RECAP: Semantic Role Enhanced Caption Generation

Abhidip Bhattacharyya, Martha Palmer, Christoffer Heckman
University of Massachusetts Amherst, University of Colorado Boulder, University of Colorado Boulder
abhidipbhatt@umass.edu, {martha.palmer, christoffer.heckman}@colorado.edu

Abstract

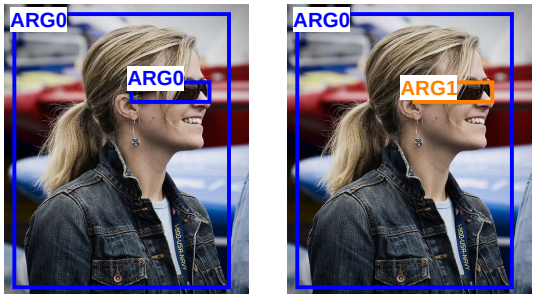
Even though current vision language (V+L) models have achieved success in generating image captions, they often lack specificity and overlook various aspects of the image. Additionally, the attention learned through weak supervision operates opaquely and is difficult to control. To address these limitations, we propose the use of semantic roles as control signals in caption generation. Our hypothesis is that, by incorporating semantic roles as signals, the generated captions can be guided to follow specific predicate argument structures. To validate the effectiveness of our approach, we conducted experiments using Flickr30k data and compared the results with a baseline model VL-BART(Cho et al., 2021a). The experiments showed a significant improvement, with a gain of 45% in Smatch score (Standard NLP evaluation metric for semantic representations), demonstrating the efficacy of our approach. By focusing on specific objects and their associated semantic roles instead of providing a general description, our framework produces captions that exhibit enhanced quality, diversity, and controllability.

Keywords: semantic role labeling, caption generation, controlled caption generation

1. Introduction

Automatic caption generation is a complex task that combines computer vision and natural language processing to generate meaningful descriptions of images. This research area has attracted considerable attention from various studies (Duygulu et al., 2002; Farhadi et al., 2010; Kulkarni et al., 2011; Vinyals et al., 2014; Karpathy et al., 2014; Karpathy and Fei-Fei, 2017). Attention-based machine translation methods (Bahdanau et al., 2015) have influenced the integration of image attention mechanisms in generating caption words (Xu et al., 2015; Huang et al., 2019; Anderson et al., 2018). Recently, with the availability of large-scale image-text datasets and advancements in transfer learning through weak supervision, transformer-based vision-language pretraining approaches (Tan and Bansal, 2019a; Lu et al., 2019; Li et al., 2019b; Su et al., 2020; Li et al., 2021, 2020; Chen et al., 2020b) have significantly enhanced caption generation performance.

The success of large-language models (LLMs) and vision+language (V+L) models has led to a shift in focus away from more fundamental components, such as semantics. The diversification of generated descriptions in caption generation is done by tweaking the model parameters such as sampling temperature, number of beams, and beam length, etc. Therefore, variation of captions is often limited to the surface level of the sentence, neglecting semantic diversification. V+L research was not always this indifferent to semantics. Semantic diversification was addressed in previous work(Cornia et al., 2019; Yao et al., 2018; Chen et al., 2020a). In this paper, we revisit the significance of semantics in caption generation.



A blond woman in sunglasses was laughing.

A blond woman was wearing sunglasses.

Figure 1: An example from Flickr30k dataset. The image on the left has `ARG0` for both the woman and the sunglass. The generated caption has an intransitive verb. In the next image the sun-glass in `ARG1`. The generated caption uses `wearing` with sun-glass as object.

The predicate-argument structure in computational linguistics is a form of semantic parsing that conveys knowledge about *who is doing what to whom when*. Semantic role annotation, based on paradigms such as PropBank or Framenet (Palmer et al., 2005; Fillmore et al., 2003), is used to train semantic parsers for natural language processing. In other words, given an action in a sentence, it identifies who is performing the action (ARG0), who is affected by the action (ARG1), what instrument is being used (ARG2), etc. to comprehend the meaning of the sentence. The task of marking word spans with semantic roles is called **S**emantic **R**ole **L**abeling (SRL) (Màrquez et al., 2008; Padó and Lapata, 2009; Kozhevnikov and Titov, 2013; Akbik et al., 2015)(Jindal et al., 2022). Recent advances in deep neural architecture has improved the quality of SRL annotation (Jia et al., 2022; Blloshmi et al.,

LREC-COLING 2024, pages 13633–13649

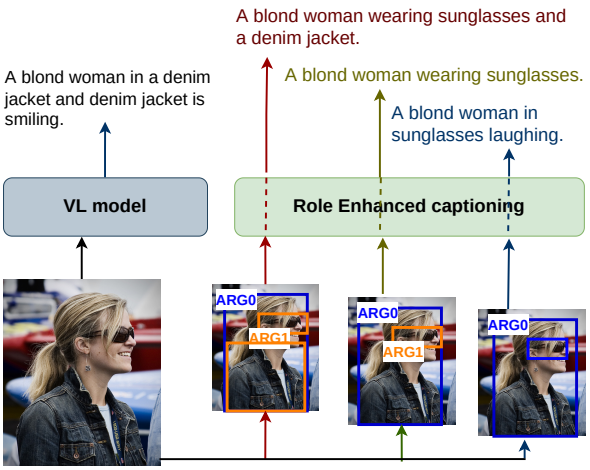


Figure 2: A typical caption generation model will present a holistic description. RECAP can generate different captions based on the semantic roles of the object bboxes in the image

2021; Fei et al., 2021).

In this paper, we propose a framework called Semantic Role Enhanced Caption Generation (RECAP), where we utilize SRL as a control signal in the image captioning process. The framework takes an input image with bounding boxes (bboxes) annotated with SRL information. This allows for the creation of variations of the same image by altering the SRLs assigned to the bboxes. The generated captions will then reflect the corresponding predicate-argument structures presented in the bbox annotations. For instance, in Figure 1, the image on the left shows a `woman` and `sunglasses` both labeled as `ARG0`. As a result, the generated caption uses an intransitive verb, with the `woman` and `sunglasses` being part of the agent. In contrast, the second image has the `sunglasses` annotated as `ARG1`, and the generated caption depicts the required predicate-argument structures by employing a transitive verb. Our key contributions can be summarized as follows.

