Visual Semantic Parsing: From Images to Abstract Meaning Representation

Mohamed A. Abdelsalam[1], Zhan Shi[1,2*], Federico Fancellu[3†], Kalliopi Basioti[1,4*], Dhaivat J. Bhatt[1], Vladimir Pavlovic[1,4], Afsaneh Fazly[1]
[1]Samsung AI Centre - Toronto , [2]Queen's University , [3]3M , [4]Rutgers University
{m.abdelsalam, d.bhatt, a.fazly}@samsung.com , z.shi@queensu.ca
f.fancellu0@gmail.com , {kalliopi.basioti, vladimir}@rutgers.edu

Abstract

The success of scene graphs for visual scene understanding has brought attention to the benefits of abstracting a visual input (e.g., image) into a structured representation, where entities (people and objects) are nodes connected by edges specifying their relations. Building these representations, however, requires expensive manual annotation in the form of images paired with their scene graphs or frames. These formalisms remain limited in the nature of entities and relations they can capture. In this paper, we propose to leverage a widely-used meaning representation in the field of natural language processing, the Abstract Meaning Representation (AMR), to address these shortcomings. Compared to scene graphs, which largely emphasize spatial relationships , our visual AMR graphs are more linguistically informed, with a focus on higher-level semantic concepts extrapolated from visual input. Moreover, they allow us to generate meta-AMR graphs to unify information contained in multiple image descriptions under one representation. Through extensive experimentation and analysis, we demonstrate that we can re-purpose an existing text-to-AMR parser to parse images into AMRs. Our findings point to important future research directions for improved scene understanding.

1 Introduction

The ability to understand and describe a scene is fundamental for the development of truly intelligent systems, including autonomous vehicles, robots navigating an environment, or even simpler applications such as language-based image retrieval. Much work in computer vision has focused on two key aspects of scene understanding, namely, recognizing entities, including object detection (Liu et al., 2016; Ren et al., 2015; Carion et al., 2020; Liu et al., 2020a) and activity recognition (Herath et al., 2017; Kong and Fu, 2022; Li et al., 2018; Gao et al., 2018), as well as understanding how entities are related to each other, e.g., human–object interaction (Hou et al., 2020; Zou et al., 2021) and relation detection (Lu et al., 2016; Zhang et al., 2017; Zellers et al., 2018).

A natural way of representing scene entities and their relations is in graph form, so it is perhaps unsurprising that a lot of work has focused on graph-based scene representations and especially on scene graphs (Johnson et al., 2015a). Scene graphs encode the salient regions in an image (mainly, objects) as nodes, and the relations among these (mostly spatial in nature) as edges, both labelled via natural language tags; see Fig. 1(b) for an example scene graph. Along the same lines, Yatskar et al. (2016) propose to represent a scene as a semantic role labelled frame, drawn from FrameNet (Ruppenhofer et al., 2016) — a linguistically-motivated approach that draws on semantic role labelling literature.

Scene graphs and situation frames can capture important aspects of an image, yet they are limited in important ways. They both require expensive manual annotation in the form of images paired with their corresponding scene graphs or frames. Scene graphs in particular also suffer from being limited in the nature of entities and relations that they capture (see Section 2 for a detailed analysis). Ideally, we would like to capture event-level semantics (same as in situation recognition) but as a structured graph that captures a diverse set of relations and goes beyond low-level visual semantics.

Inspired by the linguistically-motivated image understanding research, we propose to represent images using a well-known graph formalism for language understanding, i.e., Abstract Meaning Representations (AMRs Banarescu et al., 2013). Similarly to (visual) semantic role labeling, AMRs also represent "who did what to whom, where,

*Work done during an internship at Samsung AI Centre - Toronto

†Work done while at Samsung AI Centre - Toronto

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pages 282 - 300
December 7-8, 2022 ©2022 Association for Computational Linguistics

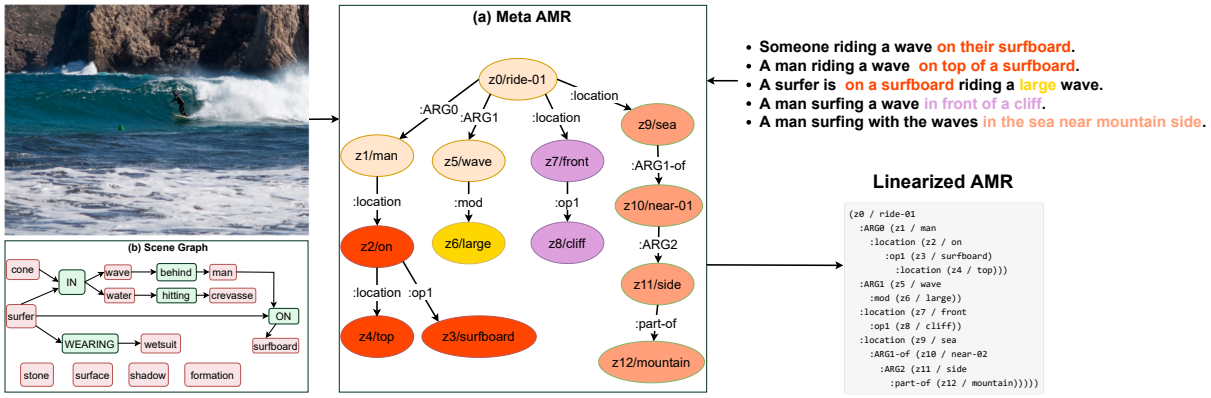


Figure 1: An image from MSCOCO and Visual Genome dataset, along with its five human-generated captions, and: (a) an image-level meta-AMR graph capturing its overall semantics, (b) its human-generated scene graph.

when, and how?" (Màrquez et al., 2008), but in a more structured way via transforming an image into a graph representation. AMRs not only encode the main events, their participants and arguments, as well as their relations (as in semantic role labelling/situation recognition), but also relations among various other participants and arguments; see Fig. 1(a). Importantly, AMR is a broadly-adopted and dynamically evolving formalism (e.g., Bonial et al., 2020; Bonn et al., 2020; Naseem et al., 2021), and AMR parsing is an active and successful area of research (e.g., Zhang et al., 2019b; Bevilacqua et al., 2021; Xia et al., 2021; Drozdov et al., 2022). Finally, given the high quality of existing AMR parsers (for language), we do not need manual AMR annotations for images, and can rely on existing image–caption datasets to create high quality silver data for image-to-AMR parsing. In summary, we make the following contributions:

- We introduce the novel problem of parsing images into Abstract Meaning Representations, a widely-adopted linguistically-motivated graph formalism; and propose the first image-to-AMR parser model for the task.
- We present a detailed analysis and comparison between scene graphs and AMRs with respect to the nature of entities and relations they capture, results of which further motivates research in the use of AMRs for better image understanding.
- Inspired by work on multi-sentence AMR, we propose a graph-to-graph transformation algorithm that combines the meanings of several image caption descriptions into image-level meta-AMR graphs. The motivation behind generating the meta-AMRs is to build a graph that covers most of entities, predicates, and semantic relations contained in the individual caption AMRs.

