or not. In the text, the words in a syntactic
component or phrase are often linguistically interdependent
and work together as a fine-grained semantic unit. Moreover,
specific meanings activated for the region or word are con-
strained by the visual or linguistic context that it is involved.
These facts inspire us to partition an image or a text for each
candidate fragment into semantic dependency, which in-
cludes the fragment itself and gathers up regions or words
that have direct semantic causalities with the fragment, and
semantic complement. When a fragment is removed, the se-
manics in the image or text that emerge from its dependency
will no longer hold due to causal disruption.

As shown in Fig. 2(a), we build a causal graph to for-
malize the causalities among variables: image or text query
Q, the semantic dependency D and complement C of a frag-
ment F, heteromodal candidate H, and image-text relevance
R between Q and H, each vertex corresponds to a variable.
each edge denotes the cause-and-effect relationship between
its end-vertices. Concretely, Q → R ← H denotes that the
relevance R is determined by how well the semantics of
query Q and candidate H overlap. D → Q ← C indicates
that, from the perspective of fragment F, image or text query
Q is composed of its dependency D and complement C or-
ganically. D gathers up the fragments that have direct sem-
antic causalities with F, that is, it recapitulates the context
of F in Q, D ⊥ C means that concepts in C cannot cause the
occurrence or not of D in terms of semantic logic. While, as
shown in Fig. 2(b), F alone cannot be free from the calling-
up from some other fragments in Q\F.

3.2 Necessity Estimation

In causal theory, given an event Y and its possible cause
X, a counterfactual interpretation of causation that effect
Y would not have occurred in the absence of X captures
how necessary the cause X is for the production of Y, i.e.,
probability-of-necessity. The potential response of Y to hy-
pothetical action X = x is denoted as Yx, and yx indicates
that Y would be y if X had been x. Then, under binary logic,
the probability-of-necessity can be defined counterfactually
as PN := P(y′| x, y), standing for the probability of y′ given
that x and y did occur, where x and y denote respectively
the events X = true and Y = true, otherwise false.
Under certain assumptions, the quantity of probability-of-
necessity can be estimated from observational data facts.

To estimate the probability-of-necessity of fragments for
the degree of semantic sharing between image and text, we
first uncover two inherent causal properties in the generation
of image-text relevance as follows.

Exogeneity of Semantic Dependency. In the causal graph,
if variable D is fixed to d, the variation in potential response
of R to D = d, Rd, will be dominated by other variables that
can affect R. However, D and R have no common ancestor
variable, i.e., no confounding. Hence variables capable of
transmitting variations to R are independent of D, and so is
Rd. For the semantic dependency of a fragment f, the way R
would respond to its occurrence D = oD or absence D = o′
D is independent of the actual value of D, thus:

\[
\{R_{oD}, R_{o'D}\} \perp D. \tag{1}
\]

In causal terms, the dependency D is exogenous relative to
image-text relevance R.

Matching Monotonicity. For images, peeling away nei-
ther salient regions together with their dependent context
nor trivial backgrounds will render the image that does not
match a description match better. Similarly, in texts, mask-
ing out syntactic components will reduce the descriptive se-
manics and make the text sketchy or even blur its logic, thus
the masked text ought not to be more relevant to image can-
didates. It can be summed up as the absence of semantic de-
pendency, D = oD, cannot make query Q that does not match
heteromodal candidate H turn to match. Furthermore, let M
denotes matching degree between Q and H, in the binary
case, m for M = true and m′ the opposite. Then:

\[
m_{oD} \land m'_{oD} = \text{false}, \tag{2}
\]

that is, the matching degree M is monotonic relative to the
occurrence of semantic dependency D.

Putting the insights above together, the probability-of-
necessity can be quantified for identifying necessary under-
takers in image-text matching.

Theorem 1 (Necessary Undertaker Identification). In im-
age-text matching, for a fragment f in image or text, with

![Figure 2: The causal graphs of image-text matching. (a) Semantic dependency (b) Fragmented view](image-url)
semantic dependency $d$, its probability-of-necessity of undertaking the degree of semantics sharing between image and text, $PN_f$, can be quantified by $(P(m \mid o_d) - P(m \mid o_d'))/P(m \mid o_d)$, where $m$ indicates the event that image and text match, $o_d$ denotes $f$’s semantic dependency occurs and $o_d'$ for its absence.

