Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4944–4954
Marseille, 20-25 June 2022

Aligning Images and Text with Semantic Role Labels for Fine-Grained Cross-Modal Understanding

Abhidip Bhattacharyya, Cecilia Mauceri, Martha Palmer, Christoffer Heckman
University of Colorado Boulder
firstname.lastname@colorado.edu

Abstract

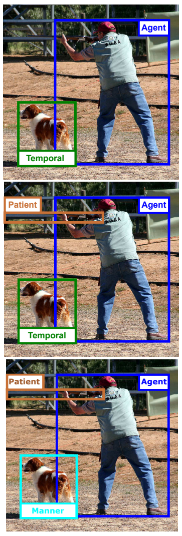
As vision processing and natural language processing continue to advance, there is increasing interest in multimodal applications, such as image retrieval, caption generation, and human-robot interaction. These tasks require close alignment between the information in the images or video and text. In this paper, we present a new semantic labeling method that combines state-of-the-art semantic annotation for language with the bounding boxes of corresponding images. This richer multimodal labeling supports cross-modal inference for applications in which such alignment is useful. Our semantic representations, developed in the natural language processing community, abstract away from the surface structure of the sentence, focusing on specific actions and the roles of their participants. This level of abstraction maps well to the objects, actions, and relationships visible in images. We utilize these representations in the form of semantic role labels in the captions and the images and demonstrate improvements in standard tasks such as image retrieval. The potential contributions of these additional labels is evaluated using a role-aware retrieval system based on graph convolutional and recurrent neural networks. The addition of semantic roles into this system provides a significant increase in capability and greater flexibility for these tasks, and could be extended to state-of-the-art techniques relying on transformers with larger amounts of annotated data.

Keywords: cross-modal retrieval, semantic role labeling,

1. Introduction

Vision processing is making exciting advances and performance is rapidly improving for tasks such as caption generation, question answering and retrieval. In parallel, NLP is making corresponding advances, fueled by both vector representations and rich semantic representations (Wang et al., 2021). In this paper, we explore the benefits of combining the rich semantic representations of NLP with image bounding boxes. A congruent description of an image should be semantically grounded by the objects presented in the image. Therefore composing a concise image description requires focusing on a few contextually salient entities, properties, or events, instead of being exhaustive. For example, in Figure 1, the example captions focus on the man. However, a description that focused on the rifle or on the dog would be equally valid. Currently, image retrieval systems can retrieve relevant results for diverse input, but they do not provide a way to intentionally inject variety into the search results by specifying a desired focus.

Semantic role labeling (SRL) (Palmer et al., 2005) is a form of semantic parsing developed for natural language processing that conveys knowledge about *who is doing what to whom when* as predicate-argument structures. In other words, given an action in a sentence, one needs to know who is performing the action (the agent), who is affected by the action (the patient), what instrument is being used, etc. to comprehend the meaning of the sentence. In the different reference captions of Figure 1, different actions are performed (aiming, standing, shooting, and holding), and different objects fulfill different roles in that action.



Retrieved Description:
[The man]_Agent is aiming to shoot [something]_Patient [while his dog watches]_Temporal

Retrieved Description:
[A man]_Agent shoots [a rifle]_Patient [while a dog looks on]_Temporal

Retrieved Description:
[A man]_Agent aiming [a rifle]_Patient [with a dog standing beside him]_Manner

Figure 1: Semantic role aware text retrieval by RARE. Each query image, on the right, has different associated SRL annotations for each bounding box. Retrieved captions match the SRL structure of the image annotations and demonstrate the variety of descriptive choices that can be made for the same image. This example focuses on differences between the Patient, Temporal, and Manner roles. Agent represents the instigator of an action. Patient is the object affected by the action. Temporal indicates a temporal relationship, i.e. when is the action occurring, what is happening at the same time. Manner is the manner in which the action is performed. A full legend of SRLs is provided in Table 1

Typical image-text retrieval systems use image features and word-embedding features as input representations. If only word features are considered, ignoring the order of the words, the sentences ‘the dog chased the cat’ and ‘the cat chased the dog’ will retrieve similar results. Semantic roles distinguish between who is the chaser and what is being chased. Word order is not always helpful in semantic decoding. For example ‘I gave the book to John’ vs. ‘I gave John the book’ are semantically equivalent. SRL provides information about the semantic roles regardless of word order.

In this paper, we enhance image-text retrieval using SRL. We train a model to recognise pairs of SRL annotated text and image bounding boxes. At test time, by encoding the SRL relationships in the query, the result becomes sensitive to the desired semantic structure. Refer to Figure 1, where SRL components are explained, and a demonstration of how different SRL inputs result in different retrieved captions is provided. In the first query image, the *rifle* was unlabeled, and hence ignored by the caption retrieval. However, in the second query image, the *rifle* was marked with the semantic role `patient`. The retrieved caption for the second query not only mentioned the *rifle*, but also labeled it with the same semantic role. The semantic annotations for the *dog* in the first two query images were `temporal`. The corresponding retrieved captions possess a temporal connotation associated with the presence of the *dog*. However, the final query image had a semantic focus on the presence of the dog as a `manner`, as did the retrieved caption. The retrieved captions corroborated the image’s annotation.

The novelty of this work is demonstrating the potential for semantic roles to contribute to multi-modal retrieval. As automated image SRL labeling is still an active field of research, we use Gold Standard image SRL and automatic text SRL. We create SRL annotated data using the Flickr30k Entity dataset (Young et al., 2014; Plummer et al., 2017). This dataset maps entity mentions in the reference descriptions to image bounding boxes. We obtain the SRL for the descriptions using an automatic SRL system (Gung and Palmer, 2021). Using entity mention mapping, we transfer the text SRL annotations to the corresponding bounding boxes.