Our analyses suggest that AMRs encode aspects of an image content that are not captured by the commonly-used scene graphs. Our initial results on re-purposing a text-to-AMR parser for image-to-AMR parsing, as well as on creating image-level meta-AMRs, point to exciting future research directions for improved scene understanding.

2 Motivation: AMRs vs. Scene Graphs

Scene graphs (SGs) are a widely-adopted graph formalism for representing the semantic content of an image. Scene graphs have been shown useful for various downstream tasks, such as image captioning (Yang et al., 2019; Li and Jiang, 2019; Zhong et al., 2020), visual question answering (Zhang et al., 2019a; Hildebrandt et al., 2020; Damodaran et al., 2021), and image retrieval (Johnson et al., 2015b; Schuster et al., 2015; Wang et al., 2020; Schroeder and Tripathi, 2020). However, learning to automatically generate SGs requires expensive manual annotations (object bounding boxes and their relations). SGs were also shown to be highly biased in the entity and relation types that they capture. For example, an analysis by Zellers et al. (2018) reveals that clothing (e.g., *dress*) and object/body parts (e.g., *eyes*, *wheel*) make up over one-third of entity instances in the SGs corresponding to the Visual Genome images (Krishna et al., 2016), and that more than 90% of all relation instances belong to the two categories of geometric (e.g., *behind*) and possessive (e.g., *have*).

One advantage of AMR graphs is that we can draw on supervision through captions associated with images. Nonetheless, the question remains as

to what types of entities and relations are encoded by AMR graphs, and how these differ from SGs. To answer this question, we follow an approach similar to Zellers et al. (2018), and categorize entities and relations in SG and AMR graphs corresponding to a sample of 50K images. We use the same categories as Zellers et al., but add a few new ones to capture relation types specific to AMRs, namely, Attribute (*small*), Quantifier (*few*), Event (*soccer*), and AMR specific (*date-entity*). Details of our categorization process are provided in Appendix A.

Figure 2 shows the distribution of instances for each Entity and Relation category, compared across SG and AMR graphs. AMRs tend to encode a more diverse set of relations, and in particular capture more of the abstract semantic relations that are missing from SGs. This is expected because our caption-generated AMRs by design capture the essential meaning of the image descriptions and, as such, encode how people perceive and describe scenes. In contrast, SGs are designed to capture the content of an image, including regions representing objects and (mainly spatial/geometric) visually-observable relations; see Fig. 1 for SG and AMR graphs corresponding to an image. In the context of Entities, and a major departure from SGs, (object/body) parts are less frequently encoded in AMRs, pointing to the well-known whole-object bias in how people perceive and describe scenes (Markman, 1990; Fei-Fei et al., 2007). In contrast, location is more frequent in AMRs.

The focus of AMRs on abstract content suggests that they have the potential for improving downstream tasks, especially when the task requires an understanding of the higher level semantics of an image. Interestingly, a recent study showed that using AMRs as an intermediate representation for textual SG parsing helps improve the quality of the parsed SGs (Choi et al., 2022), even though AMRs and SGs encode qualitatively different information. Since AMRs tend to capture higher level semantics, we propose to use them as the final image representation. The question remains as to how difficult it is to directly learn such representations from images. The rest of the paper focuses on answering this question.

3 Method

3.1 Parsing Images into AMR Graphs

We develop image-to-AMR parsers based on a state-of-the-art seq2seq text-to-AMR parser,

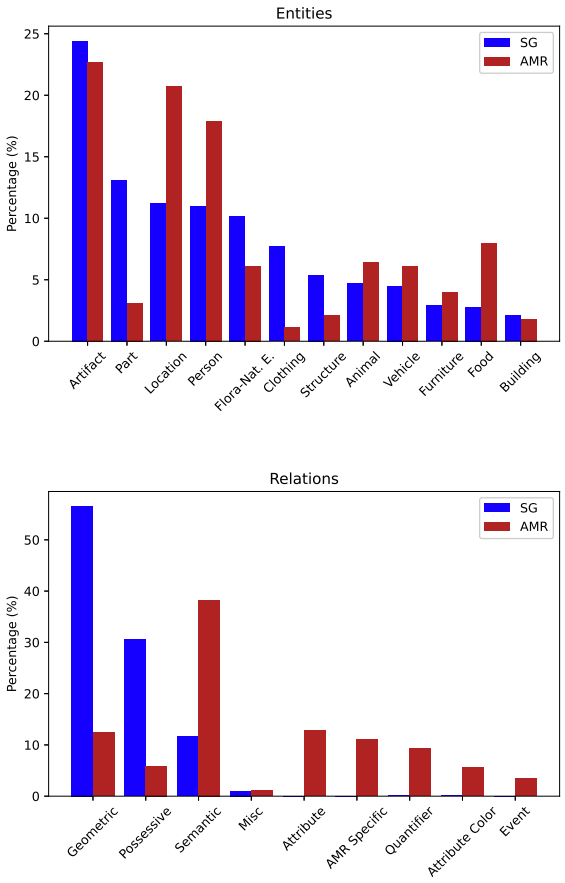


Figure 2: Statistics on a selected set of top-frequency Entity and Relation categories, extracted from the AMR and SG graphs corresponding to around 50K images that appear in both Visual Genome and MSCOCO.

SPRING (Bevilacqua et al., 2021), and a multimodal VL-BART (Cho et al., 2021). Both are transformer-based architectures with a bi-directional encoder and an auto-regressive decoder. SPRING extends a pre-trained seq2seq model, BART (Lewis et al., 2020), by fine-tuning it on AMR parsing and generation. Next, we describe our models, input representation, and training.