- We have used semantic roles as control signals to generate focused image descriptions.
- SRL-informed models can generate diverse captions based on different linguistic foci.
- The explanation of the generated caption stems from its grounding in linguistics, which can be attributed to the provided SRL annotation of the image.
- We used the Smatch score (Cai and Knight, 2013) as our reinforcement signal for better grounding of the predicate-argument structure.

2. Related Work

Initial approaches to caption generation can be categorized into two types: 1) bottom-up and 2) top-down. In bottom-up approaches, information about object relationships and attributes is detected, and this information is then used to generate captions using predefined templatess (Yang et al., 2011; Farhadi et al., 2010; Kulkarni et al., 2011). However, template-based methods often lack novelty in their generated captions. On the other hand, top-down approaches involve language modeling that is conditioned on image fragments using soft attention mechanisms (Vinyals et al., 2014; Xu et al., 2015; Mao et al., 2014; You et al., 2016; Yao et al., 2018). Anderson et al. (2018) proposed use of both bottom-up and top-down approaches (BUTD).

In recent years, the availability of large-scale image-text datasets, advancements in transfer learning through weak supervision, and the development of transformer-based V+L pretraining approaches (Tan and Bansal, 2019a; Lu et al., 2019; Li et al., 2019b; Su et al., 2020; Li et al., 2021, 2020; Chen et al., 2020b) have significantly improved the performance of caption generation. Additionally, the remarkable success of language generation models such as LLMs (Language Models)(Raffel et al., 2020; Lewis et al., 2020; Wei et al., 2022; Brown et al., 2020) and V+L models (Radford et al., 2021; Jia et al., 2021) has led to the development of socratic-style models (Zeng et al., 2023). Several models in this field (Mokady et al., 2021; Li et al., 2023; Dai et al., 2023; Alayrac et al., 2022) have employed a frozen V+L model encoder and frozen LLMs, training interface layers that map representations from the output domain of the V+L model to the input domain of the LLMs. However, these models primarily function as black boxes, limiting user control over the generation process.

In the past, attempts have been made to introduce some form of control, either through guidance over attention (Huang et al., 2019; Zhou et al., 2020) or by manipulating the object sequence (Cornia et al., 2019). Scene graphs (Xu et al., 2017; Tang et al., 2020, 2018) presented a potent tool to diversify the generated captions (Yao et al., 2018; Chen et al., 2020a). But semantic comprehension with scene graphs is insufficient for image articulation (Cho et al., 2022b; Cheng et al., 2022). On the other hand, predicate-argument structures such as semantic roles (Palmer et al., 2005; Fillmore et al., 2003) are effective computational linguistic tools in meaning representation. Given an action in a sentence, semantic role identifies who is performing the action (ARG0), who is affected by the action (ARG1), what instrument is being used (ARG2), etc. to comprehend the meaning of the sentence. The task of automatic annotation of se-

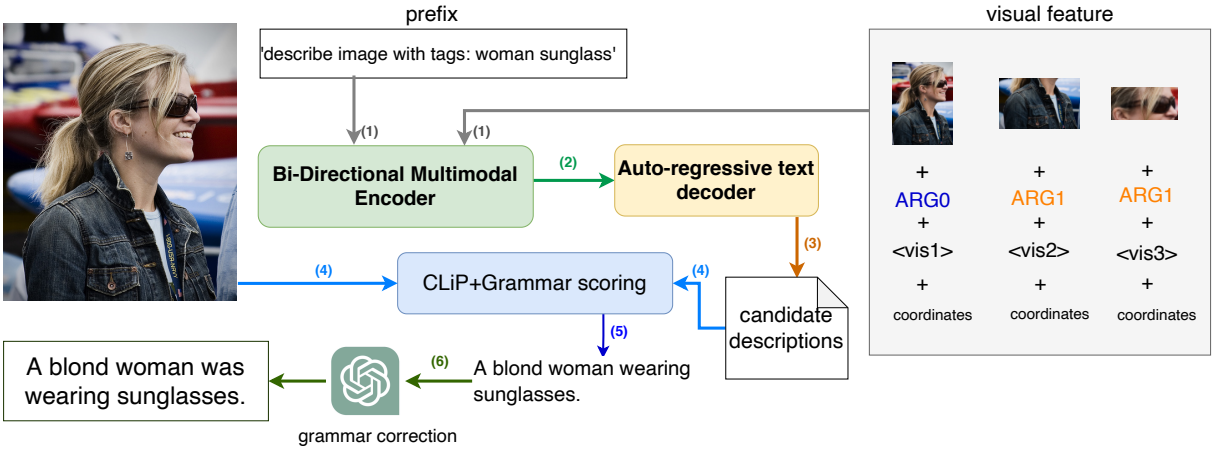


Figure 3: Overview of RECAP. 1) text embedding and visual feature embedding are fed to the encoder 2) encoded information fed to the decoder 3)decoder generates candidate descriptions 4)candidate sentences and the image is fed to CLIP-S+grammar model for scoring 5) sentence with highest score is selected and 6) fed to GPT3 for further grammar correction.

mantic roles is known as semantic role labeling, and it is a well-researched problem in the domain of Computational Linguistics and NLP (Gildea and Palmer, 2002; Pradhan et al., 2005; Zhou and Xu, 2015; Strubell et al., 2018; Blloshmi et al., 2021; Jia et al., 2022; Fei et al., 2021). In this paper we are proposing semantic **r**ole **e**nhanced **cap**tion generation (RECAP). Figure 2 shows an example of how caption generation by RECAP differs from current V+L models.