We call our method **r**ole **a**ware **r**etri**e**val system (RARE). We compare RARE retrieval to other image-text retrieval models. In comparison to other non-transformer-based models, RARE improves retrieval results by 13.7% in an image-text retrieval task. For text-image retrieval, RARE comes in second to ACMM (Huang and Wang, 2019). Our qualitative results show that when mismatches among query and retrieved results occur our model still preserves shared semantics. Transformer-based methods are the current state-of-the-art and our performance is lower than theirs. Nonetheless our method demonstrates the potential of semantic roles, thereby providing encouraging evidence for benefits to be accrued via their incorporation. To summarize, the main contributions of the current work are:

- We use semantic parsing based on semantic roles to enhance image and text representations in a shared semantic space.
- We demonstrate that semantic role labeling information can be used effectively as a control signal to retrieve specific text captions from amongst diverse descriptions for the same image.

2. Related Work

Semantic roles, described in Section 3, provide a predicate-argument structure representation of a sentence that abstracts away from syntactic variations. Early methods of automatic semantic role labeling relied heavily on syntactic parsing (Pradhan et al. (2005; Punyakanok et al. (2008; Hacioglu et al. (2004)). The first end-to-end deep neural network SRL system to report state of the art performance was authored by (Zhou and Xu, 2015). Their approach did not incorporate structural constraints of SRLs. (He et al., 2017) gained improvement over (Zhou and Xu, 2015) by introducing several modifications including highway LSTMs and structural constraints. Current approaches for SRL use transformers (Tan et al., 2017; Strubell et al., 2018).

The goal of cross-modal retrieval is to match natural language descriptions to images; this is frequently achieved by learning a latent embedding space where related images and text representations that are more similar than those of dissimilar images and text are closer to one another under some distance metric. Early work (Socher and Fei-Fei, 2010; Hodosh et al., 2013; Gong et al., 2014; Yan and Mikolajczyk, 2015; Andrew et al., 2013) aligned the two representations in latent space using **C**anonical **C**orrelation **A**nalysis (CCA) (Hotelling, 1936). CCA learns linear projections that maximize the correlation between projected vectors from the two modalities.

Driven by success in deep learning, neural network-based approaches have been deployed to learn the representations in the latent space. (Chen et al., 2020a) provide an excellent review of recent deep learning based methods. These neural network architectures generally were two-branched, with each branch dedicated to learning a representation for one modality (Faghri et al., 2018; Ma et al., 2015; Wang et al., 2018). In Wang et al. (Wang et al. (2016; Wang et al. (2018)), a bi-directional ranking loss (a hinge-based triplet ranking) (Karpathy and Fei-Fei, 2017) with neighborhood-preserving constraints was used to train the model. A hinge-based triplet ranking loss (Karpathy et al. (2014; Karpathy and Fei-Fei (2017)) is designed to force relevant image-text pairs to be closer in shared space than irrelevant pairs by a fixed margin. Phrase alignment was learned on corresponding pairs of image regions and phrases, as a separate task. The aforesaid methods calculated the loss by summing hinges over all negative samples. (Faghri et al., 2018) showed that the maximum of hinges could outperform the sum of hinges as a loss function, and many recent techniques (Lee et al., 2018; Liu et al., 2019a; Li et al., 2019) use this loss function.

Attention is a major development in deep learning, especially in language applications. Notably for image-text matching, it has improved phrase-region alignment Nam et al. (2017; Fan and Zhou (2018; Lee et al. (2018; Liu et al. (2019a). Correspondences between image regions and text tokens are learned by attending to regions with respect to text or attending to text with respect to regions. However, attention models lack the ability to discriminate irrelevant fragments from relevant fragments. Thus they learn to distribute attention over all fragments, which can lead to misalignment. Recent approaches addressed this issue by either putting a relevance function (Liu et al., 2019a) or stacking attention layers (Lee et al. (2018; Fan and Zhou (2018)). However transformer(Vaswani et al., 2017) based systems are current state of the art (Chen et al., 2020b; Ren et al., 2021). The transformers' success in learning rich semantic and structural information from large, unlabelled data sources in a self-supervised manner and their ability to transfer learning from pre-trained tasks to fine-tuned tasks make them a potential DNN architecture for vision and language tasks. Like their NLP counterparts these cross modal transformer models are also pre-trained via masked language modeling, masked region classification, and alignment.

However explicit semantic labels of images and text have not been explored much in the context of image-text matching. (Karpathy et al., 2014) used dependency parsing to express sentence fragments. (Socher et al., 2014) proposed a dependency-based RNN; in that work, word representations were created from the RNN following the latent hierarchy induced by the dependency parser. However, dependency parses are syntactic, not semantic parses. Some work (Li et al., 2019; Wang et al., 2020b; Wu et al., 2018) explores encoding images in a semantic graph. To the best of our knowledge, only (Wang et al., 2020b) incorporates semantics in both the images and text domains. Moreover representing images and text with similar semantic labels has not been explored. In this work, we use SRL (Palmer et al., 2005) as our semantic cues for both images and text. Semantic roles enable richer representation of an image and the corresponding text in the shared space. As a result, our model achieves superior performance on retrieval tasks.

3. Approach

Figure 2 depicts an overview of the RARE approach. The input consists of an image-text pair, annotated with semantic roles. We use a graph convolutional network to encode semantic relations among image regions. Following the idea of BFAN (Liu et al., 2019a), we use focal attention to align the text and visual features. Finally, two scores are computed to measure the input similarity. In the following sections, we will describe modules of our network in detail.

Image representation Following the method described in (Anderson et al., 2018) an image representation is created using Faster RCNN (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017). An image can be represented by a set of region vectors $R = \{r_1, r_2, ..., r_n\}$ detected by Faster RCNN. The representation, r_i, for each region is derived from the mean-pooled convolutional feature for that region. In our experiments, the Faster RCNN feature is 2048 dimensional and $n = 36$; the top 36 regions with the highest class detection confidence scores are selected as the image features. To save computation we pre-computed Faster RCNN features and used them in RARE.