Models. We build two variants of our image-to-AMR parser, as depicted in Fig. 3(a) and (b).

- Our first model, which we refer to as IMG2AMR$_{\text{direct}}$, modifies SPRING by replacing BART with its vision-and-language counterpart, VL-BART (Cho et al., 2021). VL-BART extends BART with visual understanding ability through fine-tuning on multiple vision-and-language tasks. With this modification, our model can receive visual features (plus text) as input, and generate linearized AMR graphs.

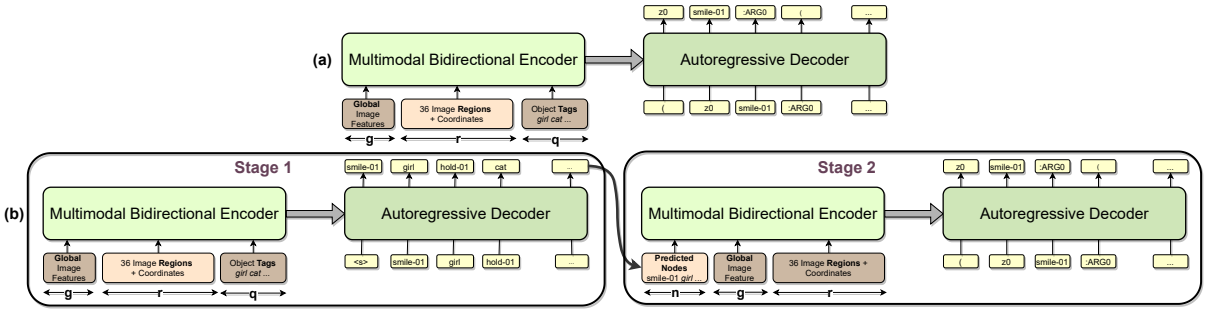


Figure 3: Model architecture for our two image-to-AMR models: (a) IMG2AMR$_{\text{direct}}$: A direct model that uses a single seq2seq encoder–decoder to generate linearlized AMRs from input images; and (b) IMG2AMR$_{\text{2stage}}$: A two-stage model containing two independent seq2seq components. $\mathbf{g}$ and $\mathbf{r}$ stand for global and region features, $\mathbf{q}$ for tag embeddings, and $\mathbf{n}$ for the embeddings of the predicted nodes. The input and output space of the decoders come from the AMR vocabulary.

- Our second model, inspired by text-to-graph AMR parsers (e.g., Zhang et al., 2019b; Xia et al., 2021), generates linearized AMRs in two stages by first predicting the nodes, and then the relations. Specifically, we first predict the nodes of the linearized AMR for a given image. These predicted nodes are then fed (along with the image) as input into a second seq2seq model that generates a linearized AMR (effectively adding the relations). We refer to this model as IMG2AMR$_{\text{2stage}}$.

Input Representation. To represent images, we follow VL-BART, which takes the output of Faster R-CNN (Ren et al., 2015) (i.e., region features and coordinates for 36 regions) and projects them onto $d = 768$ dimensional vectors via two separate fully-connected layers. Faster R-CNN region features are obtained via training for visual object and attribute classification (Anderson et al., 2018) on Visual Genome. The visual input to our model is composed of position-aware embeddings for the 36 regions, plus a global image-level feature ($\mathbf{r}$ and $\mathbf{g}$ in Fig. 3). To get the position-aware embeddings for the regions, we add together the projected region and coordinate embeddings. To get the global image feature, we use the output of the final hidden layer in ResNet-101 (He et al., 2016), which is passed through the same fully connected layer as the regions to obtain a 768-dimensional vector.

Training. To benefit from transfer learning, we initialize the encoder and decoder weights of both our models from the pre-trained VL-BART. This is a reasonable initialization strategy, given that VL-BART has been pre-trained on input similar to ours. Moreover, a large number of AMR labels are drawn from the English vocabulary, and thus the pre-training of VL-BART should also be appropriate for AMR generation. We fine-tune our models on the task of image-to-AMR generation, using images paired with their automatically-generated AMR graphs. We consider two alternative AMR representations: (a) *caption AMRs*, created directly from captions associated with images (see Section 4 for details); and (b) image-level *meta-AMRs*, constructed through an algorithm we describe below in Section 3.2. We perform experiments with either caption or meta-AMRs, where we train and test on the same type of AMRs. For the various stages of training, we use the cross-entropy loss between the model predictions and the ground-truth labels for each token, where the model predictions are obtained greedily, i.e., choosing the token with the maximum score at each step of the sequence generation.

3.2 Learning per-Image meta-AMR Graphs

Recall that, in order to collect a data set of images paired with their AMR graphs, we rely on image–caption datasets such as MSCOCO. Specifically, we use a pre-trained AMR parser to generate AMR graphs from each caption of an image. Images can be described in many different ways, e.g., each image in MSCOCO comes with five different human-generated captions. We hypothesize that these captions collectively represent the content of the image they are describing, and as such propose to also combine the caption AMRs into image-level meta-AMR graphs through a merge and refine process that we explain next.

Prior work has used graph-to-graph transformations for merging sentence-level AMRs into document-level AMRs for abstractive and multi-

Algorithm 1 META-AMR Graph Construction

Input: k human-generated image descriptions $\{c_i\}_{i=1}^{k}$ for a given image i; a set of pre-defined AMR relation types $\mathcal{R}$;
Output: A meta-AMR graph g_{meta};
Initialize: Generate AMR graphs $\{g_i\}$ for the k descriptions using a pre-trained AMR semantic parser; Initialize $g_m = (\mathcal{N}, \mathcal{E})$ to be the null graph.
$\mathcal{N} = \cup_{i=1}^{k} \mathcal{N}_i$
for $i = 1 \sim k$ **do**
  $\mathcal{E}_i = \text{getEdges}(g_i)$
  for $(n_s, n_t) : r \in \mathcal{E}_i$ **do** ▷ (n_s, n_t) is a pair of nodes connected via an edge labeled as r
    if $(n_s, n_t) \notin \mathcal{E}.\text{keys}()$ $\hookrightarrow \wedge (n_t, n_s) \notin \mathcal{E}.\text{keys}()$ $\hookrightarrow \wedge\, r \in \mathcal{R}$ **then**
      $\mathcal{E}.\text{add}(\{(n_s, n_t) : r\})$ ▷ Add a new edge when neither (n_s, n_t) nor (n_t, n_s) previously included, and r belongs to a pre-selected set of AMR relation types $\mathcal{R}$
$\mathcal{G}_m = \text{weaklyConnectedComponents}(g_m)$ ▷ Get all connected components as g_{meta} candidates since it should be a connected graph according to the definition of AMR
$g_{\text{meta}} = \text{getLargestComponent}(\mathcal{G}_m)$ ▷ Get the candidate with the largest number of nodes as it can cover most entities and predicates in the image
$g_{\text{meta}} = \text{refineNodes}(g_{meta})$ ▷ Replace node types by their frequent hypernym if available
return g_{meta}

