Improving language-supervised object detection with linguistic structure analysis

Arushi Rai
University of Pittsburgh
Pittsburgh, PA, USA
arr159@pitt.edu

Adriana Kovashka
University of Pittsburgh
Pittsburgh, PA, USA
kovashka@cs.pitt.edu

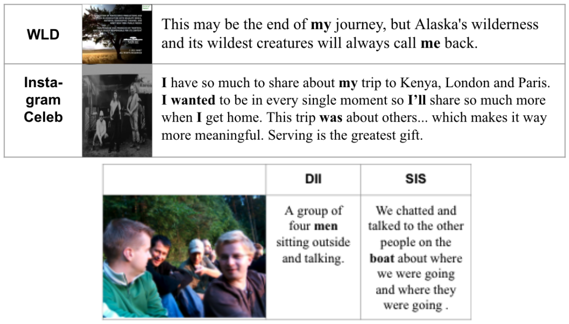
Abstract

Language-supervised object detection typically uses descriptive captions from human-annotated datasets. However, in-the-wild captions take on wider styles of language. We analyze one particular ubiquitous form of language: narrative. We study the differences in linguistic structure and visual-text alignment in narrative and descriptive captions and find we can classify descriptive and narrative style captions using linguistic features such as part of speech, rhetoric structure theory, and multimodal discourse. Then, we use this to select captions from which to extract image-level labels as supervision for weakly supervised object detection. We also improve the quality of extracted labels by filtering based on proximity to verb types for both descriptive and narrative captions.

1. Introduction

Recent progress in self-supervised and weakly-supervised computer vision has been fueled by data freely available on the web [18, 21, 26]. For semantic tasks (e.g. object detection, visual question answering), weak semantic information has been extracted from multimodal data, where visual content co-occurs with some metadata. Such metadata could come from user-generated content, constructed by users to serve their own purposes, rather than constructed for pay. When people create content, whether it is a caption for a photo posted on Instagram or a voice-over accompanying a documentary, they naturally gravitate towards narratives. Yet many multimodal semantic tasks have been supported *not* by such narrative data, but by crowdsourced, paid-for datasets datasets such as COCO [19] and Flickr30K [42]. The captions in these datasets often take on only descriptive styles [17] and also do not represent the full diversity in which humans write coherent captions to images [3].

In this paper, we analyze the differences between narrative and descriptive language accompanying images, and the impact of using these different styles of data for weakly-supervised object detection. Specifically, we analyze differences in linguistic structure and visual-text alignment of narrative and descriptive captions contained in the Visual Storytelling dataset (VIST) [17]. Our goal is to understand the challenges of using narrative or descriptive captions as a source of weak labels for object detection.



This may be the end of **my** journey, but Alaska's wilderness and its wildest creatures will always call **me** back.

I have so much to share about **my** trip to Kenya, London and Paris. **I wanted** to be in every single moment so **I'll** share so much more when **I** get home. This trip **was** about others... which makes it way more meaningful. Serving is the greatest gift.

	DII	SIS
	A group of four **men** sitting outside and talking.	We chatted and talked to the other people on the **boat** about where we were going and where they were going .

Figure 1. In-the-wild multimedia content like documentaries and visual content uploaded to content sharing platforms are accompanied by text metadata such as subtitles or captions. The top two examples are narrative captions that are written from a first person point-of-view and may refer to events in the past. The bottom example from VIST [17] provides two types of captions, traditional descriptive captions (DII) and narrative-like captions (SIS). We may extract false positive labels due to narrative artifacts, e.g. "boat" in the SIS caption.

First, we analyze how narrative and descriptive samples differ in terms of their linguistic and argumentative structure, specifically part-of-speech composition and composition of rhetoric structure theory tags. We also analyze the extent and type of visual-text alignment between an image and its corresponding narrative or descriptive caption. We conduct this analysis through the use of multimodal coherence relations, joint embedding spaces, and pretrained object detectors. We refer to this analysis as a global view, as it focuses on analysing sentences as a whole. We are interested in the extent to which supervision extracted from them by lexical matching [41] would be useful for weakly supervised object detection.

Second, we investigate whether the labels extracted with lexical matching from these captions are true (TP) or false positives (FP). Here, we conduct a local analysis: we explore why a particular type of image-text misalignment occurs (i.e. we sub-categorize the FPs) and study the correlations between TP/FP and characteristics of the words immediately around an object noun in the caption.

Third, we use the findings from our analyses to automatically select which captions should be used to extract labels. We first show the contribution of different features for inferring whether a caption has a descriptive or narrative style. We then select captions according to the resulting classifiers, demonstrating that intelligent selection results in a tangible difference in object detection results. Importantly, we also show the impact of the ability to choose individual segments of sentences (regardless of sentence style) that contain pure labels.

To summarize, our contributions include: (1) developing a model that distinguishes between descriptive and narrative captions using linguistic features; (2) evaluating the effect of extracting image-level labels from descriptive or narrative captions for weakly supervised object detection; (3) analyzing local linguistic features around false positive labels extracted from captions; and (4) evaluating the effect on detection of extracting labels from windows centered around the different linguistic features.

