Fine-Grained Visual Entailment

Christopher Thomas*†, Yipeng Zhang*, and Shih-Fu Chang

Columbia University, New York, NY 10034, USA
{christopher.thomas, zhang.yipeng, sc250}@columbia.edu

Abstract. Visual entailment is a recently proposed multimodal reasoning task where the goal is to predict the logical relationship of a piece of text to an image. In this paper, we propose an extension of this task, where the goal is to predict the logical relationship of fine-grained knowledge elements within a piece of text to an image. Unlike prior work, our method is inherently explainable and makes logical predictions at different levels of granularity. Because we lack fine-grained labels to train our method, we propose a novel multi-instance learning approach which learns a fine-grained labeling using only sample-level supervision. We also impose novel semantic structural constraints which ensure that fine-grained predictions are internally semantically consistent. We evaluate our method on a new dataset of manually annotated knowledge elements and show that our method achieves 68.18% accuracy at this challenging task while significantly outperforming several strong baselines. Finally, we present extensive qualitative results illustrating our method's predictions and the visual evidence our method relied on. Our code and annotated dataset can be found here: https://github.com/SkrighYZ/FGVE.

1 Introduction

Tasks requiring multimodal understanding across vision and language have seen an explosion of interest in recent years, driven largely by their many downstream applications. Common tasks include visual question answering [1, 20, 41], visual commonsense reasoning [55], and visual dialog [14, 15]. Moreover, tasks that had historically been studied by only the natural language processing or computer vision communities have recently received attention from both communities. For example, event extraction [10, 28, 57] and coreferencing [26, 38], longstanding information extraction and NLP tasks, have all recently been explored multimodally.

One such task is the textual entailment task, first proposed in 2005 [13]. The task requires a system to decide whether a piece of text (the hypothesis) can be logically deduced from another piece of text accepted as true (the premise). Xie et al. [50] posed a multimodal variant of the task called visual entailment, which replaces the textual premise with an image. Because of the rich cross-modal reasoning required, it has become a standard benchmark for testing joint vision-and-language understanding in many recent multimodal models [12, 23, 31].

*indicates equal contribution
†corresponding author

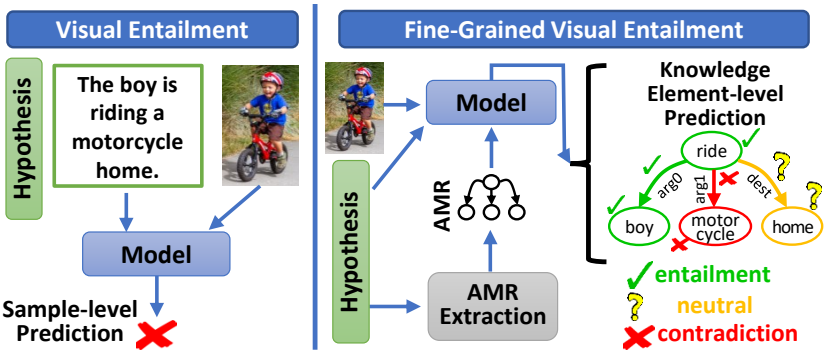

Fig. 1. In the standard visual entailment task, a model predicts the logical relationship of the entire hypothesis to an image. In our proposed task, the model predicts the relationship of each knowledge element of the hypothesis. Specifically, our model predicts a boy is present, someone is riding, the boy is riding, a motorcycle is not present, no one is riding a motorcycle, and can't conclude whether someone is riding home or not.

The standard visual entailment benchmark [50] is a three-way classification task between entailment (hypothesis is true), neutral (hypothesis could be true or false), and contradiction (hypothesis is false). Though the classification is made for the entire hypothesis, the task by its very nature requires fine-grained multimodal reasoning. For example, for a hypothesis to be labeled contradiction, the model must find at least one facet of the hypothesis that conflicts with the image. The task as posed, however does not require the model to produce fine-grained predictions which would explain its reasoning.

This lack of fine-grained predictions is significant for several important reasons. First, it limits the utility of the task to downstream applications. Fundamentally, the visual entailment task requires models to search for visual evidence necessary to make logical inferences about the text. This capability has many possible downstream use cases, from detecting image-text inconsistencies for misinformation detection [32] to ensuring answers to questions are entailed by the image [44]. However, many of these tasks require localized predictions [19], which existing methods are unable to provide. Secondly, the fine-grained predictions produced by the model naturally explain its reasoning, making its prediction interpretable.

To address the above shortcomings, we propose the **fine-grained visual entailment** task. Similar to the original visual entailment task, our goal is to predict the logical relationship of textual claims about images. But differently, we require the model to make predictions for each specific "claim" in the text. We illustrate our task in Figure 1. Our method works by decomposing the textual hypothesis into its constituent parts which we call "knowledge elements" (KEs). Knowledge elements are the claims that collectively constitute the entire hypothesis' meaning. In order to decompose the hypothesis into its constituent KEs, we represent the hypothesis as its abstract meaning representation (AMR)

[3, 25] graph. We choose AMR to represent the semantics of the hypothesis because AMR captures the semantic meaning of text irrespective of its syntax [3].

We train a multimodal transformer to make fine-grained predictions on nodes and tuples (i.e. the KEs) within the AMR graph. Our transformer takes as input visual tokens, the hypothesis, and a linearized representation [21, 35] of the AMR graph. To make predictions for each KE within the AMR graph, we introduce a novel local aggregation mechanism which learns a localized contextual representation. Our contextual representation fuses the representation of the KE's AMR with its associated visual context. These representations are then used to make KE-specific predictions.

The model described above requires KE-level supervision to train, but our dataset only contains *sample*-level supervision. Rather than rely on expensive and hard to obtain AMR graph annotations, we instead propose a novel multi-instance learning (MIL) approach. Specifically, we leverage the sample-level label to impose a set of MIL constraints on our prediction function which induce a fine-grained labeling of the graph without requiring any new annotations.

While our MIL losses ensure that knowledge element-level predictions are consistent with the sample-level label, they do not ensure that they are semantically consistent with each other. Thus, we impose both top-down and bottom-up semantic structural constraints which penalize the model for semantically inconsistent knowledge element predictions. These constraints leverage the same intuition as our sample-level MIL constraints, but instead work *between* KEs at different structures within the AMR graph.