document summarization (e.g., Liu et al., 2015; Liao et al., 2018; Naseem et al., 2021). Unlike in a summarization task, captions do not form a coherent document, but instead collectively describe an image. Inspired by prior work, we propose our graph-to-graph transformation algorithm that learns a unified meta-AMR graph from caption graphs; see Algorithm 1. Specifically, we first merge the nodes and edges from the original set of k caption-level AMRs, only including a predefined set of relation/edge labels. We then select the largest connected component of this merged graph, which we further refine by replacing non-predicate nodes by their more frequent hypernyms, when available. The motivation behind this refinement process is to reduce the complexity of the meta-AMR graphs (in terms of their size), which would potentially improve parsing performance. An example of a meta-AMR graph generated from caption AMRs is given in Appendix C.

AMR graphs of the MSCOCO training captions contain more than 90 types of semantic relations and more than 21K node types, with long-tailed distributions; see Fig. 6 in Appendix B. To refine meta-AMR graphs, we only maintain the top-20 most frequent relation types that include core roles, such as ARG0, ARG1, etc., as well as high-frequency non-core roles, such as mod and location. To further refine the graphs, we replace each non-predicate node (e.g., *salmon*) with its most frequent hypernym (e.g., *fish*) according to WordNet (Fellbaum, 1998). This results in just about 30% reduction in the number of node types (to 15K). The average complexity of graphs is also reduced from 19 nodes and 23 relations to 16 and 18, respectively.

4 Experimental Setup

Data. For our task of AMR generation from images, we use an augmented version of the standard MSCOCO image–caption dataset, which is composed of images paired with their captions, automatically generated caption-level linearized AMR graphs, and an image-level linearized meta-AMR graph. We use the splits established in previous work (Karpathy and Fei-Fei, 2015), containing 113,287 training, 5000 VALidation, and 5000 TEST images, where each image is associated with five manually-annotated captions. Following the cross-modal retrieval work involving MSCOCO (e.g., Lee et al., 2018), we use a subset of the VAL and TEST sets, containing 1000 images each. AMR graphs of the captions are obtained by running the SPRING text-to-AMR parser (Bevilacqua et al., 2021) that is trained on AMR2.0 dataset.[1] The meta-AMR graph is created from the individual AMRs through our merge and refine process described in Algorithm 1 of Section 3.

Parser implementation details. We initialize our IMG2AMR models from VL-BART, which is based on $\text{BART}_{\text{Base}}$. BART uses a sub-word tokenizer with a vocabulary size of 50,265. Following SPRING, we expand the vocabulary to include frequent AMR-specific tokens and symbols (e.g., :OP, ARG1, temporal-entity), resulting in a vocabulary size of 53,587. The addition of AMR-specific symbols in vocabulary improves efficiency by avoiding extensive sub-token splitting. The embeddings of these additional tokens are initialized by taking the average of the embeddings of their sub-word constituents. The $\text{IMG2AMR}_{\text{direct}}$ models are trained for 60 epochs, while the $\text{IMG2AMR}_{\text{2stage}}$ models are trained for 30 epochs per stage. We use a batch size of 10 with gradients being accumulated for 10 batches (hence an effective batch size of 100), the batch size was limited due to the length of the linearized meta-AMRs. The optimizer used is RAdam (Liu et al., 2020b), with a learning rate

[1] https://catalog.ldc.upenn.edu/LDC2017T10

Model	Train/Test AMRs	SMATCH	SEMBLEU-1	SEMBLEU-2
IMG2AMR$_{\text{direct}}$	meta-AMRs	37.7 ± 0.2	32.6 ± 0.8	15.2 ± 0.5
IMG2AMR$_{\text{2stage}}$	meta-AMRs	38.6 ± 0.3	30.9 ± 0.4	15.6 ± 0.3
IMG2AMR$_{\text{direct}}$	caption AMRs	52.3 ± 0.4	68.6 ± 0.4	38.4 ± 0.8

Table 1: TEST results, averaged over 3 runs, for our IMG2AMR models that follow the best setting, when trained and tested on either meta-AMRs or caption AMRs.

of 10^{-5}, and a dropout rate of 0.25. Each experiment is run on one Nvidia V100-32G GPU. Model selection is done based on the best SEMBLEU-1.

5 Results

5.1 Image-to-AMR Parsing Performance

We use the standard measures of SMATCH (Cai and Knight, 2013) and SEMBLEU (Song and Gildea, 2019) to evaluate our various image-to-AMR models. SMATCH compares two AMR graphs by calculating the F1-score between the nodes and edges of these two graphs. This score is calculated after applying a one-to-one mapping of the two AMRs based on their nodes. This mapping is chosen so that it maximizes the F1-score between the two graphs. However, since finding the best exact mapping is NP-complete, a greedy hill-climbing algorithm with multiple random initializations is used to obtain this best mapping. SEMBLEU extends the BLEU (Papineni et al., 2002) metric to AMR graphs, where each AMR node is considered a unigram (used in SEMBLEU-1), and each pair of connected nodes along with their connecting edge is considered a bigram (used in SEMBLEU-2). These metrics are calculated between the model predictions and the noisy AMR ground-truth.

We report results on generating caption AMRs (when the models are trained and tested on these AMRs), as well as meta-AMRs. When evaluating on caption AMR generation, we compare the model output to the five reference AMRs, and report the maximum of these five scores. The intuition is to compare the predicted AMR to the most similar AMR from the five references. Table 1 (top two rows) shows the performance of the models on the task of generating meta-AMRs from TEST images. We perform ablations of the model input combinations on VAL set (see Section D below), and report TEST results for the best setting, which uses all the input features for both models. The 2stage model does slightly better on this task, when looking at the SMATCH and SEMBLEU-2 metrics that take the structure of AMRs into account. Note that SEMBLUE-1 only compares the nodes of the predicted and ground-truth graphs.