2. Related Work

Vision-language datasets. Multimodal datasets commonly include vision and language data, but differ in the form each modality is presented; for example, the vision modality can be presented via static images [21, 26, 29], images in a sequence [17, 40], or long [7, 24] or short videos [32]. Similarly, language in multimodal datasets can also be represented in diverse forms such as procedural (instructional) language [13,24,40], narrative-like language [5,7,16,17,33], user-written alternative text [19,26,30,42]. Typically, instructional videos have been used for learning visual-textual grounding [23, 24]; Hessel *et al*. [14] studied the extension to using *diverse non-instructional videos* for visual-textual grounding. Instead, our work studies the effect of *extracting labels* from descriptive or narrative captions for weakly supervised object detection.

Weakly-supervised object detection. The vast ocean of multimedia on the Internet motivates many self-supervised methods for image or video representation and weakly-supervised techniques. One particular task, weakly-supervised object detection (WSOD), has been formulated as a multiple instance learning problem to train a model to localize and classify objects from image-level labels, rather than bounding box annotations [4, 10, 28, 31, 31, 35]. The first work to leverage unstructured text accompanying an image for WSOD predicted pseudo image-level labels from captions [41]. Another approach used a vision-language dataset during pretraining, and bounding boxes for only some categories, to generalize detection abilities to novel classes [43]. These language-supervised object detection models have only been tried on multimodal datasets like COCO or Flicker30K (descriptive, crowdsourced) and to a limited extent Conceptual Captions (descriptive alt-text). An exception is [7] which uses wildlife documentaries but relies on track information, which is not available in image datasets. Recent work [12,44] extends CLIP [26] for detection. We instead aim to understand the impact of in-the-wild captions for language-supervised object detection.

Coherence analysis in language and across modalities. Discourse structure relays syntactic information about how text is organized. In language, each span is connected in a meaningful and coherent manner to the next; *how* it is connected is known as the discourse or coherence relation. Discourse relations and structure are a well studied topic [22, 38, 39]. We use parsers that follow the taxonomy of discourse relations from rhetorical structure theory [22]; each span serves a function in rhetoric. This concept was extended to multimodal discourse, specifically images and captions in instructional [1] and caption generation contexts [2], but has not been used for object detection before.

3. Global View: Descriptive vs Narrative

As shown in Fig. 1, narrative or in-the-wild captions can have lower visual-text alignment due to object mentions that reference events or external narrative context not depicted in the image; this can negatively affect language supervised object detection. In this section, quantitatively analyze the differences in linguistic structure and visual-text alignment between descriptive and narrative captions.

3.1. Preprocessing

We do our analysis on the Visual Storytelling (VIST) dataset [17] with three different styles of captions: **D**escriptions of **I**mages in **I**solation (DII), **D**escriptions of **I**mages in **S**equence (DIS), **S**tory for **I**mages in **S**equence (SIS). Each image multiple captions labeled as DII and SIS, which allows us to observe the impact of descriptive or narrative captions (like in-the-wild captions) while keeping the image constant.

There may be variables beyond the characteristics of captions that could affect the analysis and application to weakly supervised object detection, e.g. class imbalance or visual differences in object appearances if the set of images are not constant between DII and SIS. To solve the latter, we sample images rather than captions. We filter about 20K images that have no corresponding DII captions, and denote the remaining subset as I, containing about 40K images. Then, we limit our label space to overlap between the most frequently occurring lexicon in VIST and PAS-

CAL VOC 2007 label set [9]; resulting in four categories: car, boat, dog, person. For each image $x_i \in I$, there are two sets of captions, C_i^{SIS} and C_i^{DII}. To extract image-level labels, we apply regular expression matching on each caption $c_i^j, \forall j \in C_i^{SIS} \cup C_i^{DII}$ with the following pattern: “(car||boat||dog||person)”, and extract labels, el_i^j from each caption c_i^j. We produce a final set of labels, el_i for each image x_i by taking the union $el_i^1 \cup el_i^2 \cup ... \cup el_i^J$ where $J = \|C_i^{SIS} \cup C_i^{DII}\|$ and retain any image x_i where el_i contains at least one category from our label set. This subset, I^* has an advantage; we retain captions without any extracted labels (preserving possible false negatives), which is useful for our analysis.

3.2. Linguistic Differences

The linguistic structure of a caption may correlate with visual-textual alignment. A caption with a low count of nouns might not mention the salient objects in objects in the image. Some prepositions such as “next to” may convey spatial information that is likely to be visible. Other prepositions such as “like” may indicate the use of non-literal language (figure of speech) and any object that follows will likely *not* to be visible. Linguistic structure beyond part of speech can also play a role in predicting whether the caption is actually aligned with the image. For example, an enablement discourse relation, where an action described in the nucleus (core) span is enabled by information provided in the satellite (helper), could imply additional contextual information is being provided; for example, “We used the gondolas one night [nucleus] to go to a ball [satellite].” Here, it is highly unlikely that *both* gondola and the ball venue would be shown.