In order to benchmark performance on this task, we densely annotate AMR graphs at the knowledge element-level. We compare our method against a number of baselines across numerous metrics. Experiments show that our approach substantially outperforms all baselines for the fine-grained visual entailment task. We also include detailed qualitative results showing our method produces semantically plausible predictions at the knowledge element-level as well as examples of the "visual evidence" chosen by our model to make its predictions.

The major contributions of this paper are as follows:

- We introduce the novel task of fine-grained visual entailment and contribute a benchmark of AMR graphs densely annotated by experts.
- We propose a novel method for this task which relies on localized cross-modal contextual representations of each knowledge element.
- We develop a number of novel loss functions to train our method to make knowledge element-level predictions with only sample-level labels.
- We perform detailed experiments and ablations of our model and loss functions which clearly demonstrate the superiority of our approach. We also present qualitative results and show the "visual evidence" used by our method.

2 Related work

Textual entailment. Textual entailment (predicting whether a hypothesis is entailed by a premise) has long been studied by the natural language processing

[13] community. Later work such as SICK [34] and the Stanford natural language inference benchmark (SNLI) [6] expanded the task definition to allow more granular labels , i.e. by adding the neutral category. The SNLI benchmark is a large-scale benchmark of crowdsourced hypotheses written for Flickr30K [52] image captions (which served as the premises). [7] further extend the SNLI dataset with human-written free-form text explanations of the sample's label. None of these works operate multimodally or make granular predictions as we do.

Visual entailment. Most related to this paper is past work in visual entailment. Xie et al. [50] introduced the visual entailment task which replaced the textual premise from SNLI (a Flickr30K image caption) with its corresponding image, while preserving the original label. Other work [11, 30] has also explored visual entailment in the video domain. [24] observed that the visual entailment dataset contained substantial label errors caused by replacing the image caption with its image. To correct this, [24] reannotated the samples labeled neutral in the test set and proposed an automatic technique to correct some mislabeled neutral train examples. [24] also use the rationales from [7] to train a text generation method to generate free-form textual "explanations" of their predictions.

Our approach offers several significant benefits over [24]'s. First, while [24] generate explanations, there is no guarantee that the generated text truly describes the model's reasoning. Moreover, the generated text may not address specific claims (or any claims) made in the hypothesis. Unlike prior work, our method decomposes the hypothesis into its constituent KEs. Our KE-level predictions naturally cover all claims within the hypothesis which may be important for downstream applications, while inherently explaining the model's reasoning.

Explainability. Our work is also related to research in producing explainable and interpretable predictions. Common examples include as saliency-map techniques [43, 58, 60] as well attention mechanism visualizations [18, 29]. One popular such example of the former category is Grad-CAM [40] which computes class-specific gradient heatmaps with the input. [48] produce fine-grained visual explanations of image regions which caused the model to predict a particular class. More recent work visualizes the attention maps in transformers [8, 9, 47]. Similar to our method, [11] produce grounded video regions as explanations for video entailment, but do not tackle the fine-grained entailment setting as we do. [19] make fine-grained predictions of image-text inconsistency using a predefined ontology, but do not consider the open domain and more granular entailment problem we do.

Multi-instance learning. Our model is required to learn which knowledge elements are entailed, neutral, or contradictory from only the sample-level label. Our work is thus related to multi-instance learning (MIL) methods where a bag of samples are assigned a single label with at least one sample in the bag being the label of the bag [16]. MIL methods have recently been explored for a variety of tasks including image classification [39, 49], object detection [17, 51, 54], scene graph generation [42], and video segment localization [33, 59]. All share the goal of learning a finer-grained prediction function than directly available from the training labels. We are the first to apply MIL techniques to learn a knowledge element-level entailment prediction model.

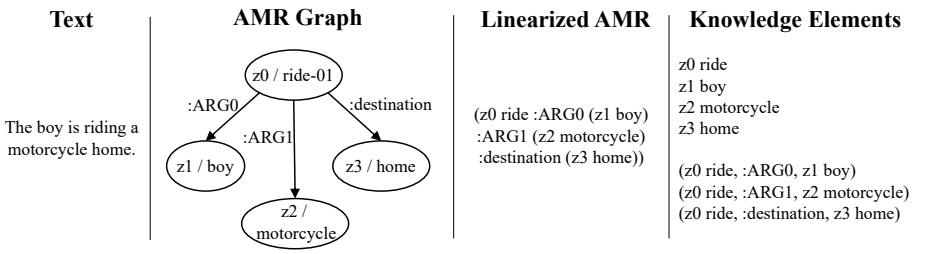

Fig. 2. Illustration of knowledge element (KE) extraction. `Arg#` tokens are role labels from PropBank [36], while `z#` tokens are IDs for nodes to facilitate node coreference. During preprocessing, we make simplifications to the linearized AMRs such as removing "/" and "-01". We include details on preprocessing and tokenization in our supplementary.

3 Fine-grained Visual Entailment

Given an image and a textual hypothesis about the image, our goal is to predict the logical relationship of the image to every "assertion" contained within the hypothesis. To do so, we transform the textual hypothesis into its abstract meaning representation (AMR) graph. We make predictions for each node and tuple (edge with its endpoint nodes) in the graph, which we call knowledge elements (KEs). KEs consist of assertions within the hypothesis about actions, entities (objects, people, etc.), colors, gender, count, etc. We propose a multimodal transformer which operates over the image, hypothesis, and a representation of the AMR graph. At a high level, our method works by locating KE tokens in the AMR graph, aggregating visual information to create a contextualized KE embedding, and performing predictions by a classifier trained with novel multi-instance learning and structural losses (which enforce semantic consistency).