**Proof.** See Appendix A for details.

Note that it does not need to guarantee that $M$ is binary in Thm. 1. For the case of continuous $M$, Thm. 1 can also measure the relative weakening in relevance between image-text pairs with $o_d$ and $o_d'$ as the probability-of-necessity.

## 4 The Proposed Implementation

We then elaborate on the implementation of $PN_f$ in Thm.1. Specifically, as shown in Fig. 3, given image $I = \{v_i\}_{i=1}^M$, text $T = \{u_i\}_{i=1}^L$, where $v_i$ denotes detected salient region, and text $T = \{u_i\}_{i=1}^L$, where $u_i$ is word embedding, we design a Necessary Undertaker Identification Framework as: (1) Vision-Language Aligning: We relate regions and words to obtain visual-semantic alignments; (2) Adaptive Representing: We adaptively represent the image and text in matching by emphasizing regions or words that are semantically bijective with what they aligned in another modality, to enable relevance measure sensitive to such match-critical fragments so that the measure can acutely reflect the change in image-text semantic overlap; (3) Necessary Undertakers Identifying: We model the $PN_f$ in two ways, one to measure the difference $\Delta P = P(m \mid o_d) - P(m \mid o_d')$ integrally and model the $PN_f$ as $\Delta P/P(m \mid o_d)$, named $PN_f$-d, and the other to measure the ratio $P' = P(m \mid o_d')/P(m \mid o_d)$ as a whole and model the $PN_f$ as $1 - P'$, named $PN_f$-r.

### 4.1 Vision-Language Aligning

To capture visual-semantic relevance at the fragment level, we obtain the semantically related fragments for one in another modality through the cross-attention mechanism. For region $v_i$, we measure its attention weight on word $u_j$ by $w_{ij}^v = e^{(c_{ij})}/\sum_{i=1}^{M} e^{c_{ij}}$, $c_{ij} = [c_{ij}]_+ / \sum_{i=1}^{M} [c_{ij}]_+$, where $\lambda$ is a constant temperature parameter and $c_{ij}$ is the cosine similarity between region $v_i$ and word $u_j$, and aggregate $v_i$’s relevant words as its linguistic context $u_i^v = \sum_{j=1}^{L} w_{ij}^v u_j$. Then we embody vision-language alignment as a vector-valued distance between $v_i$ and its attended context $u_{i}^v$:

$$a_{i}^v = l_2\text{-normalized}(\text{tanh}(W_v|v_i - u_{i}^v|^2)),$$

where $W_v \in \mathbb{R}^{P 	imes D}$ denotes a learnable projection matrix. It can be said that alignment $a_{i}^v$ is queried by $v_i$. Similarly, the alignment $a_{j}^u$ queried by word $u_j$ is $a_{j}^u = l_2\text{-normalized}(\text{tanh}(W_u|u_j - v_{j}^u|^2))$, where $v_{j}^u$ is the visual context of $u_j$ and aggregated from regions with $\sum_{i=1}^{M} w_{ij}^u v_i$.

### 4.2 Adaptive Representing

To make relevance measurement more responsively reflect the variation in how image and text overlap as their semantics change, e.g., from $o_d$ to $o_d'$, we adaptively represent the image and text by highlighting regions or words that are most semantically consistent with each other, i.e., bijective with, to which they aligned in another modality. Such regions or words are match-critical since they are exactly the grounding bases of image-text semantic overlapping.

In detail, to represent the image $I$ in matching $I$ and text $T$, for region $v_i$, we first obtain its most semantically aligned word as $\sum_{j=1}^{L} w_{ij}^{v-u} u_j$ by $w_{ij}^{v-u} = [w_{ij}^{v-u}, w_{ij}^{v-u-1}, \ldots, w_{ij}^{v-u-L}] = \text{softmax}(\tau \log(w_{ij}^v))$, where $w_{ij}^v$ is the attention distribution of $v_i$ on text words in Sec. 4.1, and $\tau$ is a temperature parameter. Then, we measure the likelihood that $v_i$’s linguistic counterpart exists in the text $T$ to indicate the degree to which the $v_i$ is match-critical by:

$$s_i^v = v_i \cdot \sum_{j=1}^{L} w_{ij}^{v-u} v_j^u,$$

which is the similarity between $v_i$ and $\sum_{j=1}^{L} w_{ij}^{v-u} v_j^u$ that serves as the visual context of $v_i$’s semantically aligned
where $P_{v}$ and the regions aligned by word $\sum_{j=1}^{L} w_{ij}u_{j}$. It measures the similarity between $v_{i}$ and the regions aligned by word $\sum_{j=1}^{L} w_{ij}u_{j}$. The more similar the two are, the more likely there is a linguistic counterpart of $v_{i}$ in the text $I$, i.e., the more match-critical $v_{i}$ is. For the image $I$, we obtain $s^{v} = [s_{i}^{1}, s_{i}^{2}, \ldots, s_{i}^{M}]$ and represent image $I$ by:

$$
\hat{i} = \sum_{i=1}^{M} \text{softmax}(s^{v})v_{i},
$$

(5)

which highlights match-critical regions within $I$ through weights $s^{v}$ adaptively. Likewise, the text $T$ is represented as $t = \sum_{j=1}^{L} \text{softmax}(s^{u})u_{j}$, where $s_{j}^{v}$ is similar to Eq. 4.

### 4.3 Necessary Undertakers Identifying

We first gather the semantic dependency of individual fragments as follows. In image $I$, the regions that are semantically dependent on region $v_{i}$ will tend to interact with it, which usually manifests as spatial proximity. For $v_{i}$, we regard regions that are relatively close to $v_{i}$, together with $v_{i}$ itself as its semantic dependency $d_{i}$. In text $T$, the phrase to which $u_{j}$ belongs is naturally its dependency $d_{j}$, while the word not belonging to any phrase is its own dependency.

Here it is ready for identifying the necessary undertakers. We model the $PN_{I\mid T}$ as $P(m \mid o_{d}) = \frac{P(m \mid o_{d})}{P(m \mid o_{d})}$ in two ways as follows.

**PN_{I\mid T}**. The difference $\Delta P = P(m \mid o_{d}) - P(m \mid o_{d})$ expresses the variance in how well image and text match caused by removing $f$’s semantic dependency, and can be measured integrally as the relevance between the content emerged from dependency and image or text candidate instance. For region $v_{i}$ with semantic dependency $d_{i}$, we represent the dependency as $d_{i} = \sum_{k \in \text{idx}_{i}} \text{softmax}(s_{\text{idx}_{i}}^{v})u_{k}$, where idx$_{k}$ denotes the index set of $d_{i}$. Then we measure the $\Delta P$ w.r.t. $v_{i}$, $\Delta P_{i}^{v}$ as:

$$
\Delta P_{i}^{v} = (1 + \frac{1}{N} \sum_{n=1}^{N} \frac{\langle d_{i} \rangle_{n} \cdot t_{n}}{\|d_{i}\|_{2} \|t_{n}\|_{2}}) / 2,
$$

(6)

which is block-wise cosine similarity with $N$ blocks. It reduces noise in similarity measure through the refinement on dimensions. Then we measure $P(m \mid o_{d})$, the probability of image and text match, by similarity between $\hat{i}$ and $t$:

$$
P(m \mid o_{d}) = (1 + \frac{1}{N} \sum_{n=1}^{N} \frac{i_{n} \cdot t_{n}}{\|i_{n}\|_{2} \|t_{n}\|_{2}}) / 2.
$$

(7)

Further, we model the probability-of-necessity of region $v_{i}$ as $PN_{i\mid d} = \Delta P_{i}^{v} / P(m \mid o_{d})$.

Similarly, we measure the $PN_{i\mid d}$ of the word $u_{j}$ with semantic dependency $d_{j}$ as $\Delta P_{j}^{d} / P(m \mid o_{d})$, where $\Delta P_{j}^{d} = \max_{k \in \text{idx}_{j}} (1 + \frac{1}{N} \sum_{n=1}^{N} \frac{(u_{i} \cdot t_{n})_{k}}{\|u_{i}\|_{2} \|t_{n}\|_{2}}) / 2$, which is the largest relevance variation can be caused by removing word in $d_{j}$.