Meta-AMR graphs tend to, on average, be longer than individual caption AMRs ($\sim$34 vs $\sim$12 nodes and relations). We thus expect the generation of meta-AMRs to be harder than that of caption AMRs. Moreover, although we hypothesize that meta-AMRs capture a holistic meaning for an image, the caption AMRs still capture some (possibly salient) aspect of an image content, and as such are useful to predict, especially if they can be generated with higher accuracy. We thus report the performance of our direct model on generating caption AMRs (when trained on caption AMR graphs); see the final row of Table 1. We can see that, as expected, performance is much higher on generating caption AMRs vs. meta-AMRs.

Given that AMRs and natural language are by design closer in the semantic space, unlike for AMRs and images, it is not unexpected that the results for our image-to-AMR task are not comparable with those of SoTA text-to-AMR parsers, including SPRING. Our results highlight the challenges similar to those of general image-to-graph parsing techniques, including visual scene graph generation (Zhu et al., 2022), where there still exists a large gap in predictive model performance.

5.2 Image-to-AMR for Caption Generation

To better understand the quality of our generated AMRs, we use them to automatically generate sentences from caption AMRs (using an existing AMR-to-text model), and evaluate the quality of these generated sentences against the reference captions of their corresponding images. Specifically, we use the SPRING AMR-to-text model that we train from scratch on a dataset composed of AMR2.0, plus the training MSCOCO captions paired with their (automatically-generated) AMRs.

Model	BLEU-4	CIDEr	METEOR	SPICE
IMG2AMR$_{\text{direct}}$ + AMR2TXT	31.7	111.7	26.8	20.4
VL-BART*	35.1	116.6	28.7	21.5

Table 2: Image captioning results on TEST set, compared with the best reported captioning results for VL-BART.

We evaluate the quality of our AMR-generated captions using standard metrics commonly used in the image captioning community, i.e., CIDEr (Vedantam et al., 2015), METEOR (Denkowski and Lavie, 2014), BLEU-4 (Papineni et al., 2002), and SPICE (Anderson et al., 2016), and compare against VL-BART's best captioning performance as reported in the original paper (Cho et al., 2021). Reported in Table 2, the results clearly show that the quality of the generated AMRs are such that reasonably good captions can be generated from them, suggesting that AMRs can be used as intermediate representations for such downstream tasks. Future work will need to explore the possibility of further adapting the AMR formalism to the visual domain, as well as the possibility of enriching image AMRs via incorporating additional linguistic or commonsense knowledge, that could potentially result in better quality captions.

5.3 Performance per Concept Category

The analysis presented in Section 2 suggests many concepts in AMR graphs tend to be on the more abstract (less perceptual) side. We thus ask the following question: What are some of the categories that are harder to predict? To answer this question, we look into the node prediction performance of our two-stage model for the different entity and relation categories of Section 2. Note that this categorization is available for a subset of nodes only. To get the per-category recall and precision values, we take the node predictions of the first stage of the IMG2AMR$_{\text{2stage}}$ model (trained to predict meta-AMR nodes) on the VAL set. For each VAL image i, we have a set of predicted nodes, which we compare to the set of nodes in the ground-truth meta-AMR associated with the image. When calculating per-category recall/precision values, we only consider nodes that belong to that category. We calculate per-image true positive, true negative, and false positive counts, which are used to obtain the recall and precision using micro-averaging. Fig. 4 presents the per-category (as well as overall) recall and precision values over the VAL set.

Interestingly, events (e.g., *festival*, *baseball*, *tennis*) have the highest precision and recall. These are abstract concepts that are largely absent from SGs, suggesting that relying on a linguistically-motivated formalism is beneficial in capturing such abstract aspects of an image content. The event category contains 14 different types, many referring to sports that have a very distinctive setup, e.g., people wearing specific clothes, holding specific objects, etc. The possibility of encoding such abstract concepts in the training AMRs (generated from human-written descriptions likely to mention the event) helps the model learn to generate them for the relevant images during inference. The next group with high precision and recall are entities (which are likely to be more closely tied to the image regions), and possessives (containing a small number of high-frequency relations, e.g., *have* and *wear*). Semantic relations have a decent performance, but contain a diverse number of types, and need to be further analyzed to disentangle the effect of category vs. frequency.

Quantifiers (many of which are related to counting), geometric relations, and attributes seem to be particularly hard to predict. Counting is known to be hard for deep learning models. Geometric relations are much less frequent in AMRs, compared to SGs. Perhaps, we do need to rely on special features (e.g., relative position of bounding boxes) to improve performance on these relations. Attributes (such as *young*, *old*, *small*) require the model to learn subtle visual cues. In addition to understanding what input features may help improve performance on these categories, we need to further adapt the AMR formalism to the visual domain.

5.4 Qualitative Samples: Generating Descriptive Captions from meta-AMRs

In Section 5.2, we showed that caption AMRs produced by our IMG2AMR model can be used to generate reasonably good quality captions via an AMR-to-text model. Here, we provide samples of how meta-AMRs can be used as rich intermediate representations for generating descriptive captions; see Fig. 5 and Section E. To get these captions, we apply the same AMR-to-text model that we trained

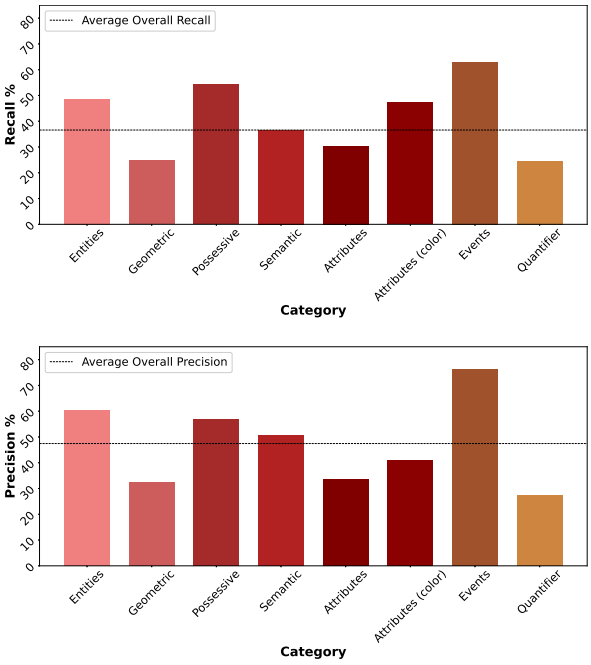


Figure 4: Node prediction performance on VAL, for the two-stage model, broken down by category.