3.1 Problem formulation

More formally, let $\mathcal{D} = \left\{\left(i^{(1)}, h^{(1)}, y^{(1)}\right), \ldots, \left(i^{(s)}, h^{(s)}, y^{(s)}\right), \ldots, \left(i^{(n)}, h^{(n)}, y^{(n)}\right)\right\}$ represent a dataset of image, hypothesis, and entailment label triples respectively, where n is the number of sample triples within the dataset. The goal of the standard visual entailment task [50] is to learn the prediction function $f_\theta(i^{(s)}, h^{(s)}) = y^{(s)}$, i.e., to predict the sample-level logical relationship of the image and hypothesis. Note that in the remainder of this text, we omit the sample index and refer to a single sample for clarity unless noted. Because we seek to make sub-hypothesis-level predictions, we first decompose each hypothesis into its constituent KEs. We denote the set of KEs extracted from h as $KE = \{ke_j\}_{j=0}^{|KE|}$. We seek a prediction function $g_\theta\left(i, h\right) = \left\{y_{ke_j}\right\}_{j=0}^{|KE|}$ where y_{ke_j} is the label of ke_j describing its specific logical relationship with the image.

3.2 Knowledge element extraction

We next describe more specifically how we extract KEs from the hypothesis. Let $\mathcal{G}$ be a text to AMR graph prediction method, then $\mathcal{G}(h) = G$ represents the conversion of a hypothesis into its AMR graph representation G, where $G = (V, E)$ and V and E represent the set of vertices (nodes) and edges within the graph, respectively. Each node $v \in V$ represents a simple, atomic statement about the image (e.g., "there is a car"; "something is walking"). In contrast, each edge $\overrightarrow{e} \in E$ is directed and defines the relationship type between nodes. Because edges are of ambiguous meaning without the nodes they connect, we do not consider edges independently as KEs. Instead, we consider the set of directed node-edge tuples denoted $T = (v_h, e, v_t)$ with $\{v_h, v_t\} \in V$, v_h and v_t being the head and tail nodes defined by the edge's direction, and $\overrightarrow{e}$ defining the relation type between the nodes. Each tuple thus represents a composite statement (e.g., "v_t is performing the action in v_h"; "v_t is the color of v_h"). Thus, the set of KEs extracted for a sample is $KE = \{V \cup T\} = \{ke_j\}_{j=0}^{|KE|}$. We show an example of what our AMR to KE extraction process looks like in Figure 2.

3.3 Architecture

Figure 3 shows the architecture of our method. In this work, we use OSCAR$^+$ [56] as our multimodal encoder $\mathcal{F}$. OSCAR$^+$ achieved recent SOTA performance on several downstream vision-language tasks [56]. Given an image-hypothesis pair (i, h), a pretrained object detector first extracts a sequence of object region features $o = o_1, \ldots, o_m$ and a sequence of predicted object tags $t = t_1, \ldots, t_p$ (object labels in text form) from the image, where m and p are the number of regions and tags respectively. Let f_φ denote a graph linearization method [35], then $r = f_\varphi(\mathcal{G}(h))$ where r is h's AMR in linearized form. The linearized AMR encapsulates all KEs within the hypothesis' AMR in a string form, while retaining their semantic structure. We extract the token embeddings from the last layer of the encoder for each of the sequences: $(\mathbf{o}, \mathbf{t}, \mathbf{r}, \mathbf{h}) = \mathcal{F}(o, t, r, h)$. Although providing h to the model is not required, it provides context and we find it slightly improves sample-level performance ($\sim 2\%$) in practice. Our method for fine-grained KE prediction to be described does not involve or require $\mathbf{h}$.

3.4 Knowledge element-contextual aggregation

Although multimodal interactions happen naturally through self-attention in the transformer's layers, we find it beneficial to apply attention to tokens inside each KE. For example, the model might pay more attention to predicates than entity labels such as "`z0`". Let $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_l)$ be the full AMR embedding sequence, where $l = |r|$. Let a subset $\mathbf{r}_{ke_j} = (\mathbf{r}_{l_1}, \ldots, \mathbf{r}_{l_k}) \in \mathbb{R}^{m_j \times d}$, where d is the hidden state dimension and $\{l_1, \ldots, l_k\} \subseteq \{1, \ldots, l\}$, be the embeddings of the m_j tokens (not necessarily consecutive) that form $ke_j \in KE$. To estimate each token's importance, we learn a function f_α that takes as input $\mathbf{r}_{ke_j}$ and outputs token-wise attention weights $\mathbf{w}_j \in R^{m_j}$. In this work, $f_\alpha(\mathbf{r}_{ke_j}) = \sigma(\mathbf{r}_{ke_j}\mathbf{w}^\alpha)$,

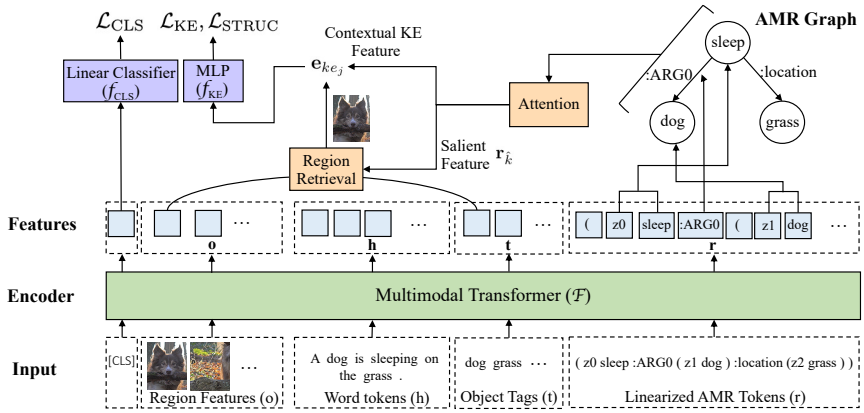

Fig. 3. Our method that makes predictions on both the sample level and the KE level. We show the processing steps for an example KE: `(z0 sleep, :ARG0, z1 dog)`.

where σ is the softmax function and $\mathbf{w}^{\alpha} \in \mathbb{R}^d$ is a learned vector shared by all KEs. We leave exploring stronger weighting mechanisms as future work.

A challenging aspect of our task requires the model to create grounded representations for individual KEs. For a single node in the AMR graph, context information from the image space could be essential for prediction. Therefore, we propose a method to retrieve relevant image region features for each $\mathbf{r}_{ke_j}$.