**PN_{I\mid R}**. The ratio $P(m \mid o_{d}) / P(m \mid o_{d})$ means the retention rate of the probability that image and text match when the semantic dependency $d$ changes from occurrence $o_{d}$ to absence $o_{d}'$. It implies the level of likeness between image-text shared semantics under $o_{d}$ and $o_{d}'$. For region $v_{i}$, we represent its semantic complement in image $I$ as $c_{i} = \sum_{k \in \text{idx}_{i}} \text{softmax}(s_{\text{idx}_{i}}^{v})u_{k}$, where idx$_{k}$ denotes the complement of idx$_{k}$ in image $I$. Then we formulate the $P(m \mid o_{d})$ through vision-language aligning in Eq.3 between $c_{i}$ and $t$:

$$
P(m \mid o_{d}) = \text{l}_{2}\text{-normalized}\left(\text{tanh}(W_{v}c_{i} - t_{i})\right),
$$

(8)

and the $P(m \mid o_{d})$ through the aligning between $i$ and $t$:

$$
P(m \mid o_{d}) = \text{l}_{2}\text{-normalized}\left(\text{tanh}(W_{v}i - t_{i})\right).
$$

(9)

Further, we measure the ratio $P(m \mid o_{d}) / P(m \mid o_{d})$ by:

$$
P^{r} = (1 + a_{i}^{v} \cdot a_{j}^{r}) / 2,
$$

(10)

which means the projection of $a_{i}^{v}$ onto $a_{j}^{r}$. Then we model the probability-of-necessity of $v_{i}$ as $PN_{i\mid r} = 1 - P^{r}$.

Likewise, we measure the $PN_{i\mid r}$ of the word $u_{j}$ with semantic dependency $d_{j}$ as $1 - P^{r}_{j}$, where $P^{r}_{j} = (1 + a_{t}^{v} \cdot a_{r}^{u}) / 2$, similar to Eq. 10.

Then we aggregate the vision-language alignments queried by the necessary regions through $PN_{i\mid d}$ or $PN_{i\mid r}$ as $a_{i} = \sum_{i} \text{softmax}(PN_{i\mid d}), a_{r}^{u}$, and the alignments queried by the necessary words through $PN_{i\mid d}$ or $PN_{i\mid r}$ as $a_{T} = \sum_{T} \text{softmax}(PN_{i\mid r}), a_{r}^{u}$, then incorporate them into image-text relevance $r$ by:

$$
r(I, T) = \text{tanh}(w_{r}([a_{i} : a_{T}]),
$$

(11)

where $w_{r} \in \mathbb{R}^{1 \times 2P}$ is a learnable vector, and the $[\cdot]$ denotes the concatenation operation.

### 4.4 Training

**Feature Encoder.** For a fair comparison with previous works, we use the ROI features of pre-trained object detector as detected regions, and transform them to $D$-dimensional $v_{i}$ via linear projection. For texts, we employ two types of extractors, Bi-GRU and pre-trained BERT (Kenton and Toutanova 2019). When using Bi-GRU, the embedding of the $i$-th word, $u_{j}$, is averaged from its forward and backward hidden states. When using BERT, we linearly map its output hidden states to $D$-dimensional embeddings.

**Objective Function.** Ranking objectives are adopted in image-text matching widely to force matched image-text pairs close to each other and pull unmatched ones away. We use the bi-directional triplet loss, focusing on the hardest negatives in-batch for efficiency:

$$
\mathcal{L}(I, T) = [\alpha - r(I, T) + r(I, T)]_{+} + [\alpha - r(I, T) + r(I, T)]_{+},
$$

(12)

where $\alpha$ is a margin constraint, $[\cdot]_{+} = \max(x, 0)$. $I_{h} = \text{argmax}f_{\alpha}(I, T)$, and $T_{h} = \text{argmax}f_{\alpha}(T, I^{*})$ are the hardest negatives, given positive matched $I$ and $T$.