3. Approach

Our model is inspired by VL-BART (Cho et al., 2021a), which is an extended version of BART$_{Base}$ (Lewis et al., 2020). In VL-BART, the text encoder is modified to accept image region embeddings as additional input. The overall framework of RECAP is illustrated in Figure 3. The encoder takes in image bbox and SRL annotations, along with object labels as a prefix. An auto-regressive decoder generates candidate descriptions, with the number of beams determined by the hyper-parameter "num_of_beam". The candidate sentences undergo CLIPS+grammar scoring(Cho et al., 2022a), and the winning sentence is selected. This winning sentence is then passed to GPT3(Brown et al., 2020) for grammar correction.

3.1. Model

Visual Embedding The image representation is obtained using Faster RCNN (Ren et al., 2015) pre-trained on Visual Genome (Krishna et al., 2017), following prior work (Anderson et al., 2018; Li et al., 2020; Tan and Bansal, 2019b; Cho et al., 2021a). We use a set of region vectors ($R = r_1, r_2, ..., r_n$) detected by Faster RCNN to represent the image. Each region's representation (r_i) is derived from the mean-pooled convolutional feature. In our experiments, we use Faster RCNN features with a dimension of 2048 and select the top 36 regions based on class detection confidence scores. To optimize computation, we pre-computed the Faster RCNN features and utilized them in RECAP. To incorporate SRLs for the bboxes, we introduced an embedding layer. The process of obtaining SRLs for the bboxes is discussed in subsection 4.1. The ultimate vision embedding ($r^v = r_1^v, ..., r_2^v$) is obtained by summing the region feature, SRL embedding, region id embedding, and bbox position.

Text Embeddings The input text for caption generation in our model follows the approach used in VL-BART. As suggested by Li et al. (2020), the text prompt (t) is accompanied by the object labels of the image bounding boxes (bboxes). We utilize a shared text embedding layer that is used by the encoder, decoder, and the language modeling head. The embedded representation for the text input t is denoted as e^t. Additionally, special visual sentinel tokens were included to the vocabulary to obtain embeddings for the image region ids. These shared embeddings play a vital role in generating captions that are grounded in the bboxes (Cho et al., 2021a).

Traning vs Inference RECAP is a transformer-based encoder-decoder model. The encoder and the decoder consists of a stack of m multimodal transformers and residual layers. The encoder takes the concatenation of r^v and e^t as input and produces a contextualized joint representation, denoted as $h = \text{Enc}(r^v, e^t)$. The decoder attends to the encoder representation h through

cross-attention and to previously generated tokens through self-attention. During training, the decoder receives a version of the expected text sentence that is right-shifted by one token position. During inference, the decoder generates tokens conditioned on the previously generated tokens in an autoregressive manner. In other words, the decoder produces a probability distribution for the next token over the vocabulary, represented as $P_\theta(y_j|y_{<j}, t, v) = \text{Dec}(y_{<j}, h)$, where θ represents the model parameters. These parameters are learned by optimizing the negative log-likelihood (NLL) with respect to the ground truth tokens in the context of the input text t and image v.

$$\mathcal{L}_{XE}(\theta) = -\sum_{j=1}^{|y|} \log P_\theta(y_j|y_{<j}, t, v) \quad (1)$$

Reinforcement Learning To address the exposure bias issue of NLL loss (Ranzato et al., 2016), self-critical sequence training (SCST) (Rennie et al., 2017) has been successful in image captioning (Huang et al., 2019; Anderson et al., 2018; Cornia et al., 2019; Li et al., 2020). In SCST, the objective is to minimize the negative expected score, starting from a model trained with cross-entropy.

$$\mathcal{L}_r(\theta) = E_{y_{1:T} \sim P_\theta}[r(y_{1:T})] \quad (2)$$

where r is the reward function. Following (Rennie et al., 2017) the gradient can be calculated as-

$$\nabla_\theta \mathcal{L}_r(\theta) \approx -(r(y^s_{1:T}) - r(\hat{y}_{1:T}))\nabla_\theta \log P_\theta(y_{1:T}) \quad (3)$$

where $y^s_{1:T}$ is a sampled caption and $\hat{y}_{1:T}$ is obtained from current model with greedy decoding. Previous methods (Huang et al., 2019; Anderson et al., 2018; Cornia et al., 2019; Li et al., 2020) primarily relied on CIDEr (Vedantam et al., 2014) as a reward score to encourage caption generation similar to the reference captions. However, our objective is to guide the model to mimic the predicate-argument structure instead.

3.2. AMRs for focusing on Predicate-Argument structure

To encourage the model to identify predicate-argument structure we required an appropriate scoring function that can reward correct predicate-argument structure closest to the ground truth (GT). To facilitate comparisons among predicate-argument structures, we selected Abstract Meaning Representation (AMR) graphs. AMR graphs can be represented as sets of triplets, where each triplet consists of a relation and two arguments connected by the relation. The measure of propositional overlap between two AMRs involves counting the number of matching triplets. Due to the various possible mappings between triplets from any two AMRs, there can be multiple instances of propositional overlap. Smatch (Cai and Knight, 2013) calculates the maximum F1 score among these different mappings. For more details on Smatch, please refer to Cai and Knight (2013).

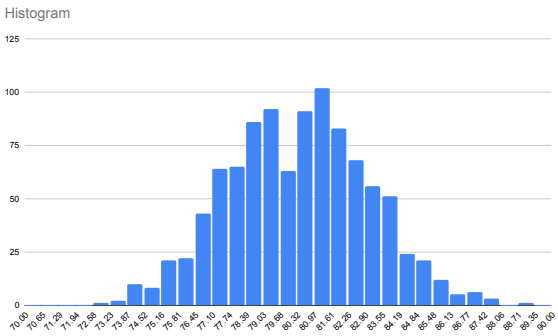


Figure 4: Histogram plot of F1 score of SRL(Shi and Lin, 2019) on Flickr30k. Human annotators corrected annotation for 4698 SRL propositions over 2676 sentences. We then randomly sampled 100 propositions and measured the F1 scores over 1000 iterations. Mean F1 score is 80.39 with std 2.66.

We used Smatch as our reinforcement reward metric. To generated AMR we employed the SPRING parser (Bevilacqua et al., 2021) trained with AMR 3.0 version of the LDC release (Knight et al., 2020). For evaluation of the parser on Flickr30k, we randomly selected 50 GT sentences from the Flickr30k dataset and generated AMRs using SPRING. Human experts also annotated the same 50 sentences. The Smatch score for the parser was 0.74. We performed a similar evaluation for the captions generated by our model, resulting in a Smatch score of 0.7.