We first select the most salient token $\hat{k}$ characterized by having the highest attention weight ($\hat{k} = \arg\max_k w_{j,k}$, where $w_{j,k}$ is the k-th element of $\mathbf{w}_j$). We then compare the cosine similarity between $\mathbf{r}_{\hat{k}}$ and each of the tag embeddings $\mathbf{t}_{k'}$, $\frac{\mathbf{r}_{\hat{k}} \cdot \mathbf{t}_{k'}}{\|\mathbf{r}_{\hat{k}}\| \|\mathbf{t}_{k'}\|}$, and retrieve the most similar tag $t_{\hat{k}'}$. With the correspondence given by the object detector, we can now retrieve the object region feature $\mathbf{o}_{\hat{k}''} \in \mathbb{R}^d$ that $t_{\hat{k}'}$ refers to.

The full contextualized embedding used for predicting ke_j's label is therefore obtained by:

$$\mathbf{e}_{ke_j} = \text{Concatenate}([\mathbf{r}_{ke_j}^{\top} \mathbf{w}_j, \mathbf{o}_{\hat{k}''}]) \quad \in \mathbb{R}^{2d} \ . \tag{1}$$

We denote our classifier's output for each KE as $f_{KE}\left(\mathbf{e}_{ke_j}\right) = \mathbf{z} = \left(\mathbf{z}_1, \ldots, \mathbf{z}_{|KE|}\right) = \left(f_{KE}\left(\mathbf{e}_{ke_1}\right), \ldots, f_{KE}\left(\mathbf{e}_{ke_{|KE|}}\right)\right)$, where $\mathbf{z}_i$ denotes the output logits of the classifier for ke_i and $f_{KE}(\cdot) \in \mathbb{R}^3$.

3.5 Multi-instance learning losses

Because we lack KE-level supervision, to train f_{KE} we leverage novel multi-instance learning (MIL) objectives which we derive from the problem semantics. Specifically, we observe that if a hypothesis is entailed by an image, all KEs within the hypothesis should themselves be entailed (denoted ent.). Formally, $(y = \text{ent}) \implies \forall_{ke_i}(y_{ke_i} = \text{ent})$. Because there is no ambiguity as to what each

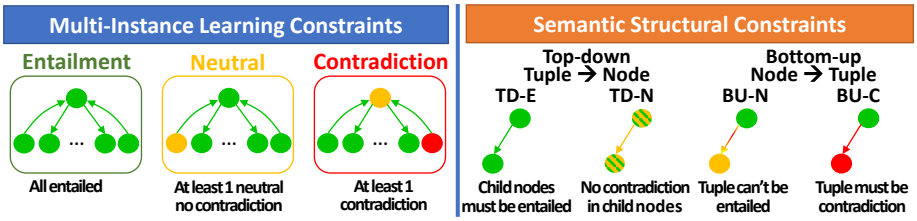

Fig. 4. Left: Our MIL constraints transfer sample-level supervision to KEs, but allow semantically inconsistent predictions (e.g. bicycle=contradiction, riding bicycle=entailed). **Right:** Structural constraints work within the graph to ensure semantically consistent predictions. Multicolor nodes / tuples indicate the KE could be either color.

KE's label should be, we impose a standard cross entropy classification loss across all KEs for entailed samples:

$$\mathcal{L}_{KE_{y=\text{ent}}} = \sum_{i}^{|KE|} -\log \frac{\exp\left(z_{i_{\text{ent}}}\right)}{\sum_{c}^{C} \exp\left(z_{i_c}\right)} \tag{2}$$

where z_{i_c} is the classifier's predicted value for class $c \in \{\text{ent, neu, con}\}$.

For a sample to be labeled neutral, we next observe that its KEs must observe the following definition: $(y = \text{neu}) \implies (\forall_{ke_i} \neg (y_{ke_i} = \text{con}) \wedge \exists_{ke_i} (y_{ke_i} = \text{neu}))$. That is, no KE may be labeled contradiction (but they may be labeled either entailment or neutral) and at least one KE should be labeled neutral. We enforce these two constraints through the following MIL loss:

$$\mathcal{L}_{KE_{y=\text{neu}}} = \sum_{i}^{|KE|} \left(-\log \left(1 - \frac{\exp\left(z_{i_{\text{con}}}\right)}{\sum_{c}^{C} \exp\left(z_{i_c}\right)} \right) \right) - \arg\max_{z_{ineu} \in \mathbf{z}} \left(\log \frac{\exp\left(z_{i_{\text{neu}}}\right)}{\sum_{c}^{C} \exp\left(z_{i_c}\right)} \right) \tag{3}$$

where $\arg\max_{z_{ineu} \in \mathbf{z}}$ selects the KE whose score in the neutral dimension is the largest. Intuitively, this amounts to selecting the KE the model is most confident is neutral and treating its label as such.

Finally, for samples labeled contradiction, we note that $(y = \text{con}) \implies \exists_{ke_i} (y_{ke_i} = \text{con})$. In other words, at least one KE must be labeled contradiction. Following the notation used above, we impose the following MIL loss:

$$\mathcal{L}_{KE_{y=\text{con}}} = -\arg\max_{z_{icon} \in \mathbf{z}} \left(\log \frac{\exp\left(z_{i_{\text{con}}}\right)}{\sum_{c}^{C} \exp\left(z_{i_c}\right)} \right) \tag{4}$$

which selects the KE the model is most confident is contradiction and enforces it to be so classified. We illustrate our MIL constraints for these three categories in Figure 4 (left).

3.6 Semantic structural constraints

The above constraints enforce that the KE predictions are consistent with the sample-level label, but they do not ensure that the KE predictions are *internally*

semantically consistent with one another. For example, a model trained with the above constraints would be free to predict the node "girl" as contradictory, but the tuple "girl on bicycle" as entailed. We call this a "bottom-up" semantic structural violation because the tuple's prediction (the parent) is inconsistent with the node's prediction (the child). Like our MIL constraints, our structural constraints flow from the semantics of the problem. We note that two types of bottom-up structural constraints should hold: BU-C) $(y_{ke_i} = \text{con}) \implies \forall_{ke_j \in \text{parent}(ke_i)} \left(y_{ke_j} = \text{con}\right)$ and BU-N) $(y_{ke_i} = \text{neu}) \implies \forall_{ke_j \in \text{parent}(ke_i)} \neg \left(y_{ke_j} = \text{ent}\right)$. BU-C requires that, if a node is contradiction, any parent tuple that contains it must also be contradiction. BU-N requires that if a node is neutral, no parent tuple may be entailed. We enforce BU-C and BU-N through the following two bottom-up structure preserving losses:

$$\mathcal{L}_{\text{STRUC}_{\text{BU-C}}} = \sum_{\substack{ke_i \in V \\ ke_j \in \text{parent}(ke_i)}}^{ke_i, ke_j} -\sigma\left(z_{i_{\text{con}}}\right) \log\left(\frac{\exp\left(z_{j_{\text{con}}}\right)}{\sum_c^C \exp\left(z_{j_c}\right)}\right) \cdot \mathbb{1}\left\{\hat{y}_i = \text{con}\right\} \tag{5}$$

$$\mathcal{L}_{\text{STRUC}_{\text{BU-N}}} = \sum_{\substack{ke_i \in V \\ ke_j \in \text{parent}(ke_i)}}^{ke_i, ke_j} -\sigma\left(z_{i_{\text{neu}}}\right) \log\left(1 - \frac{\exp\left(z_{j_{\text{ent}}}\right)}{\sum_c^C \exp\left(z_{j_c}\right)}\right) \cdot \mathbb{1}\left\{\hat{y}_i = \text{neu}\right\} \tag{6}$$

where σ is the sigmoid function, $\hat{y}_i$ represents ke_i's predicted label (i.e., the maximum scoring class in z_i), and $\mathbb{1}$ is the indicator function. Note that we weight each structural constraint with the confidence of the child's prediction which lessens the impact of incorrectly predicted KEs. We found that this significantly improved performance.

Similarly, two top-down constraints must also hold for their predictions to be logically consistent: TD-E) $(y_{ke_i} = \text{ent}) \implies \forall_{ke_j \in \text{child}(ke_i)} \left(y_{ke_j} = \text{ent}\right)$ and TD-N) $(y_{ke_i} = \text{neu}) \implies \forall_{ke_j \in \text{child}(ke_i)} \neg \left(y_{ke_j} = \text{con}\right)$. Analogous to our bottom-up constraints, we enforce our top-down constraints through the following two losses:

$$\mathcal{L}_{\text{STRUC}_{\text{TD-E}}} = \sum_{\substack{ke_i \in V \\ ke_j \in \text{child}(ke_i)}}^{ke_i, ke_j} -\sigma\left(z_{i_{\text{ent}}}\right) \log\left(\frac{\exp\left(z_{j_{\text{ent}}}\right)}{\sum_c^C \exp\left(z_{j_c}\right)}\right) \cdot \mathbb{1}\left\{\hat{y}_i = \text{ent}\right\} \tag{7}$$

and

$$\mathcal{L}_{\text{STRUC}_{\text{TD-N}}} = \sum_{\substack{ke_i \in V \\ ke_j \in \text{child}(ke_i)}}^{ke_i, ke_j} -\sigma\left(z_{i_{\text{neu}}}\right) \log\left(1 - \frac{\exp\left(z_{j_{\text{con}}}\right)}{\sum_c^C \exp\left(z_{j_c}\right)}\right) \cdot \mathbb{1}\left\{\hat{y}_i = \text{neu}\right\} . \tag{8}$$

3.7 Final loss formulation

In addition to our KE-level losses, we also include a standard sample-level cross-entropy loss performed on the CLS token of the transformer which we denote

by $\mathcal{L}_{CLS}$. Thus, our final loss formulation is given by the summation of the previous losses: $\mathcal{L} = \beta_{\text{CLS}} * \mathcal{L}_{\text{CLS}} + \beta_{\text{KE}} * \mathcal{L}_{\text{KE}} + \beta_{\text{STRUC}} * \mathcal{L}_{\text{STRUC}}$, where β are hyperparameters controlling the relative weight of each component of the loss.

3.8 Implementation details

All methods and baselines use a pretrained VinVL [56] multimodal transformer as our backbone architecture with the ResNeXt-152 C4 detector for visual features. We use a max length of 50 for $\mathbf{o}$ and 165 for $|\mathbf{t}| + |\mathbf{r}| + |\mathbf{h}|$, truncating $\mathbf{h}$. We use a batch size of 128, an initial learning rate of $5e^{-5}$ that linearly decreases, a weight decay of 0.05, and train for a max of 10 epochs. We use Spring [4] to extract AMR graphs from hypotheses. We use a depth-first approach for AMR linearization. We implement our model in PyTorch [37]. Training takes approximately two days on four Nvidia Titan RTX GPUs. Unless otherwise specified, we set $\beta_{\text{CLS}} = 0.5$ and $\beta_{\text{KE}} = \beta_{\text{STRUC}} = 1$. We include additional details in supplementary.

4 Experiments

We compare our method to several baselines on the new task of fine-grained visual entailment. Our results consistently demonstrate that our approach significantly outperforms these baselines on the fine-grained visual entailment task. We also include detailed ablations and analysis of various components of our method. We also present qualitative results illustrating that our method makes semantically meaningful predictions at the KE-level. Finally, we show the visual regions chosen by our method to make its predictions for each KE.

4.1 Dataset

The original visual entailment benchmark [50] was found to have a substantial ($\sim$39%) label error rate for the neutral class [24]. For training and testing, we therefore use the relabeled version presented in [24] which corrects this issue in the test set. The dataset contains 430,796 image, hypothesis, label triples in total. We use the original train/val/test splits [24].

In order to evaluate our method's on the KE-level prediction task, we require KE-level annotated data. To do so, we created an web annotation interface using LabelStudio [45]. Our interface shows annotators the image, hypothesis, and an image of the hypothesis AMR graph. Annotators annotate each node and tuple (the KEs) within the graph with the class that describes its relationship to the image. Note that we also allow annotators to "opt-out" of KEs that are of unclear meaning (which may occur from AMR prediction errors, etc.). These KEs are ignored in evaluation. Because the original sample labels are crowdsourced and still noisy, we also ask annotators to provide a new sample-level label.

Annotating AMR graphs is an intellectually demanding and laborious task, often requiring annotators to consult PropBank [36] or the AMR specifications [25] to understand the meaning of the KE. Because of the difficulty of the task, we

Table 1. We show KE-level accuracies at the class-level, across different types of KEs, and the overall KE-level accuracy. Finally, we show the structural constraint accuracy (see text). The best result per column is shown in bold and second best is underlined.