### 5 Experiments

#### 5.1 Datasets and Evaluation Metrics

We evaluate the proposed framework on Flickr30K (Young et al. 2014) and MSCOCO (Lin et al. 2014) datasets.
Table 1: Comparisons with state-of-the-arts on Flickr30K and MSCOCO 1K test sets. The †: the model has GloVe attached for text embedding, and *: only single model is reported. The bests are in bold.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Flickr30K</th>
<th>MSCOCO 1K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IMG → TXT</td>
<td>TXT → IMG</td>
</tr>
<tr>
<td>BUTD + Bi-GRU</td>
<td>81.7</td>
<td>95.4</td>
</tr>
<tr>
<td>GSNM(Liu et al. 2020)</td>
<td>79.7</td>
<td>94.6</td>
</tr>
<tr>
<td>GPO∗(Chen et al. 2021)</td>
<td>82.1</td>
<td>95.8</td>
</tr>
<tr>
<td>SGRAF(Diao et al. 2021)</td>
<td>80.6</td>
<td>96.1</td>
</tr>
<tr>
<td>CMCAN(Zhang et al. 2022a)</td>
<td>84.0</td>
<td>96.1</td>
</tr>
<tr>
<td>NAAF∗(Zhang et al. 2022b)</td>
<td>83.9</td>
<td>96.5</td>
</tr>
<tr>
<td>CHAN∗(Pan et al. 2023)</td>
<td>85.6</td>
<td>97.2</td>
</tr>
</tbody>
</table>

Flickr30K contains 31,000 images and each image with 5 texts. Following dataset splits in (Lee et al. 2018), 29,000 images for training, 1,000 images for validation, and 1,000 images for testing. MSCOCO contains 133,287 images and each image with 5 texts. We use 123,287 images for training, 5,000 images for validation, and 5,000 images for testing, and the results on MSCOCO are reported by both averaging over 5 folds of 1,000 test images and testing on the entire 5,000 test images. As common in the field of information retrieval, we measure the performance by recall R@k and rSum. The higher R@k indicates better performance.

5.2 Implementation Details

We utilize the BUTD features (Anderson et al. 2018) extracted from Faster R-CNN (Ren et al. 2015) with pre-trained ResNet-101 (He et al. 2016) as ROI inputs. M = 36 ROIs in each image. The dimension D = 1024 and P = 256. The temperature parameter λ = 9.0 and τ = 6.0. The number of blocks N = 16. The margin α = 0.2. In the semantic dependency gathering, we calculate polar coordinates (ρ, θ) of other regions relative to the target region and select regions with the first 2 small ρ in each of the scopes that are quartered by θ = π/4, 3π/4, −π/4, and −3π/4, and extract noun phrases from texts by the chunking function of the NLP tool spaCy. In using Bi-GRU, the dropout operation is applied on both region and word features after projecting them into 1024-dim and dropout rate is 0.4, and we employ the Adam optimizer with 0.0002 initial learning rate which is decayed by 10 times after 40 epochs on Flickr30K and after 30 epochs on MSCOCO. In using BERT, the Adam optimizer sets 0.0005 initial learning rate, and decays by 10 times after 20 epochs. Source code will be released1.

---

1https://github.com/hzhang-code/NUIF

Table 2: Comparisons with state-of-the-arts on MSCOCO 5K test set. The bests are in bold.

<table>
<thead>
<tr>
<th>Methods</th>
<th>IMG → TXT</th>
<th>TXT → IMG</th>
<th>rSum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1 R@5 R@10</td>
<td>R@1 R@5 R@10</td>
<td>rSum</td>
</tr>
<tr>
<td>BUTD + Bi-GRU</td>
<td>57.8</td>
<td>91.6</td>
<td>41.9</td>
</tr>
<tr>
<td>SGRAF</td>
<td>61.5</td>
<td>92.9</td>
<td>44.0</td>
</tr>
<tr>
<td>CMCAN</td>
<td>58.9</td>
<td>85.2</td>
<td>92.0</td>
</tr>
<tr>
<td>NAAF†</td>
<td>60.2</td>
<td>85.9</td>
<td>92.4</td>
</tr>
<tr>
<td>CHAN∗</td>
<td>60.4</td>
<td>86.2</td>
<td>92.4</td>
</tr>
<tr>
<td>NUIF-d*(ours)</td>
<td>59.3</td>
<td>85.5</td>
<td>92.0</td>
</tr>
<tr>
<td>NUIF(ours)</td>
<td>61.8</td>
<td>86.6</td>
<td>93.1</td>
</tr>
</tbody>
</table>