4. Experiments

4.1. Experimental Set up

Data Preparation. We conducted our experiments using the Flickr30k Entities dataset (Plummer et al., 2017), which is derived from the Flickr30k dataset (Young et al., 2014). The Flickr30k dataset consists of $31,000$ images, each annotated with five sentences. In the Entities dataset, each mention in the sentences is linked to one or more bboxes in the corresponding image. We utilized the provided training-dev-test splits.

Since Flickr30k does not provide GT SRLs for image bboxes, we generated SRLs for Flickr30k entities by applying automatic SRL parsing to the gold captions (Shi and Lin, 2019). The performance of the SRL parser on Flickr30k, compared to human annotation, is shown in Figure 4. For each sentence, the SRL parser produces propositions

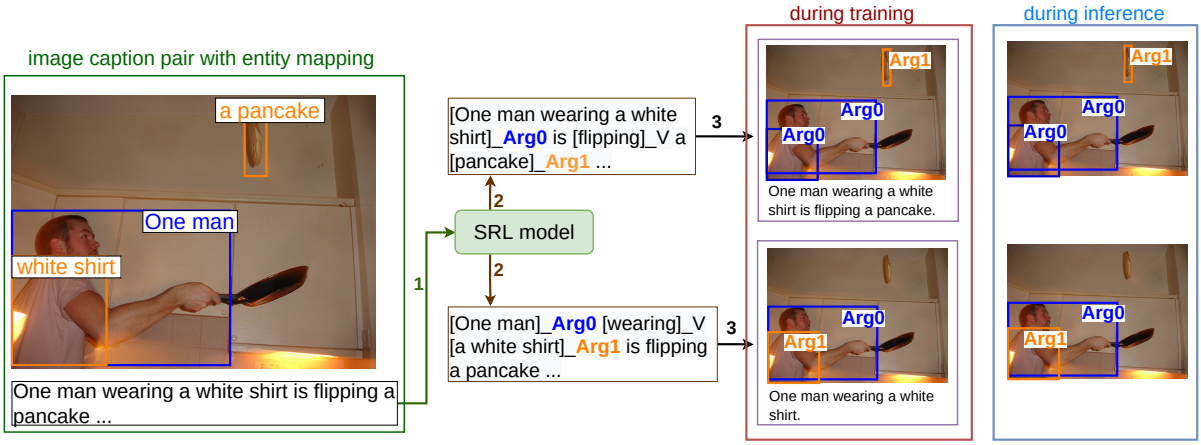


Figure 5: Process of bbox annotation creation for Flickr30k dataset using provided entity mapping. 1) GT caption is fed to SRL parser 2) SRL parser gives SRL propositions 3) Using entity mapping SRLs are transferred to bboxes and sub-sentences are created. During training, both the annotated image and sub-sentences are used. During inference we used only the SRL annotated images. Refer to subsection 4.1 for more details.

based on the number of predicates. We extract the corresponding sub-sentences for each SRL proposition. By mapping the SRLs to the bboxes using the entity mapping of the Flickr30k Entities dataset, we create annotations for the images. In our experiments, we associate each bbox with the image region that has the highest Intersection over Union (IoU) among those detected by the Faster RCNN. This results in a sub-sentence with an SRL-annotated image, which serves as our training data. The procedure is illustrated in Figure 5, where an annotated image-sentence pair is created for each predicate, such as `flipping` and `wearing`. During inference, we use the SRL-annotated images only.

At this point, our experiments are limited to Flickr30k because to the best of our knowledge, no other dataset has entity-bounding box mapping.

Implementation Details. We used the pre-trained VL-BART model ($H = 768$) and finetuned it for caption generation with Flickr30k. Our batch size was 8. We trained for 10 epochs with AdamW Optimizer (Loshchilov and Hutter, 2019) with learning rate 3×10^{-5}, $(\beta^1, \beta^2) = (0.9, 0.999)$, and $\epsilon = 1^{-8}$ with 5% linear warmup schedule. Our code will be public upon acceptance of the paper.

4.2. Results and Discussions

Quantitative Results. To investigate the impact of SRL, we conducted experiments using a basic BUTD model (Anderson et al., 2018) with SRL annotation. For ReCAP we started from the VL-BART (Cho et al., 2021a) model. Table 1 shows the performance of our caption generation system with respect to BLEU, METEOR, ROUGE and CIDEr (Papineni et al., 2002; Denkowski and Lavie, 2014; Lin, 2004; Vedantam et al., 2014). We also include Smatch scores (as previously defined in Section 3.2) in our Results table. The reader will immediately note that our results seem very mixed. Smatch and CIDEr show marked improvement, in contrast with BLEU, METEOR, ROUGE.

The introduction of SRL into the BUTD model resulted in a modest increase in the Smatch score, from 0.31 to 0.33. However, it's worth noting that a Smatch score of 0.33 can also be achieved with a vanilla BLIP model (Li et al., 2022). Furthermore, the vanilla BLIP2 (Li et al., 2023) (finetuned t5xl) model scored 0.35 Smatch score. The improvement in the Smatch score due to introduction of SRL was more significant in the case of VL-BART, with a 45% increase, raising the Smatch score from 0.35 to 0.51. These enhancements, observed in both a basic model like BUTD and an advanced model like VL-BART, highlight the effectiveness of SRL in generating the desired predicate-argument structure. Improvement in Smatch from BLIP to BLIP2 intrigued us about exploring large V+L models in context of SRL. However, at this moment we leave this as our future endeavour.

Training with SCST and utilizing the Smatch score resulted in even better performance. However, the use of CLIP-S+grammar (Cho et al., 2022a) in selecting candidate sentences did not improve performance in terms of predicate-argument structures. It's worth noting that the removal of CLIP-S selection from the pipeline slightly reduced the model's performance in other metrics. This is because the CLIP-S model was trained to select grammatically superior sentences with better object grounding, but it had no impact on SRLs.