Method	Acc_{ent}	Acc_{neu}	Acc_{con}	Acc_{node}	Acc_{tup}	Overall Acc	$\text{Acc}_{\text{STRUC}}$
VE → KE	49.77	**60.00**	<u>83.33</u>	54.36	60.51	57.17	**100**
VE+AMR → KE	55.51	<u>58.06</u>	**85.10**	58.14	64.10	60.86	96.40
w/o $\mathcal{L}_{\text{STRUC}}$	**88.87**	15.48	9.57	66.88	57.17	62.44	70.26
w/o Region Retrieval	78.20	24.51	55.67	64.61	64.87	64.73	93.57
Ours	79.64	31.29	62.76	<u>69.79</u>	<u>66.02</u>	<u>68.07</u>	96.36
Ours+CLS	<u>80.35</u>	29.35	62.76	**70.01**	**66.02**	**68.18**	<u>96.98</u>

concluded it was inappropriate for crowdsourcing. We instead employed two expert annotators who were familiar with AMR, similar to [3, 5]. Collectively, our annotators annotated 1909 KEs (1113-e, 306-n, 282-c, 208-opt-out) from 300 random samples (100-e, 100-n, 100-c) from the test set. Our annotation process took ∼75 hours of effort. We include more details in our supplementary.

4.2 Baselines

Because we are the first to tackle the fine-grained visual entailment task, there are no standard baselines for this task. We thus formulate two baselines in order to benchmark our method's performance. VE→KE is a standard sample-level visual entailment model which takes as input the visual features and the hypothesis. We replicate the model's sample-level prediction for every KE. VE+AMR→KE is similar to the previous model but also takes as input the linearized AMR of the hypothesis. At test time, we make a prediction for each KE separately by feeding the AMR corresponding to each KE into the model.

4.3 Quantitative results

We experimentally compare several variants of our method with the baselines described above. Ours is our full method described in Sec. 3. We also include an ablation of our method showing the performance of our method without our structural constraints (w/o $\mathcal{L}_{\text{STRUC}}$) and without our KE-specific region retrieval technique (w/o Region Retrieval). The latter can be directly applied on top of encoders not based on object features as well. Finally, we show a version of our method (Ours+CLS) that predicts all KEs as entailed when the sample-level is predicted entailed. Otherwise, it uses our KE-specific classifier.

KE-level performance. In Table 1 we show the performance of each method on our KE-level annotations. The first group of results shows the accuracy for each ground truth KE class (here equivalent to recall). VE→KE achieves strong performance for KEs labeled neutral and particularly contradiction. This is because contradiction KEs only appear in contradiction samples, while neutral and entailment KEs can appear in multiple categories. However, we observed that relatively few neutral KEs appear in the contradiction category, because

hypotheses usually mention untrue facts (contradiction) about something truly in the image (entailment). Thus, baseline performance is also strong for the neutral category. We note that the KE classes are unbalanced by nature of our task and we seek to locate the entailment KEs in neutral and contradiction samples. To this end, our method substantially outperforms both baselines for entailment KEs which can appear in any type of sample and are thus by far the most frequent. The baseline's apparent strength is because the first group of columns don't account for the *precision* of each KE category. In our supplementary, we show that our predictions much more closely align with the ground truth KE distribution for each sample type. This performance is reflected in the node, tuple, and overall accuracies on which our method performs best.

The second group of results shows accuracies across different types of KEs. We note our method performs better on nodes than tuples. This is unsurprising because nodes represent simpler statements about the presence of objects or actions, while tuples make more complex, composite statements about nodes and are thus harder to verify for the model. While VE+AMR→KE outperforms VE→KE overall and for most metrics, their predictions are highly similar for most samples and it is significantly outperformed by our method overall.

We next measure the number of structural violations. We consider each parent-child KE relationship a separate instance and calculate the accuracy of each method at producing semantically consistent predictions across parent-child relationships. We note that while VE→KE has no structural violations (because all KEs are predicted as the sample label), it makes no fine-grained predictions. Of the methods that make such predictions, our method performs best.

Finally, we measure the performance of our ablated method and our method's variants. Without $\mathcal{L}_{\text{STRUC}}$ our method achieves the lowest overall performance of our method, indicating that the structural constraints work synergistically with our MIL constraints to further disambiguate the KE-level labels. We further note that without $\mathcal{L}_{\text{STRUC}}$ our model has a high rate ($\sim 30\%$) of structural violations within the graph. We next show that aggregating visual features into our model's contextual embedding is important. Without region retrieval our model's accuracy drops by (-3.45% acc) because our KE's embeddings lack relevant visual context (especially for nodes). Finally, we observe that ignoring our KE-level predictions for sample's predicted entailment and predicting all KEs as entailed (Ours+CLS) slightly improves performance (+0.11% acc).

Sample-level performance. Though not our focus, we also include the performance of sample-level label prediction in Table 2. The left side shows the performance on the labels given by the MTurkers in [24], while the right shows the performance on the sample labels in our expertly annotated set (Relab.). We explore two ways of predicting the sample label. The first uses the CLS token, while the second uses the logical rules defined in Sec. 3.5 to produce the sample label from the predicted KE labels. We observe a slight drop (0.38%) in the sample-level label for our method on the crowdsourced labels. Aside from label noise, one possible reason is that the model pays less attention to the sample-level task and focuses on the KE-level task (see ablation on loss weightings in our sup-

Table 2. We show the sample-level accuracy of each method across different metrics. The best method for each metric is shown in bold and the second best is underlined.