We hypothesize the same linguistic features that differentiate between descriptive and narrative captions could contribute to visual and text alignment. We now describe the setup and the analysis for each linguistic feature.

Part of Speech (POS). We extract POS tags over all VIST captions using SpaCy [15] and observe differences in pronoun usage and verb tense between descriptive and narrative captions (Table 1a). The increased use of pronouns in SIS suggests a deviation from an impersonal, objective tone. For the same image, one SIS caption says “**we** finally arrive at the island” with a first person point-of-view compared to a DII caption which provides both count and details regarding the subject: “**a group of four men** sitting together”.

Verb tense and aspect also differ between descriptive/narrative captions (Table 1c). Both verb tenses are used in SIS captions, with more verbs referring to the past than present. Past and present tense in SIS can occur in the same sentence: “Afterwards, we take a couple photographs because we paid the photographer to do so.” An aligned image would show either show the couple posing for a photo or the transaction. This interaction between past and present occurs over 33% of SIS captions vs 16% in DII.

(a) Frequency of POS tags over DII/SIS.

Part of Speech	DII	SIS
noun	**3.85**	2.73
preposition	**1.75**	1.01
adjective	**0.98**	0.85
personal pronoun	0.18	**0.80**
verb	1.55	**2.00**
verb base	0.07	**0.37**
verb gerund	**0.61**	0.23
past tense	0.08	**0.89**
past participle	**0.21**	0.17
non-3rd person sing pres	**0.22**	0.16
3rd person sing pres	**0.36**	0.19

(b) Distribution of RST tags (top 8).

RST Tag	DII (%)	SIS (%)
Attribution	1.3	**4.8**
Background	1.3	**2.2**
Contrast	1.1	**1.7**
Enablement	0.6	**2.3**
Joint	4.1	**6.0**
Temporal	**2.2**	1.2
Elaboration	**21.4**	10.9
None	**65.1**	64.1

(c) Distribution over tense and aspect.

	past	present	progressive	perfective
DII	18.5%	**81.4%**	**94.5%**	5.4%
SIS	**52.8%**	47.2%	**75.1%**	24.9%

Table 1. We observe the difference in part of speech, verb tense and aspect composition, and rhetoric structure between descriptive and narrative captions.

Aspect is more evenly distributed in narrative than descriptive captions, but both still favor progressive aspect. Progressive aspect refers to verbs that describe an ongoing activity such as “The event is starting, there are even some dancers forming outside” (SIS) while perfective describes a completed activity, e.g. “The old part has been **removed** and now there many loose wires now” (SIS).

We hypothesize part of speech tags can be a strong feature to discriminate between DII and SIS, and DII provides more aligned captions since it describes ongoing activities more likely to be present in the accompanying visual scene.

RST Relations. Rhetorical structure theory (RST) was proposed as a taxonomy on how text is structured while being coherent [22]. There is an argument or a claim (nucleus) which is furthered by supporting spans (satellite). In this example, nucleus is in blue and satellite in cyan: “Employees are urged to complete new beneficiary designation forms for retirement or life insurance benefits whenever there is a change in marital or family status.” The support relation between satellite and nucleus is defined by a 16 discourse relations such as Elaboration (sat. provides additional in-

Category	P		R	
	DII	**SIS**	**DII**	**SIS**
Car	0.64	**0.71**	**0.38**	0.35
Person	0.94	**0.96**	0.17	**0.20**
Dog	0.76	**0.77**	**0.58**	0.57
Boat	0.50	**0.51**	**0.39**	0.33

Table 2. Using YOLOv3 [27] to estimate which extracted labels are present in the image, we calculate precision and recall of the visual presence of extracted labels over four categories.

formation, e.g. “a fox [nuc.] that is sitting on the grass near the tree [sat.]”, which is descriptive), Attribution (direct instances of reported speech e.g. “She said [sat.], ‘Hi’ [nuc.]” and indirect instances “I think [sat.] the dog is sick [nuc.]”, both of which are narrative), Joint (multinuclear relations “then we sat down [nuc.] and started drinking beer [nuc.] and talked for quite a while [nuc.]”, part of a narrative), etc. We use Wang *et al*. [37] to segment captions into spans, and the StageDP discourse parser [36] to predict the RST tags between spans.

In Table 1b we identify key discourse relations that have a higher occurrence in DII or SIS captions. The biggest difference occurs for the “Elaboration” (10.5% more in DII than SIS), “Attribution” (3.5% more in SIS), “Joint” (1.9% more in SIS) and “Enablement” (1.7% more in SIS) relations. While DII contains mainly “Elaboration” and “Temporal” tags, SIS represents a wider range of how spans relate to one another. We also observe differences in the lexicon correlated with each discourse relation and whether the caption is DII or SIS. Both DII and SIS have nearly the same number of captions flagged as temporally related, however “before” and “while” frequently occur in the flagged SIS captions and only “while” frequently occurs in the DII. This is significant because labels extracted from Temporal-SIS captions could contain either a future or past reference in either nucleus or satellite clauses, and therefore, not currently visible. This can lead to false positive labels when naively extracting from captions.