as described in Section 5.2 to the meta-AMRs predicted by our IMG2AMR$_{\text{direct}}$ model. Captions generated from meta-AMRs tend to be longer than the original human-generated captions, and contain much more details about the scene. These captions, however, sometimes contain repetitions of the same underlying concept/relation (though using different wordings), e.g., caption (a) contains both *in grass* and *in a grassy area*. We also see that our hypernym replacement sometimes results in using a more general term in place of a more specific but more appropriate term, e.g., *woman* instead of *girl* in (d). Nonetheless, these results generally point to the usefulness of AMRs and especially meta-AMRs for scene representation and caption generation.

6 Discussion and Outlook

In this paper, we proposed to use a well-known linguistic semantic formalism, i.e., Abstract Meaning Representation (AMR) for scene understanding. We showed through extensive analysis the advantages of AMR vs. the commonly-used visual scene graphs, and proposed to re-purpose existing text-to-AMR parsers for image-to-AMR parsing. Additionally we proposed a graph transformation algorithm that merges several caption-level AMR graphs into a more descriptive meta-AMR graph. Our quantitative (intrinsic and extrinsic) and qualitative evaluations demonstrate the usefulness of (meta-)AMRs as a scene representation formalism.

Our findings point to a few exciting future research directions. Our image-to-AMR parsers can be improved by incorporating richer visual features, a better understanding of the entity and relation categories that are particularly hard to predict for our current models, as well as drawing on methods used for scene graph generation (e.g., Zellers et al., 2018; Zhu et al., 2022). Our meta-AMR generation algorithm can be further tuned to capture visually-salient information (e.g., quantifiers are too hard to learn from images, and perhaps can be dropped from a visual AMR formalism).

Our qualitative samples of captions generated from meta-AMRs show their potential for generating descriptive and/or controlled captions. Controllable image captioning has received a great deal of attention lately (e.g., Cornia et al., 2019; Chen et al., 2020, 2021). It focuses on the use of subjective control, including personalization and style-focused caption generation, as well as objective control on content (controlling what the caption is about, e.g., focused on a set of regions), or on the structure of the output sentence (e.g., controlling sentence length). We believe that by using AMRs as intermediate scene representations, we can bring together the work on these various types of control, as well as draw on the literature on controllable natural language generation (Zhang et al., 2022) for advancing research on rich caption generation.

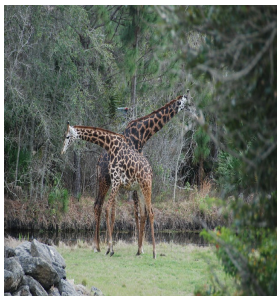
References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*.

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

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *7th Linguistic Annotation Workshop and Interoperability with Discourse*.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Association for the Advancement of Artificial Intelligence*.



(a) A couple of giraffe standing next to each other in a field near rocks walking in grass in a grassy area.

(b) A yellow and blue fire hydrant on a city street in front at an intersection sitting on the side of the road near a traffic position.

(c) A large long passenger train going across a wooden beach plate, traveling and passing by water.

(d) A woman sitting at a table eating a sandwich and holding a hot dog in a building smiling while eating.

(e) A white area filled with lots of different kinds of donuts with various toppings sitting on them.

(f) A group of people sitting around at a dining table with water posing for a picture.

(g) A person in a red jacket cross country skiing down a snow covered ski slope with a couple of people riding skis and walking on the side of the snowy mountain.

(h) A person in black shirt sitting at a table in a building with a plate of food with and smiling while having meal.

Figure 5: A sample of images, along with descriptive captions automatically generated from the meta-AMRs predicted by our IMG2AMR$_{\text{direct}}$ model. Refer to Section E for the generated meta-AMRs. The url and license information for each of these images is available in Section E. Faces were blurred for privacy.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *51st Annual Meeting of the Association for Computational Linguistics*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*.

Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. 2021. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*.

Woo Suk Choi, Yu-Jung Heo, Dharani Punithan, and Byoung-Tak Zhang. 2022. Scene graph parsing via Abstract Meaning Representation in pre-trained language models. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control, and tell: A framework for generating controllable and grounded captioning. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Vinay Damodaran, Sharanya Chakravarthy, Akshay Kumar, Anjana Umapathy, Teruko Mitamura, Yuta Nakashima, Noa Garcia, and Chenhui Chu. 2021. Understanding the role of scene graphs in visual question answering. *arXiv preprint arXiv:2101.05479*.

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

Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim, and Ramon Fernandez Astudillo. 2022. Inducing and using alignments for transition-based AMR parsing. In *North American Chapter of the Association for Computational Linguistics*.

Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. 2007. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1).

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ruohan Gao, Bo Xiong, and Kristen Grauman. 2018. Im2flow: Motion hallucination from static images for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Samitha Herath, Mehrtash Harandi, and Fatih Porikli. 2017. Going deeper into action recognition: A survey. *Image and vision computing*, 60.

Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. 2020. Scene graph reasoning for visual question answering. *arXiv preprint arXiv:2007.01072*.

Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. 2020. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*.

J. Johnson, R. Krishna, M. Stark, L. J. Li, D. A. Shamma, M. S. Bernstein, and F. F. Li. 2015a. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015b. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

A. Karpathy and L. Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yu Kong and Yun Fu. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5).

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *European Conference on Computer Vision*.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Association for Computational Linguistics*.

Xiangyang Li and Shuqiang Jiang. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, 21(8):2117–2130.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2018. Flow-grounded spatial-temporal video prediction from still images. In *Proceedings of the European Conference on Computer Vision*.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *the 27th International Conference on Computational Linguistics*.

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

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *North American Chapter of the Association for Computational Linguistics*.

Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2020a. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128(2).

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020b. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*.

Ellen M. Markman. 1990. Constraints children place on word meanings. *Cognitive Science*, 14.