Method	Acc_{CLS}	$\text{Acc}_{\text{KE}\rightarrow\text{CLS}}$	$\text{Acc}^{\text{Relab.}}_{\text{CLS}}$	$\text{Acc}^{\text{Relab.}}_{\text{KE}\rightarrow\text{CLS}}$	$\text{Acc}^{\text{Relab.}}_{\text{Best}}$
VE → KE	**80.37**	-	**79.73**	-	79.73
VE+AMR → KE	-	**79.15**	-	<u>80.06</u>	<u>80.06</u>
w/o $\mathcal{L}_{\text{STRUC}}$	79.78	75.17	<u>79.40</u>	79.73	79.73
w/o Region Retrieval	79.58	75.21	<u>79.40</u>	<u>80.06</u>	<u>80.06</u>
Ours	<u>79.99</u>	76.26	78.73	**80.73**	**80.73**
Ours+CLS	<u>79.99</u>	<u>77.70</u>	78.73	77.74	78.73

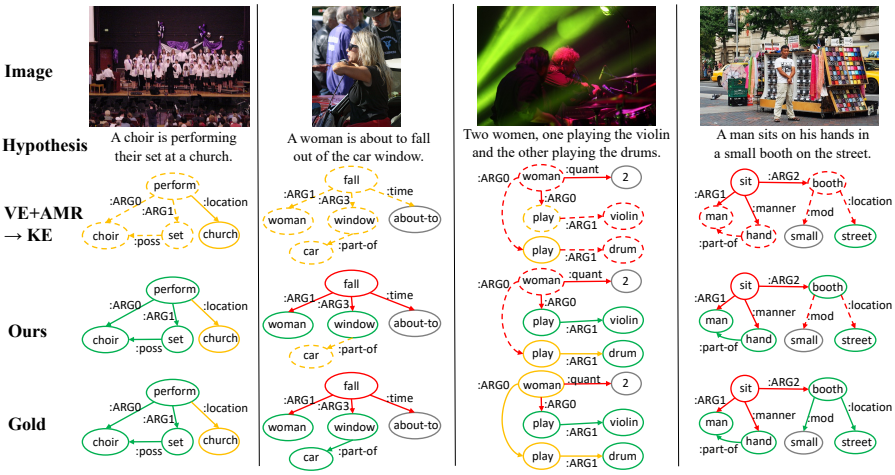

Fig. 5. Qualitative results showing our KE-level predictions on AMR graphs compared to the baseline. Nodes and edges (representing tuples) are colored based on their predicted label (ent, neu, con, opt-out). Wrong predictions are denoted by dashed lines.

plementary). We show that on our set of expertly annotated samples, producing the sample-level label using our KE predictions outperforms all baselines.

4.4 Qualitative results

In this section, we present qualitative results showcasing our method's KE-level predictions. We also illustrate the visual region our method selects for each KE.

KE-level prediction results. In Figure 5, we show KE-level prediction results. We observe that the baseline often predicts many KEs the same label. In contrast, our model's predictions are more diverse and reasonable. In the first column, our model correctly concludes the location can't be determined and marks "`church`" neutral. In the next column, our model struggles to detect a car in the image so incorrectly marks "`car`" as neutral, but correctly infers that "`fall`" is contradictory. In the rightmost column, our model "incorrectly" says the booth

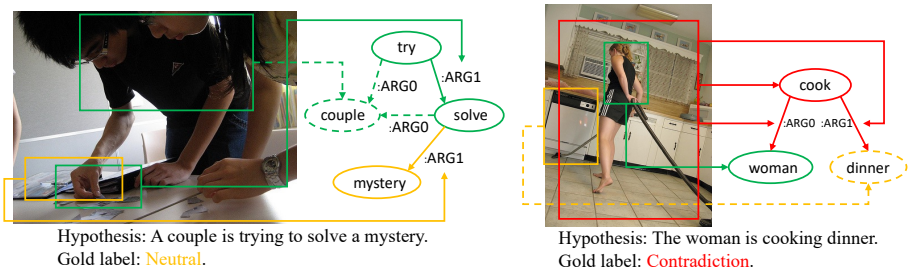

Fig. 6. We show the image region our model selects for each KE prediction. KEs are colored based on their predicted label. Wrong predictions are denoted by dashed lines.

is not small, but this is subjective. Similarly, our model says the booth is not located on the street, which is feasible because the booth is on the sidewalk.

Region selection results. In Figure 6 we show the image regions selected by our model for each KE. We observe relevant KE→image localization results. For example, on the left we observe that "`(try, :ARG1, solve)`" retrieves the puzzle being assembled and "`couple`" retrieves the two people. On the right, the model selects the upper body of the woman for "`woman`" and a large region of the image to conclude the action "`cook`" is incorrect.

5 Conclusion

We introduced the novel problem of fine-grained visual entailment where the goal is to predict the logical relationship of knowledge elements extracted from a piece of text with an image. We proposed a model for this task which fuses relevant visual features with the representation of each knowledge element. Because of a lack of fine-grained annotations, we proposed novel multi-instance learning losses to transfer sample-level supervision to the knowledge element-level. We also proposed novel semantic structure preserving constraints. Experiments conducted on a new benchmark show that our approach significantly outperforms relevant baselines, and more importantly, produces interpretable predictions.

There are several possible directions for future work. For example, pretraining encoders with human-created AMR inputs [4, 22] may better prepare encoders for our task. While we use object labels in our work, object attributes are also potentially helpful for predicting KEs that involves attributes [56]. Finally, region retrieval [2, 27, 53] and graph networks [46] may also be exploited.

Acknowledgements This research is based upon work supported by DARPA SemaFor Program No. HR001120C0123. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Babar, S., Das, S.: Where to look?: Mining complementary image regions for weakly supervised object localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1010–1019 (2021)
3. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N.: Abstract meaning representation for sembanking. In: Proceedings of the 7th linguistic annotation workshop and interoperability with discourse. pp. 178–186 (2013)
4. Bevilacqua, M., Blloshmi, R., Navigli, R.: One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In: Proceedings of AAAI (2021)
5. Bonial, C., Badarau, B., Griffitt, K., Hermjakob, U., Knight, K., O'Gorman, T., Palmer, M., Schneider, N.: Abstract meaning representation of constructions: The more we include, the better the representation. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
6. Bowman, S., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 632–642 (2015)
7. Camburu, O.M., Rocktäschel, T., Lukasiewicz, T., Blunsom, P.: e-snli: Natural language inference with natural language explanations. Advances in Neural Information Processing Systems **31** (2018)
8. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 397–406 (2021)
9. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 782–791 (2021)
10. Chen, B., Lin, X., Thomas, C., Li, M., Yoshida, S., Chum, L., Ji, H., Chang, S.F.: Joint multimedia event extraction from video and article. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 74–88 (2021)
11. Chen, J., Kong, Y.: Explainable video entailment with grounded visual evidence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
12. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
13. Dagan, I., Glickman, O., Magnini, B.: The pascal recognising textual entailment challenge. In: Machine Learning Challenges Workshop. pp. 177–190. Springer (2005)
14. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M.F., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
15. Das, A., Kottur, S., Moura, J.M., Lee, S., Batra, D.: Learning cooperative visual dialog agents with deep reinforcement learning. In: Proceedings of the IEEE international conference on computer vision. pp. 2951–2960 (2017)

16. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)
17. Dong, B., Huang, Z., Guo, Y., Wang, Q., Niu, Z., Zuo, W.: Boosting weakly supervised object detection via learning bounding box adjusters. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2876–2885 (2021)
18. Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Attention branch network: Learning of attention mechanism for visual explanation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10705–10714 (2019)
19. Fung, Y., Thomas, C., Reddy, R.G., Polisetty, S., Ji, H., Chang, S.F., McKeown, K., Bansal, M., Sil, A.: Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. 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). pp. 1683–1698 (2021)
20. Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Vqa-lol: Visual question answering under the lens of logic. In: European conference on computer vision. pp. 379–396. Springer (2020)
21. Goodman, M.W.: Penman: An open-source library and tool for amr graphs. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 312–319 (2020)
22. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)
23. Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., Fu, J.: Seeing out of the box: End-to-end pre-training for vision-language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12976–12985 (2021)
24. Kayser, M., Camburu, O.M., Salewski, L., Emde, C., Do, V., Akata, Z., Lukasiewicz, T.: e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1244–1254 (2021)
25. Knight, K., Badarau, B., Baranescu, L., Bonial, C., Bardocz, M., Griffitt, K., Hermjakob, U., Marcu, D., Palmer, M., O'Gorman, T., Schneider, N.: Abstract meaning representation (AMR) annotation release 3.0. `https://catalog.ldc.upenn.edu/LDC2020T02` (2020)
26. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are you talking about? text-to-image coreference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3558–3565 (2014)
27. Kumar, V., Namboodiri, A., Jawahar, C.: Region pooling with adaptive feature fusion for end-to-end person recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2133–2142 (2020)
28. Li, M., Zareian, A., Zeng, Q., Whitehead, S., Lu, D., Ji, H., Chang, S.F.: Cross-media structured common space for multimedia event extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2557–2568 (2020)
29. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Transactions on Image Processing **28**(5), 2439–2450 (2018)

30. Liu, J., Chen, W., Cheng, Y., Gan, Z., Yu, L., Yang, Y., Liu, J.: Violin: A large-scale dataset for video-and-language inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10900–10910 (2020)
31. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10437–10446 (2020)
32. Luo, G., Darrell, T., Rohrbach, A.: Newsclippings: Automatic generation of out-of-context multimodal media. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6801–6817 (2021)
33. Luo, Z., Guillory, D., Shi, B., Ke, W., Wan, F., Darrell, T., Xu, H.: Weakly-supervised action localization with expectation-maximization multi-instance learning. In: European conference on computer vision. pp. 729–745. Springer (2020)
34. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A sick cure for the evaluation of compositional distributional semantic models. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 216–223 (2014)
35. Matthiessen, C.M., Christian, M., Bateman, J.A., Matthiessen, M.: Text generation and systemic-functional linguistics: experiences from English and Japanese. Burns & Oates (1991)
36. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Computational linguistics **31**(1), 71–106 (2005)
37. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
38. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive image-language cues. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1928–1937 (2017)
39. Rymarczyk, D., Borowa, A., Tabor, J., Zielinski, B.: Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1721–1730 (2021)
40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
41. Sheng, S., Singh, A., Goswami, V., Magana, J., Thrush, T., Galuba, W., Parikh, D., Kiela, D.: Human-adversarial visual question answering. Advances in Neural Information Processing Systems **34** (2021)
42. Shi, J., Zhong, Y., Xu, N., Li, Y., Xu, C.: A simple baseline for weakly-supervised scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16393–16402 (2021)
43. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMLR (2017)
44. Si, Q., Lin, Z., yu Zheng, M., Fu, P., Wang, W.: Check it again: Progressive visual question answering via visual entailment. 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). pp. 4101–4110 (2021)

45. Tkachenko, M., Malyuk, M., Shevchenko, N., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2021), `https://github.com/heartexlabs/label-studio`, open source software available from https://github.com/heartexlabs/label-studio
46. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations (2018)
47. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: 57th Annual Meeting of the Association for Computational Linguistics. pp. 5797–5808. ACL Anthology (2019)
48. Wagner, J., Kohler, J.M., Gindele, T., Hetzel, L., Wiedemer, J.T., Behnke, S.: Interpretable and fine-grained visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9097–9107 (2019)
49. Wu, J., Yu, Y., Huang, C., Yu, K.: Deep multiple instance learning for image classification and auto-annotation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3460–3469 (2015)
50. Xie, N., Lai, F., Doran, D., Kadav, A.: Visual entailment: A novel task for fine-grained image understanding. arXiv preprint arXiv:1901.06706 (2019)
51. Yang, H., Wu, H., Chen, H.: Detecting 11k classes: Large scale object detection without fine-grained bounding boxes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9805–9813 (2019)
52. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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)
53. Yuan, S., Bai, K., Chen, L., Zhang, Y., Tao, C., Li, C., Wang, G., Henao, R., Carin, L.: Weakly supervised cross-domain alignment with optimal transport. In: Proceedings of the British Machine Vision Conference (2020)
54. Yuan, T., Wan, F., Fu, M., Liu, J., Xu, S., Ji, X., Ye, Q.: Multiple instance active learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5330–5339 (2021)
55. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6720–6731 (2019)
56. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5579–5588 (2021)
57. Zhang, T., Whitehead, S., Zhang, H., Li, H., Ellis, J., Huang, L., Liu, W., Ji, H., Chang, S.F.: Improving event extraction via multimodal integration. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 270–278 (2017)
58. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2131–2145 (2018)
59. Zhou, Y., Sun, X., Liu, D., Zha, Z., Zeng, W.: Adaptive pooling in multi-instance learning for web video annotation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 318–327 (2017)
60. Zunino, A., Bargal, S.A., Volpi, R., Sameki, M., Zhang, J., Sclaroff, S., Murino, V., Saenko, K.: Explainable deep classification models for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3233–3242 (2021)