5.3 Comparisons with State-of-the-art Methods

We compare our proposed NUIF with recent state-of-the-art methods on the Flickr30K and MSCOCO benchmarks. The experimental results are cited directly from respective papers. When using the BUTD+Bi-GRU encoder, for a fair comparison with more previous methods, we report performances without the pre-trained GloVe representation attached to text embedding. Quantitative results on Flickr30K and COCO 1K test sets are shown in Tab. 1. NUIF outperforms state-of-the-art methods on most metrics with large margins clearly and achieves consistent superiority in different encoder settings. Comparisons on COCO 5k test set are shown in Tab. 2, and our method also performs best on al-
Table 3: Ablation studies of PN_f’s modeling on Flickr30K.

<table>
<thead>
<tr>
<th>Methods</th>
<th>COCO 1K, BUTD + Bi-GRU</th>
<th>COCO 5K, BUTD + Bi-GRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN_f</td>
<td>76.0 94.9 97.9</td>
<td>54.2 82.5 89.2</td>
</tr>
<tr>
<td>w/o dropout</td>
<td>78.3 96.2 98.8</td>
<td>58.0 84.7 91.8</td>
</tr>
<tr>
<td>w/o PN_f</td>
<td>79.9 96.7 99.9</td>
<td>59.3 85.5 92.0</td>
</tr>
<tr>
<td>w/o PN_f (w/ PN_f-d)</td>
<td>80.0 96.4 98.8</td>
<td>59.9 85.3 92.1</td>
</tr>
<tr>
<td>w/o PN_f (w/ PN_f-r)</td>
<td>81.7 97.0 99.9</td>
<td>61.8 86.6 93.1</td>
</tr>
<tr>
<td>w/o PN_f (w/ PN_f-d)</td>
<td>82.9 97.0 98.8</td>
<td>62.9 87.4 92.6</td>
</tr>
<tr>
<td>w/o PN_f (w/ PN_f-r)</td>
<td>83.3 97.3 98.9</td>
<td>63.3 87.7 92.9</td>
</tr>
<tr>
<td>w/o PN_f (w/ PN_f-d)</td>
<td>84.2 97.2 99.1</td>
<td>64.7 97.5 99.1</td>
</tr>
<tr>
<td>w/o PN_f (w/ PN_f-r)</td>
<td>84.7 97.5 99.1</td>
<td>65.8 97.5 99.1</td>
</tr>
<tr>
<td>w/o PN_f (w/ PN_f-d)</td>
<td>64.8 88.2 94.2</td>
<td>65.2 88.8 94.2</td>
</tr>
<tr>
<td>w/o PN_f (w/ PN_f-r)</td>
<td>65.6 89.1 94.3</td>
<td>67.8 89.9 94.8</td>
</tr>
</tbody>
</table>

Table 4: Ablation studies of PN_f’s modeling on MSCOCO.

The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

causality (due to its much smaller proportion of not-so-short text) rather than scale (see Fig. 4). Short texts have weak causality since they are generally rough, while not-so-short ones have rich causality since their fine-grained details, e.g., many regions in a sunset image are aligned with “beautiful sunset”, removing some regions (e.g., cloud) will not affect the degree of semantic sharing between sunset image and text, resulting in ambiguous (weak) causality. While for text “beautiful sunset with white clouds over a river”, the necessity of certain regions (e.g., cloud) is enhanced.

5.5 Visualization

To further verify our method’s ability to specify the necessary undertakers of the degree of semantic sharing between image and text, we visualize the learned PN_f-d of regions and words in Fig. 5. In row 2 column 2, due to the spatial proximity strategy in gathering region dependency, the girl closer to the fish region that semantically corresponds to the match-critical word “fish” is wrongly considered more necessary than the other. However, on the whole, our method effectively captures the regions and words necessary for judging semantic consistency or contradiction in matching.
6 Conclusion

In this paper, we revisit image-text matching in the causal view, and propose a novel theoretical prototype for estimating the probability-of-necessity of fragments for the degree of semantic sharing by means of counterfactual inference. Further, we implement a Necessary Undertaker Identification Framework (NUIF) for image-text matching to formalize the probability-of-necessity of fragments in two ways, which intuitively specifies the contribution of fragments to image-text relevance. Our method attributes the degree of image-text semantic sharing to constituent semantics. Extensive experiments demonstrate the superiority of our proposed NUIF. Future works include designing effective semantic dependency gathering, to reasonably infer fragments’ necessity in specific scenarios.