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

Tahira Naseem, Austin Blodgett, Sadhana Kumaravel, Tim O'Gorman, Young-Suk Lee, Jeffrey Flanigan, Ramón Fernandez Astudillo, Radu Florian, Salim Roukos, and Nathan Schneider. 2021. DocAMR: Multi-sentence AMR representation and evaluation. In *arXiv*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Josef Ruppenhofer, Miriam R. L Petruck Michael Ellsworth, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*.

Brigit Schroeder and Subarna Tripathi. 2020. Structured query-based image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 178–179.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80.

Linfeng Song and Daniel Gildea. 2019. SemBleu: A robust metric for AMR parsing evaluation. In *57th Annual Meeting of the Association for Computational Linguistics*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1508–1517.

Qingrong Xia, Zhenghua Li, Rui Wang, and Min Zhang. 2021. Stacked AMR parsing with silver data. In *Findings of the Association for Computational Linguistics: EMNLP*.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10685–10694.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Cheng Zhang, Wei-Lun Chao, and Dong Xuan. 2019a. An empirical study on leveraging scene graphs for visual question answering. *arXiv preprint arXiv:1907.12133*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv: https://arxiv.org/abs/2201.05337*.

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019b. AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. 2020. Comprehensive image captioning via scene graph decomposition. In *European Conference on Computer Vision*, pages 211–229. Springer.

Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, et al. 2022. Scene graph generation: A comprehensive survey. *arXiv preprint arXiv:2201.00443*.

Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. 2021. End-to-end human object

interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

A AMR vs. SG: Entity and Relation Categorization Details

The analysis provided in Section 2 requires us to annotate the entities and relations of a sample of AMRs and SGs into a pre-defined set of categories. We first select all images that appear in both MSCOCO (Lin et al., 2014) and Visual Genome, so we have access to ground-truth scene graphs, as well as captions from which we can generate AMR graphs for the same set of images. We use a single AMR per image, generated from the longest caption, but include all SGs associated with an image in our analysis. For each SG and AMR graph, we consider the entities and relations corresponding to ∼900 most frequent types (around 1.3M entity and 1M relation instances for SGs; and around 130K entity and 150K relation instances for AMRs). We annotate these into a pre-defined set of entity and relation categories, including those defined by (Zellers et al., 2018) plus a few we add to cover new AMR relations. Table 5 provides a breakdown of the categories, as well as examples of word types we considered to belong to each category. The table also provides the total number of word types per category and percentages of instances across all types for each category.

Next, we describe our annotation process. SG nodes (entities) come with their most common WordNet sense annotations, which we use to identify their categories. For SG relations, we manually annotate their categories. To annotate AMR entities and relations, we follow a similar procedure, by automatically finding the most common WordNet sense for non-predicate AMR nodes (assuming most of these will be entities) and correcting them if needed. For example, the automatically-identified most common sense of *mouse* is the Animal sense, whereas in our captions, almost all instances of the word point to the computer mouse (Artifact). For any remaining concepts, including predicate nodes (e.g., *eat*, *stand*) and entities for which a category cannot be assigned automatically, we manually identify their categories.

B Distribution of AMR Node Types

Fig. 6 shows the distribution of the 90 AMR role/edge types in our training data. As we can see, keeping the top-20 types is justified given the skewed distribution of the types. Future work will need to examine the nature of the less frequent relations, and the implications of removing them from AMR graphs.

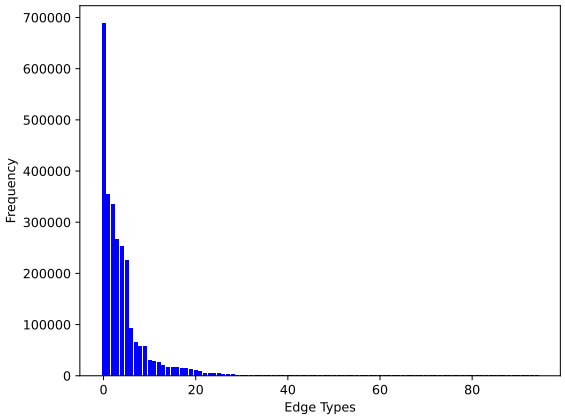


Figure 6: Frequency of the 90 AMR role/edge types prior to the refinement process, which exhibits the characteristics of a long-tail distribution.

C Meta-AMR Construction Example

Fig. 7 shows an example of how a meta-AMR is constructed from five caption-level AMRs. The corresponding captions are provided in red, and the AMR graphs are given in PENMAN notation.

D Ablations

Effect of input on node prediction performance. Table 3 presents performance of meta-AMR node prediction (first stage of IMG2AMR$_{2stage}$) with different input combinations, in terms of Precision and Recall (when predicted and ground-truth nodes are taken as sets), and BLEU-1 (when the order of nodes in the final linearized AMR is taken into consideration). These results suggest that an overall best performance is achieved by using all input features, namely regions, tags and global image feature.

{r}	{q}	{g}	Recall	Precision	BLEU-1
✓	-	-	34.5	47.1	33.1
-	✓	-	30.4	42.8	29.7
-	-	✓	30.6	39.9	29.1
✓	✓	-	35.8	**49.0**	34.3
✓	-	✓	35.1	47.5	33.9
-	✓	✓	32.9	46.5	32.1
✓	✓	✓	**36.7**	48.4	**35.6**

Table 3: VAL performance of meta-AMR node prediction (first stage of IMG2AMR$_{2stage}$) with different input combinations.

Effect of input on parsing performance. We train our IMG2AMR models with different inputs to

the encoders, and evaluate on VAL set. Specifically, the input to the model may contain the global image feature $\mathbf{g}$, region embeddings $\mathbf{r}$, tag embeddings $\mathbf{q}$ (for the first encoder), and node embeddings $\mathbf{n}$ (for the second encoder of IMG2AMR$_{\mathrm{2stage}}$). Table 4 reports the VAL results of our two models (trained and tested with meta-AMRs) with different input combinations (region embeddings, tag embeddings, global image features) for the direct model, and (node embeddings, global image features, region embeddings) for the second encoder of the $\mathrm{2stage}$ model. For IMG2AMR$_{\mathrm{2stage}}$, we fix the input of the first encoder to the best combination according to Table 3 above, and ablate over the input of the second encoder. Both models are trained and tested with meta-AMRs. As we can see, richer input generally results in better performance. We can also see a big drop in the performance of IMG2AMR$_{\mathrm{direct}}$ when only region features are used as input, suggesting that tags can help associate mappings between regions and AMR concepts.