3.3. Visual-Text Alignment

Next, we analyze quantitative measures of visual-text alignment, ranging from the alignment of extracted labels and visual objects to multimodal discourse relations. Insights regarding visual-text alignment can help identify sources of noise and determine the suitability of extracting labels from each source.

Using a pretrained object detection model. Since VIST is intended for image captioning, not object detection, it does not come with ground-truth bounding boxes nor image-level labels. Thus we extract image labels using YOLOv3 [27]. We would expect that both precision and recall of the extracted labels would be high in DII compared to SIS, and that the precision would be higher than recall. Surprisingly, Table 2 shows precision is estimated to be lower for DII, contrary to our expectations. We investigate this in Sec. 4.

Using CLIP. We also observe the alignment between images and captions on a semantic embedding level. We use CLIP [26] to extract image and caption joint embeddings, and use cosine similarity to measure their similarity. DII caption-image embeddings have statistically significant ($p < 0.0001$) higher (0.30 ± 0.03) cosine similarity than SIS (0.26 ± 0.04). This aligns with our expectation that descriptive captions are more likely to be directly aligned with its visual content, and contradicts the previous observation of lower precision of extracted labels from DII captions.

CLUE image-text discourse relations. While descriptive captions commonly appear to have a redundant relationship with their corresponding visual information, other forms of language can convey additional or complementary information to a visual aid. Our last linguistic analysis breaks down image-text alignment into discrete categories through multimodal discourse relations. We use the image-text discourse relations defined by Alikhani *et al*. [2] such as: Visible (elements in text directly related to the image, akin to descriptive captions), Subjective (text provides unverifiable information “**the most beautiful** horse”), Story (text provides some narrative context to the image with risk of “context hallucinations” and extracted labels not present in the corresponding image), Action, Meta, and Irrelevant (no visual-textual grounding). Using a multimodal late-fusion model trained to predict CLUE discourse labels, [2] we observe that for about 50% of the SIS captions, there are no CLUE discourse labels predicted, compared to only about 14% of DII captions. Ignoring captions with no predictions, DII captions have slightly more “Visible, Action” (0.5%) and “Story” (0.3%) tags. There may be image-text relations unexplored in narrative-like captions due to 50% of captions having no predicted label; this is plausible given that CLUE was annotated on a subset of Conceptual Captions [30], a multimodal dataset with descriptive (alt-text) captions.

4. Local View: Exploring False Positives Labels Extracted from Descriptive/Narrative Captions

As noted in Sec. 3.3, we expected high precision of labels extracted from DII, since annotators were tasked with describing *exactly* the image contents and the selected categories (dog, car, etc) are highly visible categories. To determine if the reported low precision was a result of an imperfect classifier or if the labels were indeed visually absent, we gathered a set of extracted labels from 100 captions which were flagged as containing a false positive (i.e. the extracted label did not appear in the YOLOv3 predictions for that image). These extracted labels were annotated as either a *vi-*

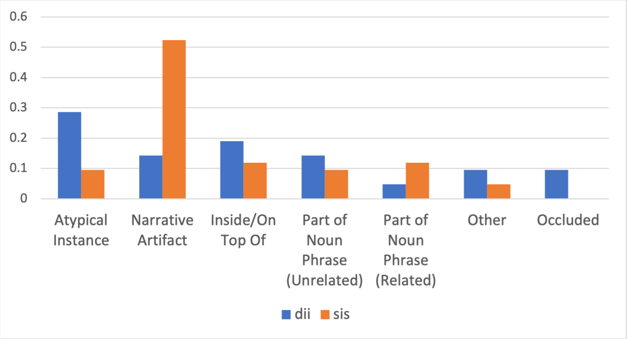


Figure 2. The distribution of the type of VAEL. VAEL from SIS are mainly comprised of narrative artifacts (53%) compared to DII (14%).

sually present extracted label (VPEL) or a visually absent extracted label (VAEL). For those annotated as VAEL, we also annotated the type of VAEL using the scheme in Table 3. The agreement was high, $\kappa = 0.697$ over the annotated set and two annotators (authors). For Fig. 2, we used the subset for which there was agreement between the annotators. We conclude that *the amount of VAELs is actually lower in DII (50%) compared to SIS (60%)*, therefore the lower precision of extracted labels from DII is due to missed predictions from YOLOv3, not because of the quality of extracted labels from DII.