The performance of ReCAP did not surpass the baseline in terms of BLEU, METEOR, and ROUGE

Model	Smatch	CIDEr	BLEU-4	METEOR	ROUGE
BLIP (base_coco) (Li et al., 2022)	0.33	0.761	0.274	0.237	0.507
BLIP(large_coco) (Li et al., 2022)	0.33	0.793	0.289	0.243	0.518
BLIP2(finetuned t5 xl) (Li et al., 2023)	0.35	0.947	0.337	0.264	0.557
TDBU (Anderson et al., 2018)	0.31	0.53	0.253	0.214	0.512
+SRL	0.33	0.384	0.206	0.185	0.481
RECAP (ours)					
VL-BART (Cho et al., 2021a)	0.35	0.66	0.27	0.23	0.50
+SRL	0.4	0.95	0.12	0.17	0.35
+CLIP-S(Cho et al., 2022a)	0.4	0.945	0.118	0.167	0.346
+amr scst	0.51	0.857	0.104	0.177	0.344
+amr scst −CLIP-S	0.51	0.823	0.099	0.175	0.341

Table 1: Comparison of RECAP with respect to different metrics discussed in subsection 4.2

Roles	Count	+SRL	+CLIP-S	+SCST	−CLIP-S
ARG0	10566	0.864	0.852	0.856	0.862
ARG1	9024	0.757	0.755	0.777	0.782
ARG2	2269	0.696	0.699	0.711	0.724
ARGM-LOC	1457	0.595	0.609	0.653	0.659
ARGM-TMP	945	0.692	0.702	0.695	0.692
ARGM-ADV	709	0.290	0.298	0.375	0.379
ARGM-DIR	532	0.682	0.678	0.651	0.663
ARGM-PRD	114	0.127	0.152	0.127	0.085
ARGM-COM	72	0.405	0.405	0.388	0.388
ARGM-PRP	69	0.386	0.454	0.367	0.326

Table 2: Comparison with respect to SRL grounding in the generated caption (Recall based scoring)

GT sentence: A woman kayaking in a yellow kayak with her dog.

a

Without SCST: A person kayaking on a lake with a dog.

b

With AMR SCST:A person in a blue wetsuit kayaking on a lake with a black and yellow dog.

c

Figure 6: Captions with low CIDEr score can have desired predicate-argument structure. a) GT caption. b)caption generated by RECAP without SCST. c) caption generated by RECAP with SCST

metrics. Moreover, recent systems like BLIP (Li et al., 2022) can achieve better results with these metrics while maintaining a similar Smatch score compared to the baseline model. It is important to note that these metrics evaluate the generated captions based on n-gram overlap with the GT captions. VL-BART and most V+L language models tend to provide a comprehensive description of the image, which often leads to better performance in terms of these metrics. Additionally, each generated caption is compared to five reference captions, resulting in a higher likelihood of n-gram overlaps. One surprising observation is that, while the BLIP large model demonstrated some improvement over the BLIP base model in terms of n-gram based metrics, their Smatch score remained the same. This suggests that these large models, despite generating captions with detailed object descriptions, may not capture predicate-argument structures any better.

In contrast, our model was trained to be sensitive to the image's SRL annotation. To achieve this, we used SRL-specific annotation for the image and corresponding sub-sentences as discussed in Section 4.1. As a result, the captions generated by RECAP are highly specific to the particular predicate-argument structure in the image, and we do not have five reference sentences for a particular SRL annotation. Therefore, the amount of n-gram overlap with the GT captions will be lower. This is also reflected in the numbers presented in Table 1.

However, the integration of SRL into VL-BART has led to a substantial 24% improvement in the CIDEr score, increasing it from 0.66 to 0.82. It's worth noting that the CIDEr metric takes into account not only n-gram overlaps but also the tf-idf scores of each n-gram, which helps to mitigate the impact of uninformative n-grams. Therefore, the improvement in the CIDEr score suggests that RECAP captures valuable information and generates captions that are more relevant and of higher quality, thanks to the incorporation of semantic roles.

To further support our claim regarding the inefficacy of CLIP-S with respect to predicate-argument structure, we measured the SRL recall of RECAP's output, as shown in Table 2. To assess the grounding of a semantic role, we parsed the generated sentence using the SRL parser and collected the corresponding text spans for each role. Similarly, we gathered the bboxes from the input image associated with the same semantic role. Using spaCy (Honnibal and Montani, 2017) we created the embeddings of text spans and object labels. If the similarity score between the text span and the object label of the bbox exceeded a threshold of 0.5, we considered it as a correct grounding. We used spaCy to compute the similarity score. Inclusion of CLIP-S on top of the SRL-incorporated model did not significantly improve the SRL grounding.

One noteworthy observation is that SCST training of the model improves its performance in terms of the Smatch score, albeit with a slight decrease in the CIDEr score. It is important to note that even though a sentence may have a higher n-gram overlap, it can still fail to capture the desired predicate-

argument structure. This is illustrated in Figure 6 with an example. In the image, the captions labeled as a, b, and c represent the GT caption, the caption generated with SRL without AMR SCST and the caption generated with AMR SCST, respectively. It is worth noting that captions a and b are very similar in terms of word choices and predicate-argument structure. However, in caption c, RECAP over-generates the frame of wearing (indicated by the phrase 'in a blue wetsuit') and the description of the dog. It is important to note that, while these descriptions are not incorrect and do not change the main predicate, they do reduce the n-gram overlap with the GT.

Linguistic Analysis. RECAP generates diverse captions based on the semantic roles of objects, providing control over the caption generation process. This SRL-conditioned generation produces different captions based on various linguistic features. Thus, RECAP offers improved interpretability compared to diversifying generation solely through adjustments to model parameters such as the number of beams, beam size, and sampling temperature. In the following sections, we will explore RECAP's ability in terms of verb valency, different predicate selection, diverse agent selection, event structures, and more.

The valency refers to number of argument and their corresponding types that a verb can take. In the case of RECAP, the generation of predicates is influenced by the valency indicated by the SRL annotation of the input image, as previously illustrated in Figure 1. Depending on the variety of arguments, RECAP selects the appropriate predicate to accommodate the required valency.