Semantic Role Labels The semantic role label of a word is determined with respect to a specific verb or predicate in the sentence. Table1 provides definitions for the most commonly used SRLs in Flickr30k. If there is more than one predicate in the sentence, there will be more than one set of SRL annotations, one for each predicate. For example, the sentence "a man shoots a rifle while a dog looks on" has two predicates, one focused on the verb "shoots" and one focused on "looks on". The agent SRL of "shoots" is "man". The agent SRL of "looks on" is "dog". Each set of SRL annotations for one predicate is called a proposition.

Different bounding boxes are salient with respect to each SRL proposition. More formally consider an image-text pair is given by (I, T). Image I is represented by region set R. The caption sentence T consists of tokens $< t_1, t_2, .., t_k >$. With respect to the j-th predicate, each token is labeled with a SRL $< s_1^j, s_2^j, ..., s_k^j >$. The image regions are also assigned SRLs with respect to the j-th predicate, $l_1^j, l_2^j, ..., l_n^j$. A data-point is created as a quadruplet of regions, region-labels, tokens, and token-labels, (R, L_j, T, S_j). During training each quadruplet is considered as a separate training sample. At the time of inference, a similarity score over all the quadruplets for an image-sentence pair is summed.

An embedding layer is used to encode the SRL in the visual input. We use a semantic role vocabulary established by enumerating over all the SRLs in the training set. The most frequent labels are listed in Table 1. Every region extracted from Faster RCNN is annotated with a semantic role. To accomplish this we used the SRL annotation of ground truth bounding boxes and transferred them to Faster RCNN bounding boxes based on a intersection over union (IoU) threshold. To create a joint region-SRL representation, region vectors are passed through a fully connected layer and concatenated with the SRL embeddings. This combined representation is projected in the hidden space via a fully connected layer. We use a visual SRL embedding dimension of 512 and the fully connected layer has di-

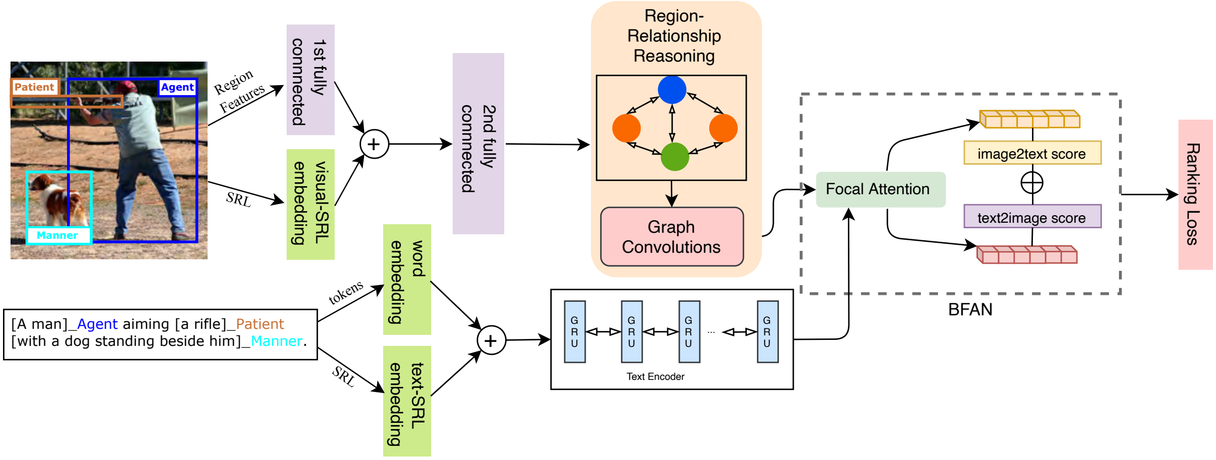


Figure 2: Overview of RARE. A preprocessing step annotates the text SRL automatically and transfers the generated SRL to entity bounding boxes. In the first step of the pipeline, the image and SRL annotated bounding boxes are fed to a fully connected layer and an embedding layer. Tokens and SRL for text descriptions are also processed with embedding layers. After embeddings are computed, the region relationship reasoning module takes the image vectors and creates a graph. For text, a bi-directional GRU is used as text encoder. Finally, the encoded image representation and text representation are input to focal attention.

mension of 512, resulting in a concatenated vector of dimension 1024.

For joint text-SRL representations, token embeddings are concatenated with SRL embeddings. This embedding layer is separate from the image-SRL embedding as the text annotation is over spans. We used BIO tagging to annotate text spans with SRL (He et al., 2017). In our experimental model, the text-SRL embedding is 150 dimensional.

A sequence of concatenated vectors for (T, S_j) is processed by a gated recurrent unit (GRU) to produce a proposition representation. In our experimental model, initial word-embeddings are 300-dimensional and the dimension of text representation from the bidirectional GRU is 1024-dimensional.

Region Encoder To capture relationships among regions we used a region relationship reasoning model. Highlighted by the orange box in Figure 2, the design of our region encoder is inspired by (Li et al., 2019).

A fully connected graph $G = (R, E)$ is created where R is a set of regions and E is an edge between regions. The edge weights are described using an affinity matrix D. Affinity matrix is computed by projecting region vectors in a latent space and then computing the inner product between two regions. A graph convolutional network (GCN) is applied to this graph to encode the relationships among the nodes. As suggested in Li et al., we also have used residual connections to the original GCN as

$$R^{GCN} = (DRW_g)W_r + R, \tag{1}$$

where $W_g \in \mathbb{R}^{d\times d}$ is a weight matrix for a GCN layer. $W_r \in \mathbb{R}^{d\times d}$ is the weight matrix for the residual layer and D is the affinity matrix of shape $n \times n$. The output, R^{GCN}, is the relationship enhanced representation for image regions. R^{GCN} is further processed with a GRU layer. We take the output at each time-step as the image representation. For our model dimensions of W_r, W_g and GRU are 1024.