We found there may be specific linguistic structural cues indicative of the presence of a VAEL and its type. For example, lexical cues (italicized) associated with atypicality may indicate a Atypical Instance VAEL (bolded): "She looked very happy with some of the gifts she got. Her favorite was a **dog** *toy*." In cases like narrative artifact, there may be an interaction between multiple spans and the use of past tense that may indicate a VAEL: "*Earlier that day*, my **dog** peed on a flower." Other examples of narrative artifact in SIS include the use of present tense verbs and prepositions: "What better feeling that *walking* on grass *after* a long **car** ride", "As [female] was *walking* her **dog**, she met a stranger." Recognizing when and where visually absent contextual information appears can still allow for label extraction by extracting outside of the contextual sequence. *This can increase the quality of labels extracted from noisier, narrative captions* instead of using only descriptive captions as a proxy for clean labels. We investigate linguistic cues (verb type and tense, nucleus/satellite, object/subject) and their effect on extracted labels for WSOD in Sec. 6.

5. Methods

The analysis in Secs. 3 and 4 indicates descriptive captions may contain labels that are more aligned with the corresponding image compared to narrative captions. This motivates learning to predict whether a caption is of descriptive style or narrative style. We train a DII/SIS classifier using captions and their split metadata (DII or SIS) from Visual Storytelling Dataset (VIST) [17]. The goal of this DII/SIS classifier (Sec. 5.1) is to attribute descriptiveness scores to all captions corresponding to an image. Then, the scores will be used to select one entire caption (global view) from which labels will be extracted. These labels will be used to train a WSOD model (Sec. 5.2).

Following the local view analysis in Sec. 4, we also create a rule-based module to observe the impact of extracting labels based on proximity to some linguistic feature: verb type, nucleus/satellite, and subject/object. This module limits extracting labels from windows centered around a linguistic feature in all captions for the image as opposed selecting a single caption.

5.1. Extracting Labels

DII/SIS Classifier In our prior analysis, we observed linguistic features like part of speech and discourse relations like RST [22] and CLUE [3] are distinctive between descriptive and narrative captions. We train a logistic regression classifier to predict the descriptiveness of the caption.
Rule Based Classifier The global view method, described above, filters out entire captions. The local view method, on the other hand, chooses to extract labels based a lexical match between the selected categories and a five-word window of a particular linguistic feature in a caption. We illustrate this in Fig. 3. For nucleus/satellite local view classifiers, lexical match is directly applied to the span associated with nucleus or satellite, rather than a 5-word window.

5.2. Weakly Supervised Object Detection

We follow prior literature [4, 41] to train a network to predict the labels identified in Sec. 5.1. A convolutional neural network base encoder $h(x_i)$ is used to extract a feature map for an image, x_i. We use VGG-16 [20] with the fully connected layers removed as our base encoder, initialized with pre-trained ImageNet [8] weights. Region of interest (ROI) pooling [11] is then applied to the feature map and the regions of interest $R_i \in \mathbb{R}^{4\times M}$ to generate a feature embedding $\phi(x_i)_m$ for each region: $\phi(x_i) = ROIPool(h(x_i), R_i)$. We initialize two parallel fully-connected layers f_c and f_d whose outputs will be normalized to give a classification score, the probability that an object c is present in that region, and detection score, the probability that $R_{i,m}$ contributes to the image-level class prediction, respectively.

$$p^{cls}_{m,c} = \sigma(f_c(\phi(x_i)_m)),\ \ p^{det}_{m,c} = \frac{\exp\left(f_d(\phi(x_i)_m)\right)}{\sum_{j=1}^{M} \exp\left(f_d(\phi(x_i)_j)\right)} \tag{1}$$

The class and detection predictions for each region are multiplied and summed over M proposals to produce one

Type of VAEL	Definition
Atypical Instance	Object is present in an atypical form (e.g. clay model, toy)
Inside/On Top Of	Image is taken inside or on top of the object
Occluded	Full view of object is limited because object is occluded
Part of Phrase (Rel)	Extracted label is part of a related phrase ('car' in 'car show')
Part of Phrase (Unrel)	Extracted label is part of unrelated phrase ('dog' in 'hot dog')
Missing From Scene (Narrative Artifact)	Object completely missing from the scene but was mentioned in the one of the captions to further the story told in the caption, e.g. "She was returning from the **car** when she pets the dog"
Missing (Other)	Missing from scene for none of the reasons described above

Table 3. Types of visually absent extracted labels.

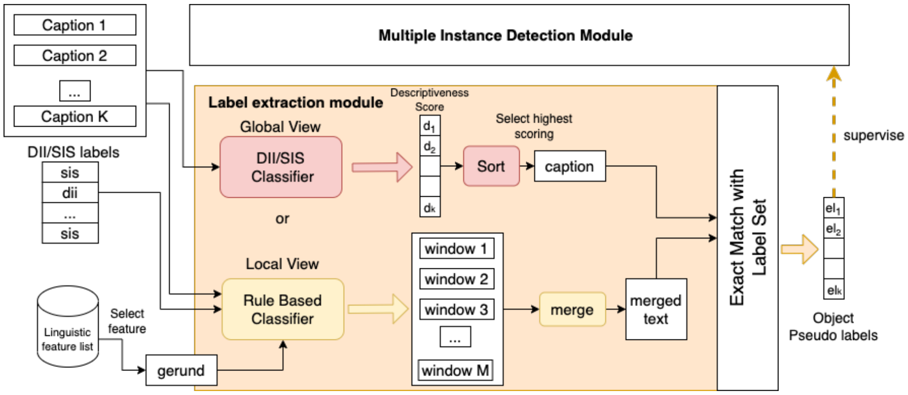


Figure 3. Global and local view considerations while extracting labels. The local view classifier takes **a** feature among: gerund, verb base form, past tense, past participle, non-3rd person singular present, 3rd person singular present, nucleus, satellite, subject, object.