RECAP selects different predicates based on the presence of different participants. In Figure 7, while (a) and (c) feature the verb playing, the main predicate for (d) and (e) is wearing. Moreover, despite both captions featuring the predicate wearing it should be noted that the participants in (d) and (e) of Figure 7 are distinct, corresponding to the different bboxes.

With the use of SRL, we gain control over the information that is included in the generated captions. For example, in Figure 7, the annotation for (a) included the phrase two girls. In (b), the SRL annotation provided additional information about the ball, and RECAP incorporated this information into the generated caption by modifying the main verb accordingly. In the following image, (c), the attire of the individuals became part of ARG0 in the SRL annotation. RECAP incorporates this information by extending the noun phrase to include the gerund as a modifier, resulting in the caption "Two girls wearing uniforms are playing soccer." This demonstrates how SRL enables us to explicitly control and incorporate specific details into the generated captions.

RECAP can understand event structures, such as activities and state (Vendler, 1957), based on the SRLs in the input image. Activity verbs in these images typically represent physical actions like running, jumping, and swimming, while state verbs convey a state or condition like sitting or standing. Even with the same participants, RECAP can generate different verbs based on their semantic roles. For example, when ARG0 indicates agency and volition of the participant towards the predicate, RECAP will generate a caption reflecting an activity. Conversely, if ARG1 represents prototypical patients, RECAP will choose a stative predicate when ARG1 is the subject. In Figure 8, (a) shows all humans marked as ARG0, resulting in a caption with the main verb walking. In contrast, in (b), the same bboxes are annotated with ARG1, leading to a sentence with the stative verb standing.

RECAP can differentiate between different ARGM roles, which capture various semantic roles in Propbank style annotation, such as direction, temporal, location, and more. Depending on the specific type of ARGM, RECAP can paraphrase the caption accordingly. For instance,Figure 9 depicts four annotations of the same image. In subfigure (a), the background is annotated as ARGM-TMP (temporal), while in subfigure (b), the same bbox is annotated as ARGM-DIR (direction). Thus, for (a), RECAP generates ""... as he walks down a snow covered road" (noting that the main predicate here is wearing, as the clothing bbox is ARG1). Conversely, in (b), RECAP produces "... is walking down a snow..." In (c), the image has both ARGM-DIR and ARGM-LOC, resulting in the generated caption "walking along a trail in the snow."

5. Limitations

One major limitation of RECAP is its tendency to omit copula verbs in generated sentences, as seen in the caption of Figure 10(d). The reason for this is that during training, SRL annotated images are paired with proposal-specific sub-sentences, as described in subsection 4.1. To address this issue, we implemented a solution by generating multiple sentences per image and selecting the one with better grammar and object grounding using CLIP-S and grammar scores (Cho et al., 2022a). The selected caption was then passed to the OpenAI ChatGPT3 (Brown et al., 2020) for grammar correction. For example, the corrected sentence for Figure 10(d) would be 'A man **is** wearing...'. Similarly, the sentence in Figure 10(a) would be corrected as 'A man wearing a black shirt **is** cleaning...'. Despite the addition of the verb wearing in the corrected sentence, the predicate-argument structure remains

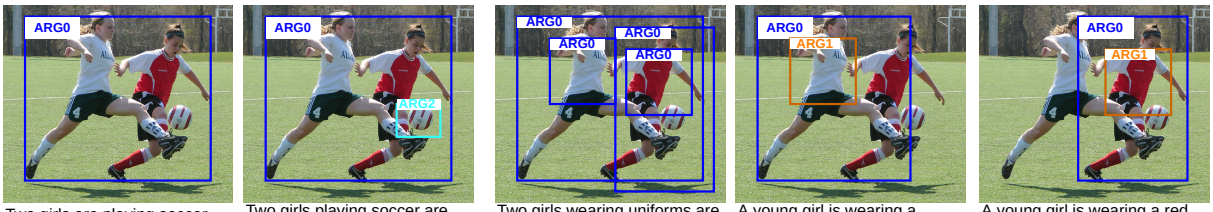


Figure 7: Diversification of captions based on SRL. (a) and (c) depict inclusion of gerund in the noun phrase based on added bbox annotation. (b) differs in the main verb based on annotation of the ball. (c) and (d) demonstrate the alternation of the agent and patient in the generated caption for the same verb. a and e highlight the contrast in the amount of information conveyed in the generated captions.

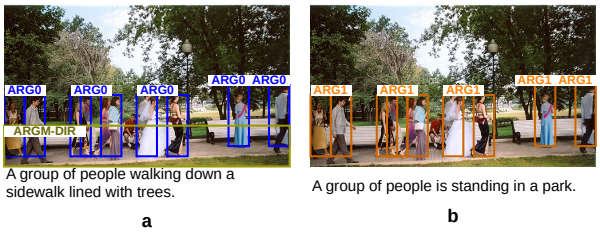


Figure 8: Depending on the SRL, the generated caption can produce either an activity or a stative verb.

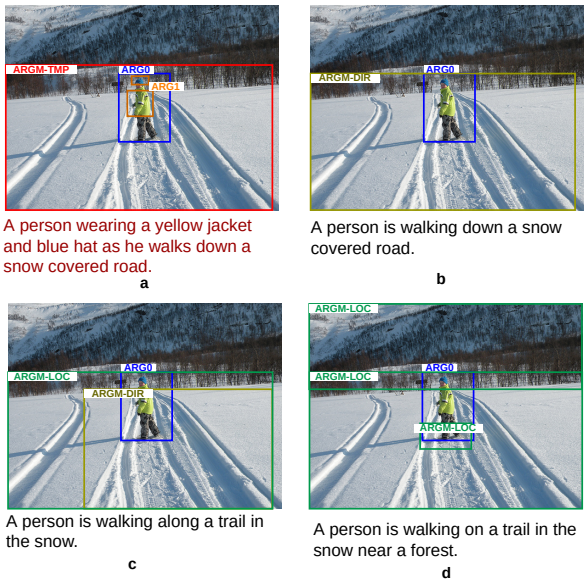


Figure 9: RECAP has the ability to distinguish among different ARGM roles. RECAP can effectively paraphrase the caption depending on the specific type of ARGM.

unchanged.