Focal Attention To curb the problem of semantic misalignment described in the related works section, we have used bi-directional focal attention (Liu et al., 2019a) to align our visual and text representations. To obtain shared semantics for images and text, the representation of one modality is fixed and used to attend the other modality. We use the term “fragment” to refer to a single modality’s representation, either an image region or a text token. Relevant fragments are found using the attention weights in three stages:

i) *Pre-assign Attention*. First, the attention score for each fragment with respect to a fixed representation from the other modality is initialised by computing cosine similarities between fragments and normalizing them using softmax activation.

Without loss of generality, assume u_i denotes a fragment representation for the fixed modality and v_j is the fragment representation for the other domain. Each weight, $w_{i,j}$, in the focal attention weight matrix, $W_A \in \mathbb{R}^{n\times m}$, is initialized as

$$w_{i,j} = \sigma(\alpha \frac{u_i^T v_j}{\|u_i\|\|v_j\|}), i \in [1, .., m], j \in [1, .., n], \tag{2}$$

where σ is a softmax function and α is a scaling factor.

ii) *Identify relevant fragments*. The relevance of a fragment is determined by comparing its attention score with other fragments. The relevance score, $H(w_{i,j})$, for v_j with respect to u_i is calculated as:

$$F(w_{i,j}) = \sum_{t=1}^{n} |w_{i,j} - w_{i,t}| \times g(w_{i,j})$$
$$H(w_{i,j}) = \mathbb{I}(F(w_{i,j}) > 0), \tag{3}$$

where $g(.)$ denotes confidence of the fragment being compared and it is derived as $\sqrt{w_{i,j}}$. $\mathbb{I}(.)$ is an indicator function.

iii) *Reassign attention.* Attention scores are recalculated as:

$$w'_{i,j} = \frac{w_{i,j} H(w_{i,j})}{\sum_{t=1}^{n} w_{i,t} H(w_{i,t})}, \tag{4}$$

The attended representation with respect to fixed domain fragment u_i is obtained by $v'_i = \sum_{j=1}^{n} w'_{i,j} v_j$. The global relevance of u and v is measured as:

$$S(u, v) = \frac{1}{m} \sum_{i=1}^{m} D(u_i, v'_i), \tag{5}$$

where $D(.)$ is Cosine similarity. This method is applied for both image-to-text and text-to-image direction.

Objective Function Following previous work (Nam et al., 2017; Lee et al., 2018; Liu et al., 2019a; Li et al., 2019) we have used a structured ranking loss (Karpathy et al., 2014; Karpathy and Fei-Fei, 2017) with maximum of hinges (Faghri et al., 2018) as the objective function. Instead of considering all the negatives, this triplet based loss function will focus on hard negatives. For a matching pair of image-text (I, T), loss, L, is computed as

$$L = max(\delta - S_{IT} + S_{\bar{I}T}, 0) + max(\delta - S_{IT} + S_{I\bar{T}}, 0) \tag{6}$$

where $\bar{I}$ and $\bar{T}$ are the hard negatives, and δ is the margin. For computational efficiency hard negatives are found within each mini-batch, instead of the entire training set.

Cross-modal Retrieval At inference time, the query is divided into its component propositions. Each proposition is used for the retrieval task and generates a similarity score for their retrieved candidates. The final score for a candidate is calculated as the sum of scores over all propositions.

4. Experiments

4.1. Experimental Set up

Data Preparation. We used the Flickr30k Entities dataset (Plummer et al., 2017) for all our experiments. This dataset is built upon the Flickr30k dataset (Young et al., 2014) which contains $31,000$ images annotated with five sentences each. In the Entities dataset, each mention in each sentence is linked to one or more bounding boxes in the image. We use the training-validation-test splits provided. Flickr30k does not provide ground truth SRLs. We generate semantic roles for Flickr30k entities using automatic SRL parsing on the gold captions(Gung and Palmer, 2021). The corresponding bounding boxes are marked with the detected SRL of the text mentions. In our experiments, we automatically associate each bounding box with the image region with maximum IoU among those detected by the faster RCNN. The distribution of the most frequent semantic roles in Flickr30k is presented in Table 1. To evaluate performance of text SRL (Gung and Palmer, 2021) we randomly sampled 500 sentences and divided them into 5 sets. Each set was then assigned to human annotators (groups of 2) for manual checking and correction. Assuming these human corrected SRLs as gold-SRLs, the overall F1 score of the SRL annotator was 90.9.

The other standard benchmarking dataset for image-text matching is MSCOCO (Lin et al., 2014). To the best of our knowledge, MSCOCO does not have any mappings between the entity mentions and image bounding boxes, obviating our use of this data for our current experiments.

Evaluation. We evaluate our performance based on Recall@K(K=1, 5, 10) for text-to-image and image-to-text retrieval tasks. Recall@K is computed as the proportion of correct images or text segments being retrieved among the top K results.