image-level class prediction vector.

$$\hat{p_c} = \sigma\left(\sum_{m=1}^{M} p_{m,c}^{det} o_{m,c}^{cls}\right) \tag{2}$$

To train, we apply the multiple instance detection loss to the extracted image-level label and the predicted image-level label [4]:

$$L_{mid} = \frac{1}{C}\sum_{c=1}^{C}\left[y_c \log \hat{p_c} + (1-y_c)\log(1-\hat{p_c})\right] \tag{3}$$

6. Experiments

We experiment with different ways of selecting our training data. We train the object detection model using subsets of VIST, and evaluate on PASCAL.

Implementation Details We extract labels for each caption individually and decide to sample or merge depending on whether the global or local view methods are used. Our processed VIST subset, I^*, retained images with at least one mention in at least one of its corresponding captions of an object from the object category set defined in Sec. 3.1. We merge over extracted labels from the DII and SIS captions for each image and use these estimated image-level labels to sample at least 500 instances per category from I^*. We use this semi-balanced set as our training set. We extract 2000 bounding box proposals per image using Selective Search [34] from OpenCV [6]. We use a PyTorch [25] implementation of WSDDN [4] for our weakly supervised model. The longest side in the image is randomly resized to 480, 576, 688, 864, or 1200. Horizontal flip is randomly applied. During evaluation, the longest side is scaled to 576. We use the feature map produced by the final convolutional layer of VGGNet [20] and apply ROI pooling [11] to extract proposal feature vectors. After proposal scores are computed, we apply non maximum suppression with an IoU threshold of 0.4. We sample 500 bounding box proposals at train time. We use the SGD optimizer with a momentum of 0.9, weight decay of 0.0005, learning rate of 1e-4, and multi-step learning rate schedule where the LR is decayed with gamma 0.1 at two milestones (2^{nd} and 6^{th} epochs). We train the model for 40,000 iterations or 20 epochs with a batch size of 1 on 1 GPU (GTX 1080 Ti). The best model is selected using a validation set (200 samples) from the PASCAL VOC2007 evaluation set.

6.1. Linguistic features for DII/SIS classifier and WSOD training

How distinguishable are descriptive and narrative captions? In our analysis in Sec. 3.2, we see clear lin-

Linguistic features	PREC	REC
POS	**0.8823**	**0.8782**
RST	0.6281	0.5170
CLUE	0.6813	0.6411
CLIP [26]	0.7215	0.7170
POS+RST	0.8838	0.8787
POS+CLUE	0.8911	0.8879
POS+RST+CLUE	0.8916	0.8884
CLIP+POS	0.8925	0.8893
CLIP+RST	0.7360	0.7312
CLIP+CLUE	0.7523	0.7480
CLIP+POS+RST+CLUE	**0.8982**	**0.8960**
BERT	**0.9570**	**0.9560**

Table 4. We evaluate precision/recall of DII/SIS classifiers on a VIST holdout.

Descriptive Classifier	DII	Random	SIS
GT Labels	**0.0195**	0.0105	0.0050
POS	**0.0304**	0.0105	0.0047
POS+RST	**0.0187**	0.0105	0.0046
POS+RST+CLUE	**0.0187**	0.0105	0.0038
CLIP	**0.0173**	0.0105	0.0043

Table 5. Effect of descriptive classifiers for filtering on WSOD. We evaluate mAP performance of WSOD on VOC-07 [9]. The random classifier (trained once) is independent of descriptiveness classifiers. Bold indicates best performance in row.

guistic structural differences in the captions of each source. In addition, we observe descriptive captions have a higher alignment with their corresponding image. First, we train a model to predict whether a caption is DII or SIS and then evaluate on a holdout set and use the performance as an indicator of the predictive power of combinations of these features. In Table 4, we see that the linguistic features used in our analysis earlier are quite competitive given their representation size (ranging from 33-55 dimensions). POS is the best single-feature-type performer (top of table), followed by the cosine similarity between the visual and text CLIP embeddings. Combinations of linguistic features only achieve a slight boost over POS alone. Finally, using BERT embeddings (512-dimensional) achieves high precision and recall; this indicates significant language differences between DII and SIS, but this memory-intensive classifier is impractical to use on-the-fly unlike our simpler classifiers.