However, it is important to note that the correction made by ChatGPT sometimes change the desired verb order thereby changing semantic roles for the sentence generated by RECAP. For instance, the main verb in the generated caption for Figure 9(a)

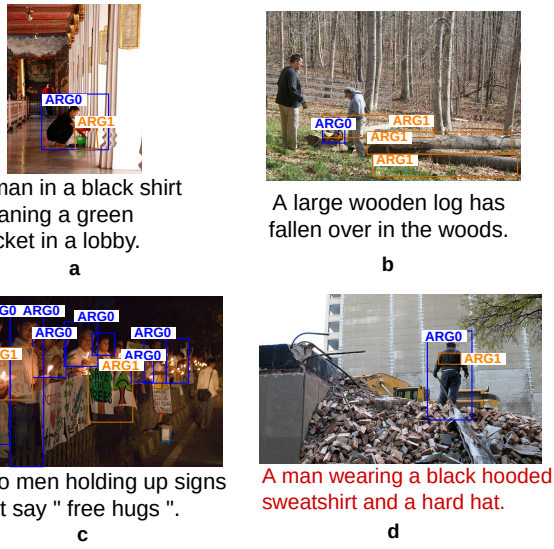


Figure 10: Limitations of RECAP. a) wrong verb b) ARG0 is not included in the sentence c) hallucinate wrong description d)over generation

is wearing, but it lacks a copula verb. Note that as the bbox for snow covered road is annotated as ARGM-TMP, walking is the sub-ordinate event. The ChatGPT correction produced: "A person wearing a yellow jacket and a blue hat is walking down a snow covered road." This corrected sentence has similar predicate argument structure compared to the original sentence generated by RECAP (such as wearing and walking) but walking became the main verb. Moreover the corrected sentence did not incorporate ARGM-TMP.

Despite generating sentences with correct SRLs, RECAP sometimes misidentifies the predicate. In Figure 10(a), the person is using the green bucket instead of cleaning it. Additionally, RECAP sometimes disregards the bbox and its annotation, as seen in Figure 10(b). The underlying language model of BART can occasionally hallucinate and generate excessive descriptions. In Figure 10(c), RECAP hallucinates about "signs that say free hugs." Lastly, in Figure 10(d), the caption includes

the `sweatshirt`, even though the image only has an annotation for the hat.

6. Conclusion

In this paper, we introduce RECAP, a caption generation system that incorporates visual SRL to enhance the process. Existing approaches for caption generation using visual and language models yield accurate descriptions of images but lack user controllability and interpretability. By utilizing semantic roles with the input image, RECAP not only generates diverse captions but also provides linguistically informed interpretability. Additionally, semantic roles allow for directing the generated descriptions to adhere to specific predicate-argument structures. We propose the use of SMATCH as a metric to assess the quality of descriptions in terms of predicate-argument structure. However, obtaining semantic role annotated images still poses a significant challenge, which can be addressed by leveraging advancements in grounded situation recognition (Pratt et al., 2020; Cho et al., 2022b; Cheng et al., 2022; Bhattacharyya et al., 2023). We are also interested in incorporating semantic role information into existing language and vision models (Li et al., 2023; Dai et al., 2023).

7. Acknowledgements

The authors thank the support of DARPA KAIROS Program No. FA8750-19-2-1004. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the U.S. government. We are grateful to James Pustejovsky for suggesting the name of the project.

8. Bibliographical References

Harsh Agrawal*, Karan Desai*, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *International Conference on Computer Vision (ICCV)*.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. 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 397–407, Beijing, China. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages III–1247–III–1255.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of AAAI*.

Abhidip Bhattacharyya, Martha Palmer, and Christoffer Heckman. 2023. CRAPES:cross-modal annotation projection for visual semantic role labeling. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 61–70, Toronto, Canada. Association for Computational Linguistics.

Rexhina Blloshmi, Simone Conia, Rocco Tripodi, and Roberto Navigli. 2021. Generating senses and roles: An end-to-end model for dependency- and span-based semantic role labeling. pages 3786–3793.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,

Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

J. Chen, L. Zhang, C. Bai, and K. Kpalma. 2020. Review of recent deep learning based methods for image-text retrieval. In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 167–172.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020a. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *ECCV*.

Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander Hauptmann. 2022. Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3272–3281, New York, NY, USA. Association for Computing Machinery.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021a. Unifying vision-and-language tasks via text generation. In *ICML*.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022a. Fine-grained image captioning with clip reward. In *Findings of NAACL*.

Junhyeong Cho, Youngseok Yoon, and Suha Kwak. 2022b. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junhyeong Cho, Youngseok Yoon, Hyeonjun Lee, and Suha Kwak. 2021b. Grounded situation recognition with transformers. In *British Machine Vision Conference (BMVC)*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning*, pages 160–167, New York, NY, USA. ACM.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

S. Datta, K. Sikka, A. Roy, K. Ahuja, D. Parikh, and A. Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. pages 2601–2610.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Emily L Denton, Soumith Chintala, arthur szlam, and Rob Fergus. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems 28*, pages 1486–1494.

Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C. Zitnick. 2015. Exploring nearest neighbor approaches for image captioning.

P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. pages 97–112, London, UK, UK. Springer-Verlag.

Fartash Faghri, J. David Fleet, Ryan Jamie Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. *British Machine Vision Conference*.

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

Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. 2019. Bilinear attention networks for person retrieval.

Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. 2009. Describing objects by their attributes. pages 1778–1785. IEEE Computer Society.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. pages 15–29, Berlin, Heidelberg. Springer-Verlag.

Hao Fei, Meishan Zhang, Bobo Li, and Donghong Ji. 2021. End-to-end semantic role labeling with neural transition-based model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:12803–12811.

Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2121–2129, Red Hook, NY, USA. Curran Associates Inc.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles.

Daniel Gildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 239–246, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. pages 580–587, Washington, DC, USA. IEEE Computer Society.