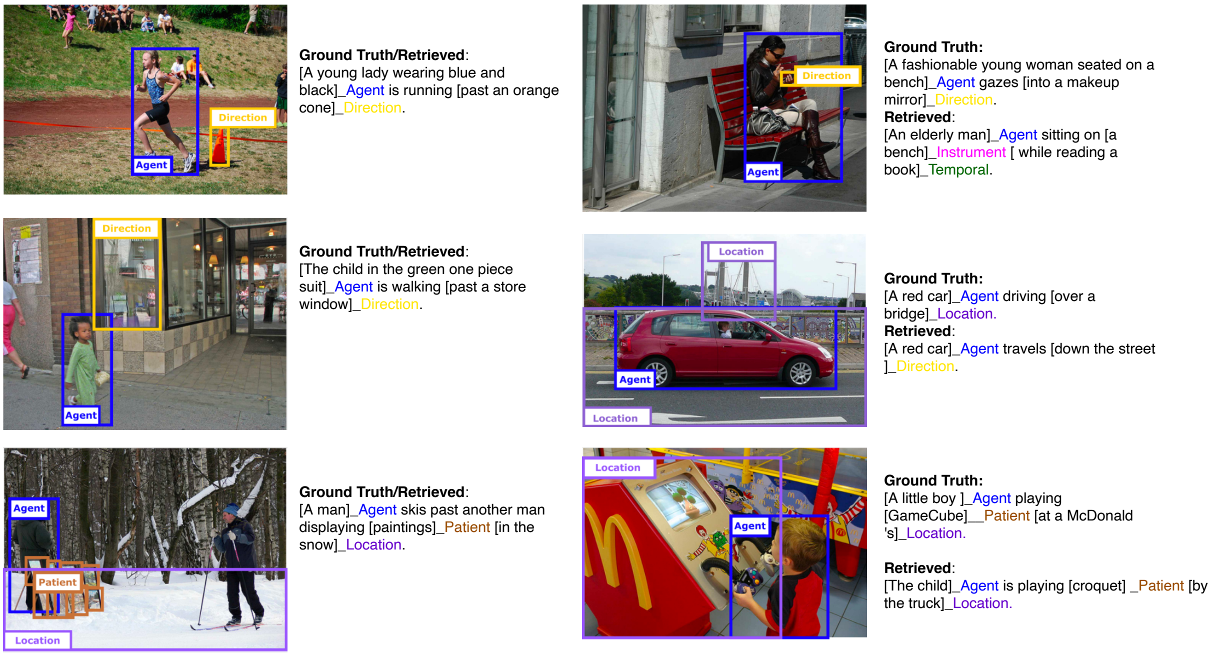
Setting. RARE is trained for 30 epochs with a 0.0002 learning rate for the first 15 epochs and then the learning rate is decayed by 0.1. We used the Adam optimizer (Kingma and Ba, 2014). We set the margin hyper-parameter δ in Equation 6 as 0.2. The best model over 30 epochs is selected based on the sum of the recalls on the validation set. Experiments are conducted using Nvidia Titan Xp GPUs.[1]

4.2. Quantitative Results

Comparison with Non-transformer Based Methods A quantitative comparison with recent approaches on the Flickr30k Entities benchmark is presented in Table 2. While there are many existing cross-modal retrieval results, we chose the subset in Table 2 based on state-of-the-art performance (Huang and Wang, 2019; Li et al., 2019) and their relatedness to our method (Liu et al., 2019a; Wang et al., 2020b; Liu et al., 2020). Transformer-based methods are discussed in the following section. For text-to-image retrieval (a.k.a image retrieval), RARE has the best $R@1$ score with a relative improvement of 13.7% compared to the next best VSRN model (Li et al., 2019).

For image-to-text retrieval (a.k.a text retrieval) RARE outperformed all other methods except ACMM (Huang and Wang, 2019). VSRN is one of the few architectures that encodes semantics. However, (Li et al., 2019) only computed global semantic reasoning in image-space, and not text representation. Syntactic information is explored in other approaches (Karpathy et al., 2014;

[1] code can be found at https://github.com/abhidipbhattacharyya/SRL_aware_ret.



Ground Truth/Retrieved:
[A young lady wearing blue and black]_Agent is running [past an orange cone]_Direction.

Ground Truth/Retrieved:
[The child in the green one piece suit]_Agent is walking [past a store window]_Direction.

Ground Truth/Retrieved:
[A man]_Agent skis past another man displaying [paintings]_Patient [in the snow]_Location.

(a) Correct Image-to-Text Retrievals

Ground Truth:
[A fashionable young woman seated on a bench]_Agent gazes [into a makeup mirror]_Direction.
Retrieved:
[An elderly man]_Agent sitting on [a bench]_Instrument [while reading a book]_Temporal.

Ground Truth:
[A red car]_Agent driving [over a bridge]_Location.
Retrieved:
[A red car]_Agent travels [down the street]_Direction.

Ground Truth:
[A little boy]_Agent playing [GameCube]__Patient [at a McDonald 's]_Location.

Retrieved:
[The child]_Agent is playing [croquet] _Patient [by the truck]_Location.

(b) Incorrect Image-to-Text Retrievals

Figure 3: Image-to-Text retrieval by RARE. For the incorrect retrievals in Fig 3b, either the SRL does not match the query, or in the case of the bottom right image, the SRL matches but there are incorrectly identified objects.

Role	Description of Role	Dataset	Image to Text		Text To Image	
		N	N	R@1	N	R@1
Agent	object which instigates the verb	158969	4690	0.96	4985	0.94
Patient	object which is affected by the verb	161841	4187	0.96	5025	0.82
Instrument	object which affects the verb	63853	1468	0.89	1967	0.72
Location	location of object or action	47866	910	0.85	1482	0.60
Temporal	describes time	17458	406	0.93	574	0.67
Direction	direction of motion	18933	316	0.84	600	0.50
Manner	manner of performing an action	15503	306	0.73	457	0.56
Predication	adjunct of an action	3698	74	0.81	101	0.64
Purpose	purpose of an action	2999	58	0.85	108	0.48
Companion	who an action was done with	1618	47	0.85	55	0.69
Start	starting position of action	1705	32	0.81	47	0.53

Table 1: Table summarizes SRL distribution over Flickr30k dataset and performance of RARE for specific SRLs. N denotes number of occurrences of an SRL in query.

Liu et al., 2020). Text graphs in GSMN (Liu et al., 2020) use syntactic rather than semantic labels, which may explain the better performance of RARE on image retrieval. ACMM, the current best performing non-transformer-based method, uses a more advanced network and memory unit to address less frequent fragments. However, even with a simpler architecture, RARE outperforms ACMM on image retrieval.