A Necessary Undertaker Identification

Proof. The semantic dependency \(d\) of the fragment \(f\) gathers up those fragments in image or text that have direct semantic causalities with \(f\), containing \(f\) itself. Removing \(f\) will break these causalities and then distort the semantics emerged from dependency \(d\). This is equivalent to the original semantics of \(d\) being altered from the image or text. That is, as \(o_f\) changes to \(o_f'\), \(o_d\) changes to \(o_d'\) on semantics. Combining the definition of necessary cause (see Sec. 3.2) in causal inference (Pearl 2009; Glymour, Pearl, and Jewell 2016), we express the probability that fragment \(f\) is necessary to the degree of semantic sharing between image and text, probability-of-necessity, as:

\[
PN_f = P\left( m'_d \mid m, o_d \right),
\]

which means the probability that, given that \(d\) did occur and \(M = \text{true}\) in reality, the potential response of matching degree \(M\) to \(d\)’s hypothetical erasure is \(M = \text{false}\). Since the matching degree \(M\) is determined by image-text relevance \(R\) monotonically and uniquely, and \(D\) is exogenous relative to the relevance \(R\), i.e., \(\{R_{od}, R_{o'd}\} \not\perp D\), then:

\[
\{M_{od}, M_{o'd}\} \not\perp D,
\]

which implies:

\[
P(m_{od}) = P(m_{o'd}|o_d) = P(m|o_d),
\]

that is:

\[
o_d \land m = o_d \land m_{od}.
\]

Then, for Eq. 13, we have:

\[
P\left( m'_{od} \mid m, o_d \right) = \frac{P\left( m'_{od}, m, o_d \right)}{P\left( m, o_d \right)} = \frac{P\left( m'_{od}, m_{od}, o_d \right)}{P\left( m, o_d \right)} \times \frac{P\left( o_d \mid m_{od} \right)}{P\left( o_d \mid m \right)}.
\]

Obviously, \(m'_{od} \land m'_{o'd} = \text{true}\), then:

\[
m_{od} = m_{od} \land (m'_{od} \lor m'_{o'd})
\]

Considering the matching monotonicity, \(m_{od} \land m'_{od} = \text{false}\):

\[
m_{od} = m'_{od} \land m_{od}.
\]

Substituting Eq. 20 into Eq. 18, it holds that:

\[
m_{od} = m_{od} \lor \left( m'_{od} \land m_{od} \right).
\]

Since the disjointness of \(m_{od}'\) and \(m_{od}'\), and of \(m_{od}'\) and \(m_{od}\) (exogeneity of \(d\)), we obtain:

\[
P(m_{od}) = P(m_{od}') + P(m_{od}, m'_{od}).
\]

Then taking the exogeneity of \(d\), it yeilds:

\[
P(m|o_d) = P(m|o_d') + P(m_{od}, m'_{od}).
\]

Combining Eq. 17 and Eq. 23, we have:

\[
P\left( m'_{od} \mid m, o_d \right) = \frac{P\left( m \mid o_d \right) - P\left( m \mid o_d' \right)}{P\left( m \mid o_d \right)}.
\]

Thus, we obtain:

\[
PN_f = \left( P\left( m \mid o_d \right) - P\left( m \mid o_d' \right) \right) / P\left( m \mid o_d \right).
\]

which concludes the proof. It is worth noting that, for a matched image-text pair, the matching degree \(m\) \((M = \text{true})\) means they are semantically related and \(m'\) \((M = \text{false})\) means the relationship of matched is no longer valid. For unmatched image and text, the \(m\) indicates the semantic relevance they achieve is being maintained and \(m'\) indicates a weakening of the relevance.

Acknowledgements

This work is supported by the National Science Fund for Excellent Young Scholars under Grant 62222212.

References


Conference on Research and Development in Information Retrieval, 1–10.