Ross B. Girshick. 2015. Fast R-CNN. pages 1440–1448.

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014a. A multi-view embedding space for modeling internet images, tags, and their semantics. *Int. J. Comput. Vision*, 106(2):210–233.

Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014b. Improving image-sentence embeddings using large weakly annotated photo collections. pages 529–545.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, volume 37, pages 1462–1471, Lille, France. PMLR.

Ankush Gupta and Prashanth Mannem. 2012. From image annotation to image description. In *International Conference on Neural Information Processing*, pages 196–204, Berlin, Heidelberg. Springer-Verlag.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. pages 770–778.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning.

Yan Huang and Liang Wang. 2019. Acmm: Aligned cross-modal memory for few-shot image and sentence matching.

Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. pages 7254–7262.

Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning semantic concepts and order for image and sentence matching. pages 6163–6171.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zixia Jia, Zhaohui Yan, Haoyi Wu, and Kewei Tu. 2022. Span-based semantic role labeling with argument pruning and second-order inference. In *AAAI Conference on Artificial Intelligence*.

Wenhao Jiang, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. 2018. Learning to guide decoding for image captioning. *CoRR*, abs/1804.00887.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *International Conference on Neural Information Processing Systems*, pages 1889–1897.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *AAAI Conference on Artificial Intelligence*, pages 2741–2749. AAAI Press.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2014a. Multimodal neural language models. In *International Conference on Machine Learning*, pages II–595–II–603. JMLR.org.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014b. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014a. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *CoRR*, abs/1411.7399.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014b. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, Sofia, Bulgaria. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA. Curran Associates Inc.

G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. pages 1601–1608.

Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–368, Jeju Island, Korea. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. 2021. Semvlp: Vision-language pre-training by aligning semantics at multiple levels. *CoRR*, abs/2103.07829.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019a. Visual semantic reasoning for image-text matching.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530. Association for Computational Linguistics.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *ACM International Conference on Multimedia*, page 3–11.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. *CoRR*, abs/1803.08314.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, volume 29.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal convolutional neural networks for matching image and sentence. pages 2623–2631.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1064–1074.

Elman Mansimov, Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. 2015. Generating images from captions with attention. *CoRR*, abs/1511.02793.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Explain images with multimodal recurrent neural networks. *CoRR*, abs/1410.1090.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Special issue introduction: Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Alexander Mathews, Lexing Xie, and Xuming He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *CoRR*, abs/1708.00107.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696, USA. Omnipress.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical multimodal lstm for dense visual-semantic embedding.

Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Girish Kulkarni, Margaret Mitchell, Kota Yamaguchi, Karl Stratos, Amit Goyal, Jesse Dodge, Alyssa Mensch, Hal Daumé III, Alexander C. Berg, Choi Yejin, and Tamara L. Berg. 2015. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision*, 119:1–14.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. In *International Conference on Neural Information Processing Systems*, pages 1143–1151.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, 36(1):307–340.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1756–1765. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vision*, 123(1):74–93.

Sameer Pradhan, Kadri Hacioglu, Wayne Ward, James H. Martin, and Daniel Jurafsky. 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. Association for Computational Linguistics.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. *ArXiv*, abs/2003.12058.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. Publisher Copyright: © ICLR 2016: San Juan, Puerto Rico. All Rights Reserved.; 4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016a. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. JMLR.org.

Scott E. Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016b. Learning what and where to draw. *CoRR*, abs/1610.02454.

Scott E. Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. 2016c. Learning deep representations of fine-grained visual descriptions. *CoRR*, abs/1605.05395.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 91–99. Curran Associates, Inc.

Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. 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. Association for Computational Linguistics.

Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. 2016. Joint image-text representation by gaussian visual-semantic embedding. In *ACM International Conference on Multimedia*, page 207–211.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *International Conference on Neural Information Processing Systems*, pages 2234–2242.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

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

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. *CoRR*, abs/1804.08199.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019a. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019b. Lxmert: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing*.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. pages 3713–3722.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2018. Learning to compose dynamic tree structures for visual contexts.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. CIDEr: Consensus-based image description evaluation. *CoRR*, abs/1411.5726.

Zeno Vendler. 1957. Verbs and times. *The Philosophical Review*, 66(2):143–160.

Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2016. Captioning images with diverse objects. *CoRR*, abs/1606.07770.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. pages 5005–5013.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:394–407.

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval.

Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2019. Position focused attention network for image-text matching. In *International Joint Conference on Artificial Intelligence*, pages 3792–3798.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Yiling Wu, Shuhui Wang, and Qingming Huang. 2018. Learning semantic structure-preserved embeddings for cross-modal retrieval. In *ACM International Conference on Multimedia*, page 825–833.

Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning fragment self-attention embeddings for image-text matching. In *ACM International Conference on Multimedia*, page 2088–2096.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057. JMLR.org.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. pages 1316–1324.

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

Yezhou Yang, Ching Lik Teo, Hal Daumé, III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Stroudsburg, PA, USA. Association for Computational Linguistics.

Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. 2016. Stacked attention networks for image question answering. pages 21–29.

Zhilin Yang, Ye Yuan, Yuexin Wu, Ruslan Salakhutdinov, and William W. Cohen. 2016. Encode, review, and decode: Reviewer module for caption generation. *CoRR*, abs/1605.07912.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. *CoRR*, abs/1809.07041.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. pages 4651–4659.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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.

D. Yu, J. Fu, T. Mei, and Y. Rui. 2017. Multi-level attention networks for visual question answering. pages 4187–4195.

Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2023. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris N. Metaxas. 2016. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1612.03242.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 2017. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1710.10916.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CoRR*, abs/2101.00529.

Jie Zhou and Wei Xu. 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. Association for Computational Linguistics.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*.

Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020. More grounded image captioning by distilling image-text matching model.

9. Language Resource References

Charles Fillmore, Christopher Johnson, and Miriam Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal proposition bank 2.0. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France. European Language Resources Association.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and Nathan Schneider. 2020. Abstract meaning representation amr annotation release 3.0 ldc2017t10.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. 31(1):71–106.