Comparison with Transformer Based Methods Table 3 represents comparison with recent transformer based methods. Success of transformers (Devlin et al., 2019; Liu et al., 2019b) in learning rich semantic and structural information from large, unlabelled data sources in a self-supervised manner and their ability to transfer learning from pre-trained tasks to fine-tuned tasks make them a potential architecture for the cross modal retrieval task (Ren et al., 2021; Chen et al., 2020b; Wen et al., 2021; Li et al., 2021). While these methods have superior performance to RARE on the standard retrieval task, there are no explicit ways to control the retrieved results. Moreover transformer based systems require larger amounts of training data and more significant computing resources. For example, many of these techniques require tens of thousands of hours of training using clusters of GPUs (Ren et al., 2021; Wen et al., 2021; Li et al., 2021). We believe that the current work demonstrates the potential of SRL in cross-modal retrieval tasks and could easily be ex-

Model	Text to Image			Image To Text		
	R1	R5	R10	R1	R5	R10
PFAN (Wang et al., 2019)	50.4	78.7	86.1	70	91.8	95.1
GVSE (Ren et al., 2016)	50.6	79.8	87.6	68.5	90.9	95.5
BFAN (Liu et al., 2019a)	50.8	78.4	-	68.1	91.4	-
SGM (Wang et al., 2020b)	53.5	79.6	86.8	71.8	91.7	95.5
ACMM (Huang and Wang, 2019)	53.8	79.8	-	**85.2**	**96.7**	-
VSRN (Li et al., 2019)	54.7	81.8	88.2	71.3	90.6	96
CVSE (Wang et al., 2020a)	52.9	80.4	87.8	73.5	92.1	95.8
GSMN (Liu et al., 2020)	57.4	82.3	89.0	76.4	94.3	97.3
RARE (ours)	**67.8**	**83.0**	**88.4**	76.3	93.4	96.6

Table 2: Comparison with other approaches on Flickr30k. Results sorted from worst to best R@1 on the Text-to-Image task. RARE also has second best R@1 performance on the Image-to-Text task.

Model	Text to Image			Image To Text		
	R1	R5	R10	R1	R5	R10
RARE (ours)	67.8	83.0	88.4	76.3	93.4	96.6
(Chen et al., 2020b)	76.0	**93.4**	**96.7**	85.8	97.8	98.8
(Ren et al., 2021)	**76.3**	93.3	95.6	**88.3**	**98.6**	99.3

Table 3: Comparison with transformer based approaches on Flickr30k.

Model	Text to Image			Image To Text		
	R1	R5	R10	R1	R5	R10
Model Ablation						
BFAN base model	53.5	79.6	73.4	72.6	93	96
+ SRL encodings	65.1	79.8	86.9	74.2	93.1	96.5
+ GCN	**67.8**	**83**	**88.4**	**76.3**	**93.4**	**96.6**
Input Ablation						
Image SRL only	40.9	45	58	43.8	76.5	86.3
Text SRL only	36.9	36.9	49	40.8	69.6	80.8
Both	**67.8**	**83**	**88.4**	**76.3**	**93.4**	**96.6**

Table 4: Ablation Studies on Flickr30k showing the effect of alterations to the network architecture and partial input information on retrieval recall

tended to use transformer-based features.

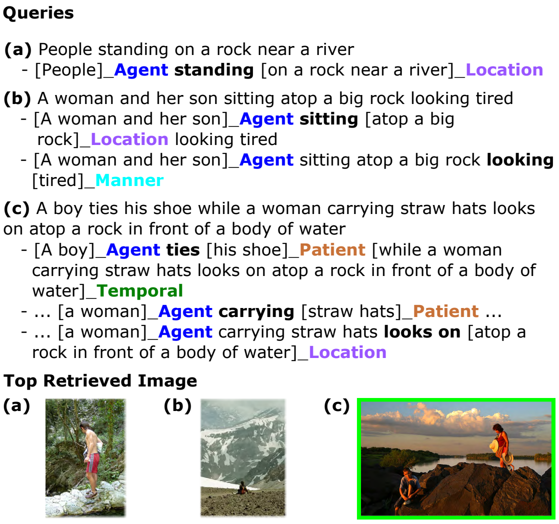
Contribution of SRLs Table 1 breaks down the retrieval results by SRL. The adherence of SRLs from the query to those in the retrieved results is better in image-to-text than text-to-image. We hypothesize that this improved adherence enables the better performance for image-to-text, as evidenced by the ablation study below. One possible factor accounting for the difference could be the graph convolutional network we use for image encoding. Our text encoder, in contrast, is a bi-directional GRU which may retain less information about important global relationships. In addition, the poorest text-to-image retrievals often involve infrequent peripheral event modifiers, like *Manner* or *Purpose*. In text these are often quite vague, indicating fairly implicit, non-concrete referents in the image that fail to generalize.

Ablation Study To validate each component of RARE, we present an ablation study in Table 4. We started with re-implementation of a BFAN-prob based system (Liu et al., 2019a). In the next model, we add semantic role encoding to the base system. Introduction of SRL boosts the performance on $R@1$ by 1.6% for text retrieval and 11.7% in image retrieval. Our final system incorporates the graph-CNN region encoder. This brings an additional 2% performance gain in $R@1$. In our input ablation study we provide SRL for the text or image component of the input only. The results shown in Table 4 confirm that SRL are needed for both input modalities to improve the alignment. When presented with SRL information for only one modality, the system was unable to match the performance of BFAN-prob based system. In the absence of corroborating SRL information from the other modality, we believe the SRLs introduce noise.

4.3. Qualitative Results

The main attraction of RARE is control over the retrieved results, which can differ depending on the SRL provided in the query. A unique feature of RARE is its ability to inject SRLs in the query and control the retrieval result at a fine-granularity for both text-to-image and image-to-text retrieval. Returning to Figure 1, we see a simple example of fine-grained text retrieval. In all three retrievals, the query image is the same, but the SRLs are different. All retrieved descriptions belong to the list of five ground truth descriptions for that image in Flickr30k, but each of the queries retrieves a different ground truth description which also matches the provided SRL.