Now, we ask **how good is an object detection model trained using labels extracted from global-view selected captions**? We choose some of the promising DII/SIS classifiers (from Table 4). Each assigns descriptiveness scores; the highest-scoring caption is selected per image if descriptive captions are desired, or the lowest-scoring caption if we want narrative captions. When using the (ground truth) caption source directly, we randomly select one caption depending on the type of caption wanted. We extract labels from the selected captions as pseudo image-level labels to train the multiple instance detection network and measure mAP over the selected categories (car, boat, dog, person).

	DII		**SIS**	
Linguistic Features	VACR	VPCR	VACR	VPCR
past tense	0.000	0.000	0.167	**0.172**
gerund	0.095	**0.304**	0.024	**0.034**
non-3rd person sing pres	**0.048**	0.000	**0.024**	0.000
3rd person sing pres	0.143	**0.196**	0.000	**0.052**
verb base	**0.095**	0.000	0.024	**0.034**
past participle	0.000	**0.109**	**0.024**	0.000
subject	0.143	**0.261**	0.214	**0.293**
object	**0.667**	0.457	0.667	**0.707**
nucleus	**0.238**	0.217	**0.238**	0.207
satellite	**0.143**	0.043	**0.238**	0.069

Table 6. We evaluate different rule-based classifiers on our small annotated set.

Extracting image-level labels from DII-classified captions always performs better than SIS-classified captions, and SIS-classified captions perform worse than a randomly selected caption (any type). Using a POS-based classifier to select DII-like captions yields a better set of labels compared to using any other classifier for the same. Note the POS-based classifier recorded lower precision than the POS+RST+CLUE one on predicting DII captions in Table 4, meaning that several SIS captions get classified as DII, yet in Table 5, the POS-chosen captions perform best. A possible explanation is that the POS-based classifier depends on characteristics of descriptive captions that are directly tied to alignment. More importantly, the high-scoring captions chosen by the POS classifier perform significantly better than the ground-truth DII captions, indicating that *the source of a caption alone is not a sufficient indicator of extracted label quality, rather the characteristics of the caption are important*. Note the low results overall are because of the very small training sets (in turn due to the size of the VIST dataset), but this is not a limitation of our approach.

6.2. Rule-based local classifier to filter out false positives for WSOD

Our second set of experiments extend the local view analysis in Sec. 4, where we observed through qualitative examples that certain verbs and lexical or other linguistic features may be indicative of visually absent extracted labels (VAEL): an object mentioned in a caption, but missing in the corresponding image. Specifically, we test if extracting labels based on *proximity* to verbs or other linguistic features can be predictive of label quality. We first evaluate the ability of local view classifiers to capture VAEL and then we will evaluate the impact on WSOD of using local view classifiers to extract labels. We use the VAEL/VPEL annotated set of extracted labels described in Sec. 4. In contrast to the previous subsection, we want to assess if spe-

Verb	$D^* \cup S^*$ (Count)	D^* (Count)	S^* (Count)
baseline	0.0047 (583)	0.0026 (456)	0.0024 (145)
verb base	0.0014 (128)	0.0010 (42)	0.0014 (86)
past tense	0.0031 (461)	0.0015 (83)	**0.0033** (403)
gerund	**0.0053** (976)	**0.0046** (909)	0.0014 (136)
non-3rd person sing pres	0.0013 (94)	0.0012 (77)	0.0011 (18)
3rd person sing pres	**0.0068** (985)	**0.0067** (888)	0.0016 (193)
past participle	0.0028 (332)	**0.0028** (270)	0.0014 (100)

Table 7. mAP on VOC-07 [9]. Bold signifies mAP higher than baseline (top row).

cific linguistic features are likely to extract more VAEL or VPEL and if this differs between descriptive and narrative captions. Each local-view classifier takes in one linguistic feature from the following list: verb type, dependency tag of nouns (subject/object), nucleus/satellite.

We define two metrics to assess which linguistic features are indicative of label quality or lack thereof. VAEL Capture Rate (VACR) measures the rate of VAELs captured by the local classifier over the total number of VAELs in the VAEL/VPEL annotated subset (DII or SIS), and VPEL Capture Rate (VPCR) rate is calculated similarly, except using VPEL counts. In Table 6 we see VAELs from DII and SIS tend to be near non-3^{rd} person present tense verbs, nucleus, satellite, and objects, while in SIS VAELs tend to be found also near past participle verbs. DII VAELs tend to be near base verbs. Similarly, we find that VPELs are more common near gerunds, subjects, and 3^{rd} person present tense verbs for both DII and SIS.