A text description can contain multiple propositions and therefore multiple SRL label sets. RARE can also use multiple sets as part of the query. In Figure 5, the bounding boxes of the first three images correspond to three distinct propositions, and all three are used together to retrieve a description. This 3-part query correctly retrieves the ground truth caption. When presented with less annotated information, in Query 2, the system retrieved a shorter caption accordingly, as shown in the last image of Figure 5. Similar control can be exerted in the text-to-image retrieval direction.



Queries

(a) People standing on a rock near a river
- [People]_**Agent** **standing** [on a rock near a river]_**Location**

(b) A woman and her son sitting atop a big rock looking tired
- [A woman and her son]_**Agent** **sitting** [atop a big rock]_**Location** looking tired
- [A woman and her son]_**Agent** sitting atop a big rock **looking** [tired]_**Manner**

(c) A boy ties his shoe while a woman carrying straw hats looks on atop a rock in front of a body of water
- [A boy]_**Agent** **ties** [his shoe]_**Patient** [while a woman carrying straw hats looks on atop a rock in front of a body of water]_**Temporal**
- ... [a woman]_**Agent** **carrying** [straw hats]_**Patient** ...
- ... [a woman]_**Agent** carrying straw hats **looks on** [atop a rock in front of a body of water]_**Location**

Top Retrieved Image

Figure 4: Fine grained text-to-image retrieval by RARE. From top to bottom, the descriptions are more informative, with an increasing number of propositions. Verbs for each proposition are shown in bold. All three captions correspond to image (c) in the ground truth, but the other retrieved images also reflect the queries.

An example is shown in Figure 4. The query sentences (a,b,c) have increasingly complex structure. The third, most precise query, correctly retrieves the ground truth image, while the other queries retrieve other relevant top results.

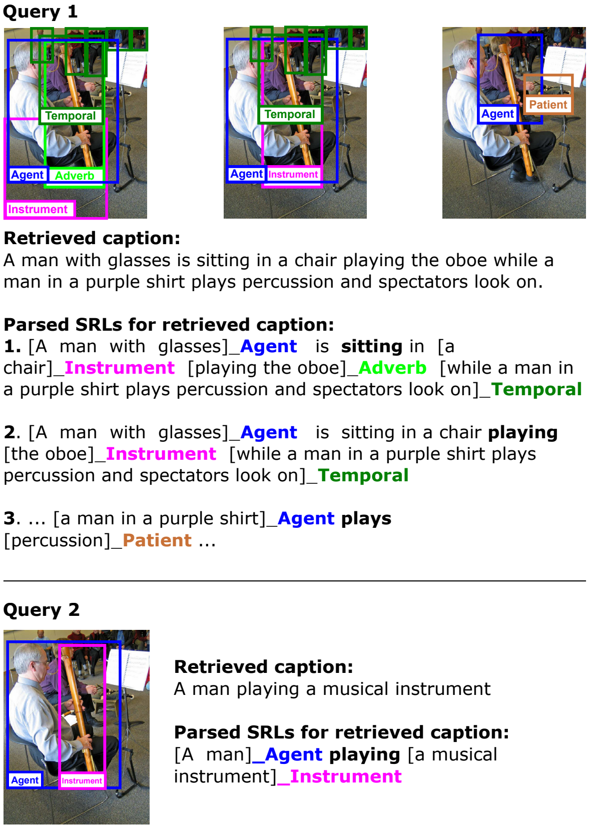
Errors made by RARE are reasonable. Figure 3 depicts some examples for text retrieval. Figure 3a shows examples where the system retrieved the ground truth caption. Figure 3b depicts the quality of retrieved captions in unsuccessful retrievals. Despite the mismatches, retrieval results for these images are coherent with the query images. Specifically for the bottom right example, RARE is able to match the semantics of `play` and `location` although it misidentifies the game.

5. Future Work

A crucial next step is implementation of a fully automatic visual SRL detector for previously unseen images, to enable multimodal retrieval without human annotation. For example, using a visual SRL detector together with the text SRL parser used in this work, one could use a dataset which lacks any SRL annotations, e.g. the MSCOCO dataset, as an additional benchmark for RARE. To address infrequent SRLs, we will further investigate the applicability of a memory network in RARE. We would also like to explore application of SRL in more advanced architectures (Liu et al., 2020; Wehrmann et al., 2020) for richer representations.

6. Conclusion

In this paper, we propose **r**ole **a**ware **r**etri**e**val (RARE) for cross-modal retrieval. This work demonstrates that incorporating semantic role labeling can improve the performance of cross-modal retrieval. When we evaluate RARE on Flickr30k, RARE achieves competitive performance for the image retrieval task against the best non-transformer based system. Although transformer based methods are the current state-of-the-art for the retrieval task, they do not allow explicit control signals to diversify their recommendations. The incorporation of semantic roles into those architectures could provide that benefit, but will require automatic annotation of large amounts of silver standard training data which is one of our goals. Our qualitative results presented here have shown the potential of RARE for fine-grained retrieval achieved by injecting semantic role labels into the retrieval query, to guide the retrieved caption or image to the desired focus. The semantic roles allow the system to choose more concise salient descriptions, or alternatively, to retrieve long complex multi-predicate descriptions. Humans can effortlessly generate a large variety of descriptions for images. Now, RARE provides an automated solution for retrieving more varied and fine-grained retrieval results.



Query 1

Retrieved caption:
A man with glasses is sitting in a chair playing the oboe while a man in a purple shirt plays percussion and spectators look on.

Parsed SRLs for retrieved caption:

1. [A man with glasses]_**Agent** is **sitting** in [a chair]_**Instrument** [playing the oboe]_**Adverb** [while a man in a purple shirt plays percussion and spectators look on]_**Temporal**

2. [A man with glasses]_**Agent** is sitting in a chair **playing** [the oboe]_**Instrument** [while a man in a purple shirt plays percussion and spectators look on]_**Temporal**

3. ... [a man in a purple shirt]_**Agent** **plays** [percussion]_**Patient** ...

Query 2

Retrieved caption:
A man playing a musical instrument

Parsed SRLs for retrieved caption:
[A man]**_Agent** **playing** [a musical instrument]**_Instrument**

Figure 5: Semantic role aware text retrieval by RARE. In Query 1, RARE uses three sets of SRL annotations on the image as the query. A caption is retrieved that contains three propositions. In Query 2, a simpler query on the same image uses a single set of SRL annotations. The retrieved caption is likewise simpler and contains a single proposition.

7. Bibliographical References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. pages III–1247–III–1255.

Chen, J., Zhang, L., Bai, C., and Kpalma, K. (2020a). Review of recent deep learning based methods for image-text retrieval. pages 167–172.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020b). Uniter: Universal image-text representation learning. In *ECCV*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Faghri, F., Fleet, J. D., Kiros, R. J., and Fidler, S. (2018). VSE++: Improving visual-semantic embeddings with hard negatives.

Fan, H. and Zhou, J. (2018). Stacked latent attention for multimodal reasoning. pages 1072–1080.

Gong, Y., Ke, Q., Isard, M., and Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. 106(2):210–233, January.

Hacioglu, K., Pradhan, S., Ward, W., Martin, J. H., and Jurafsky, D. (2004). Semantic role labeling by tagging syntactic chunks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 110–113, Boston, Massachusetts, USA, May 6 - May 7. Association for Computational Linguistics.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483. Association for Computational Linguistics.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. 47(1):853–899, May.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3-4):321–377, 12.

Huang, Y. and Wang, L. (2019). Acmm: Aligned cross-modal memory for few-shot image and sentence matching. October.

Karpathy, A. and Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. 39(4):664–676, April.

Karpathy, A., Joulin, A., and Fei-Fei, L. (2014). Deep fragment embeddings for bidirectional image sentence mapping. pages 1889–1897.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. 12.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. 123(1):32–73, May.

Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. (2018). Stacked cross attention for image-text matching. September.

Li, K., Zhang, Y., Li, K., Li, Y., and Fu, Y. (2019). Visual semantic reasoning for image-text matching.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. (2021). UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. 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)*, pages 2592–2607, Online, August. Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context.

Liu, C., Mao, Z., Liu, A.-A., Zhang, T., Wang, B., and Zhang, Y. (2019a). Focus your attention: A bidirectional focal attention network for image-text matching. page 3–11.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., and Zhang, Y. (2020). Graph structured network for image-text matching. pages 10921–10930.

Ma, L., Lu, Z., Shang, L., and Li, H. (2015). Multimodal convolutional neural networks for matching image and sentence. pages 2623–2631.

Nam, H., Ha, J.-W., and Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. July.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. 123(1):74–93.

Pradhan, S., Hacioglu, K., Ward, W., Martin, J. H., and Jurafsky, D. (2005). Semantic role chunking combining complementary syntactic views. In *Proceedings of the Ninth Conference on Computational*

Natural Language Learning (CoNLL-2005), pages 217–220, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Punyakanok, V., Roth, D., and Yih, W.-t. (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. volume 28, pages 91–99.

Ren, Z., Jin, H., Lin, Z., Fang, C., and Yuille, A. (2016). Joint image-text representation by gaussian visual-semantic embedding. page 207–211.

Ren, S., Lin, J., Zhao, G., Men, R., Yang, A., Zhou, J., Sun, X., and Yang, H. (2021). Learning relation alignment for calibrated cross-modal retrieval. 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)*, pages 514–524, Online, August. Association for Computational Linguistics.

Socher, R. and Fei-Fei, L. (2010). Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. pages 966–973.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. 2:207–218.

Strubell, E., Verga, P., Andor, D., Weiss, D., and McCallum, A. (2018). Linguistically-informed self-attention for semantic role labeling. *CoRR*, abs/1804.08199.

Tan, Z., Wang, M., Xie, J., Chen, Y., and Shi, X. (2017). Deep semantic role labeling with self-attention. *CoRR*, abs/1712.01586.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wang, L., Li, Y., and Lazebnik, S. (2016). Learning deep structure-preserving image-text embeddings. pages 5005–5013.

Wang, L., Li, Y., and Lazebnik, S. (2018). Learning two-branch neural networks for image-text matching tasks. 41:394–407.

Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., and Fan, X. (2019). Position focused attention network for image-text matching. pages 3792–3798, 7.

Wang, H., Zhang, Y., Ji, Z., Pang, Y., and Ma, L. (2020a). Consensus-aware visual-semantic embedding for image-text matching. pages 18–34. Springer.

Wang, S., Wang, R., Yao, Z., Shan, S., and Chen, X. (2020b). Cross-modal scene graph matching for relationship-aware image-text retrieval. March.

Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, R. H., Liu, W., Chauhan, A., Guan, Y., Li, B., Li, R., Song, X., Fung, Y., Ji, H., Han, J., Chang, S.-F., Pustejovsky, J., Rah, J., Liem, D., ELsayed, A., Palmer, M., Voss, C., Schneider, C., and Onyshkevych, B. (2021). COVID-19 literature knowledge graph construction and drug repurposing report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 66–77, Online, June. Association for Computational Linguistics.

Wehrmann, J., Kolling, C., and Barros, R. C. (2020). Adaptive cross-modal embeddings for image-text alignment. pages 12313–12320. AAAI Press.

Wen, K., Xia, J., Huang, Y., Li, L., Xu, J., and Shao, J. (2021). Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2188–2197.

Wu, Y., Wang, S., and Huang, Q. (2018). Learning semantic structure-preserved embeddings for cross-modal retrieval. page 825–833.

Yan, F. and Mikolajczyk, K. (2015). Deep correlation for matching images and text. pages 3441–3450, June.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. 2:67–78.

Zhou, J. and Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China, July. Association for Computational Linguistics.

8. Language Resource References

Gung, J. and Palmer, M. (2021). Predicate representations and polysemy in verbnet semantic parsing. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 51–62, Groningen, The Netherlands (online), June. Association for Computational Linguistics.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada, July. Association for Computational Linguistics.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. 31(1):71–106, March.