Seeing that the local view classifiers pick cleaner or dirtier labels depending on the feature and style of the caption, we ask if we can **leverage these local view classifiers to improve the quality of extracted labels for WSOD.**

Since verbs had the most significant effect, we limit the next experiment to verb-based local view classifiers. We denote all extracted labels for an image from DII and SIS captions as D^* and S^*, respectively. Our baseline reflects the average effect of using any verb-based local view classifier; we choose a verb tag at random. We include count information to determine if the relative gain in performance is due to simply more samples or clean labels. For DII, gerunds and 3^{rd} person present tense verbs perform well above the baseline, however gerunds perform about 50% worse than 3^{rd} person present tense verbs despite having slightly more samples. Past participle verbs have a slightly higher performance than the baseline, while extracting 40% less labels than the baseline. This means that for DII, 3^{rd} person present tense verbs and past participle verbs correlate with better extracted labels which is consistent with Table 6.

For SIS, only past tense verbs perform above the baseline; this verb tag is frequently present in SIS, so it is not surprising more labels are extracted. However, although SIS past tense verbs have less examples (403) than the baseline for DII (456), SIS past tense verbs perform 0.0007 better than that baseline. When comparing DII+SIS past tense verbs and SIS past tense verbs, we see adding labels extracted from DII past tense verbs actually deteriorates the performance by 0.0020. This could mean past tense verbs in DII could be tied to more VAELs. From Table 6 alone, we could not make this conclusion, since past tense verbs did not appear near any labels in DII captions within the small annotation set. In SIS, past participle and gerund verbs perform only 0.0003 better than non-3^{rd} person present tense verbs despite having 5-7x more labels. We also evaluated subject/object and nucleus/satellite local view classifiers and found there was nearly no difference for subject/object and the difference between nucleus and satellite (nucleus 3x better) can be attributed to the number of extracted labels.

We also show that verb forms can be used to *construct windows to extract labels from descriptive or narrative captions*. This is important as it is a first step to enable extraction of quality labels from narrative captions as well.

7. Conclusion

There are key differences in linguistic structure between descriptive and narrative captions. We leveraged this to classify between these styles, and found that using descriptiveness predictions to select captions for WSOD training performed far better than sampling a caption using ground truth source information, indicating that some mixture between descriptive and narrative captions may be beneficial when extracting labels for WSOD. While it appears more narrative captions are needed to perform on par with descriptive captions, due to the abundance of in-the-wild image and caption pairs it may be possible to get those pairs. We also show we can improve the quality of labels extracted from narrative captions through local view classifiers. We used VIST due to each image having captions with different styles, but hope to extend our analysis to other datasets.

Acknowledgement: This work was supported by a University of Pittsburgh Computer Science fellowship, and by Google/Amazon/Adobe gifts.

References

[1] Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. Cite: A corpus of image–text discourse relations. In *Proceedings of NAACL-HLT*, pages 570–575, 2019. 2

[2] Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, 2020. 2, 4

[3] Malihe Alikhani and Matthew Stone. "Caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, page 58–67. Association for Computational Linguistics, Jun 2019. 1, 5

[4] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 2, 5, 6

[5] Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. Good news, everyone! context driven entity-aware captioning for news images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12467, 2019. 2

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 6

[7] Kai Chen, Hang Song, Chen Change Loy, and Dahua Lin. Discover and learn new objects from documentaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[9] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. 88(2), 2010. 3, 7, 8

[10] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[11] Ross B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 5, 6

[12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 2

[13] Jack Hessel, Lillian Lee, and David Mimno. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. 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 2034–2045, 2019. 2

[14] Jack Hessel, Zhenhai Zhu, Bo Pang, and Radu Soricut. Beyond instructional videos: Probing for more diverse visual-textual grounding on youtube. In *EMNLP*, 2020. 2

[15] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 3

[16] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*, pages 709–727. Springer, 2020. 2

[17] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, 2016. 1, 2, 5

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2

[20] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015. 5, 6

[21] Dhruv Kumar Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 1, 2

[22] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281, 1988. 2, 3, 5

[23] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2020. 2

[24] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, 2019. 2

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. 6

[26] 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. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 4, 7

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 4

[28] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[29] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *Data Centric AI NeurIPS Workshop 2021*, 2021. 2

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2, 4

[31] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[32] Yansong Tang, Jiwen Lu, and Jie Zhou. Comprehensive instructional video analysis: The coin dataset and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3138–3153, 2021. 2

[33] Christopher Thomas and Adriana Kovashka. Predicting the politics of an image using webly supervised data. In *NeurIPS*, 2019. 2

[34] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171, 2013. 6

[35] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[36] Yizhong Wang, Sujian Li, and Houfeng Wang. A two-stage parsing method for text-level discourse analysis. In *ACL*, 2017. 4

[37] Yizhong Wang, Sujian Li, and Jingfeng Yang. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, 2018. 4

[38] Bonnie Lynn Webber and Aravind K. Joshi. Anchoring a Lexicalized Tree-Adjoining Grammar for discourse. In *Discourse Relations and Discourse Markers*, 1998. 2

[39] Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2), 2005. 2

[40] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *EMNLP*, 2018. 2

[41] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, page 9685–9694, Oct 2019. 1, 2, 5

[42] Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. 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, 2014. 1, 2

[43] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 14393–14402, 2021. 2

[44] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2