Model Input	SMATCH	SEMBLEU-1	SEMBLEU-2
IMG2AMR$_{\mathrm{direct}}$			
$\{\mathbf{r}\}$	30.3	18.6	5.4
$\{\mathbf{r}, \mathbf{q}\}$	**39.1**	32.9	16.2
$\{\mathbf{r}, \mathbf{q}, \mathbf{g}\}$	39.0	**33.7**	**16.4**
IMG2AMR$_{\mathrm{2stage}}$			
$\{\mathbf{n}\}$	39.3	31.3	16.1
$\{\mathbf{n}, \mathbf{g}\}$	39.6	31.9	16.3
$\{\mathbf{n}, \mathbf{g}, \mathbf{r}\}$	**40.4**	**32.6**	**16.9**

Table 4: Ablation over model inputs on VAL, for both IMG2AMR models. For IMG2AMR$_{\mathrm{2stage}}$ we use all features $\{\mathbf{r}, \mathbf{q}, \mathbf{g}\}$ as the 1st encoder input.

Cateogry	Example Types per Category	#Types AMR	#Types SG	%Tokens AMR	%Tokens SG
	ENTITIES				
Artifact	clock, umbrella, bottle	128	128	22.7	24.4
Part	eyes, finger, wing	21	44	3.1	13.1
Location	beach, mountain, kitchen	86	52	20.7	11.2
Person	man, women, speaker	30	19	17.9	11
Flora/Nature	ocean, tree, flower	20	34	6.1	10.2
Clothing	dress, scarf, suit	11	31	1.1	7.7
Food	orange, donut, bread	52	23	8	2.8
Animal	horse, bird, cat	16	20	6.4	4.7
Vehicle	car, motorcycle, bicycle	18	17	6.1	4.5
Furniture	table, chair, couch	9	10	4.0	2.9
Structure	window, tower, circle	13	18	2.1	5.4
Building	brick, house, cement	6	6	1.8	2.1
	RELATIONS				
Geometric	down, edge, between	48	122	12.4	56.6
Possessive	have, wear, contain	5	42	5.9	30.6
Semantic	attempt, carry, eat	183	275	38.3	11.6
Attribute Color	color, white, blue	13	8	5.6	0.1
Attribute	young, small, colorful	82	-	12.8	-
AMR specific	and, or, date-entity	8	-	11.1	-
Quantifier	more, both, few	31	1	9.3	0.1
Event	soccer, party, festival	14	-	3.4	-
Misc	they, something, you	6	13	1.1	1.0

Table 5: The list of AMR and SG entity and relation categories, as well as examples of word types, number of types, and percentage of tokens per category.

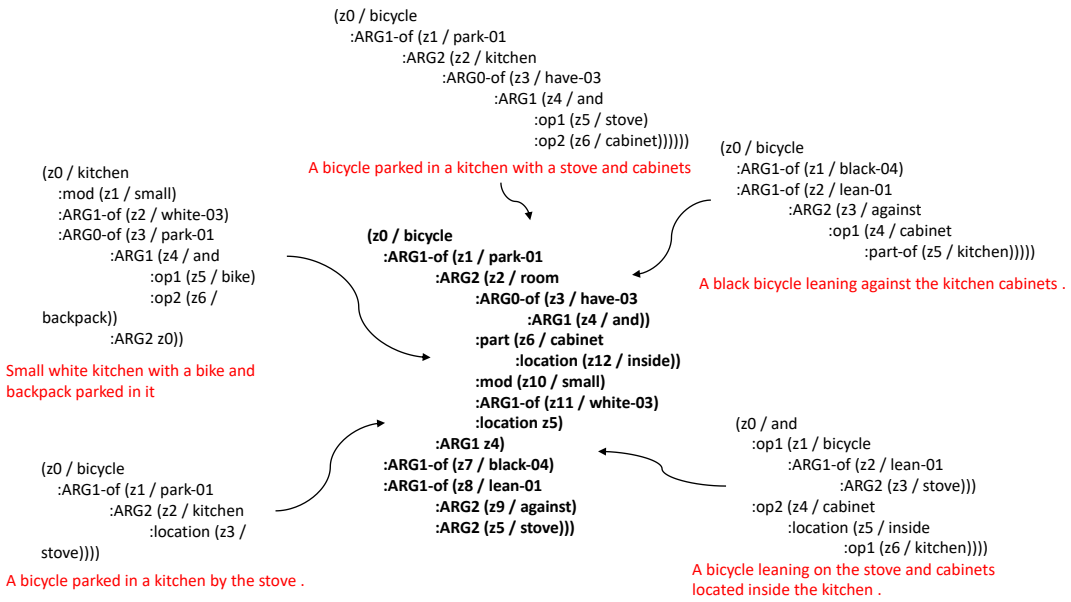
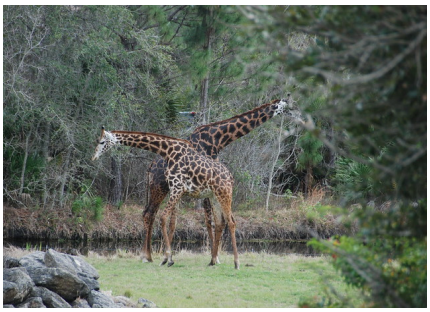


Figure 7: An example of five caption AMRs and their corresponding meta-AMR. Captions are marked as red.

E Generated AMRs for the Qualitative Samples



(a) A couple of giraffe standing next to each other in a field near rocks walking in grass in a grassy area.

(b) A yellow and blue fire hydrant on a city street in front at an intersection sitting on the side of the road near a traffic position.

(c) A large long passenger train going across a wooden beach plate, traveling and passing by water.

Figure 8: Images used in Section 5.4, along with their predicted AMRs and generated captions. Refer to Section 5.4 for more details.

(a) A woman sitting at a table eating a sandwich and holding a hot dog in a building smiling while eating.

(b) A white area filled with lots of different kinds of donuts with various toppings sitting on them.

(c) A group of people sitting around at a dining table with water posing for a picture.

Figure 9: (cont) Images used in Section 5.4, along with their predicted AMRs and generated captions. Refer to Section 5.4 for more details.

(a) A person in a red jacket cross country skiing down a snow covered ski slope with a couple of people riding skis and walking on the side of the snowy mountain.

(b) A person in black shirt sitting at a table in a building with a plate of food with and smiling while having meal.

Figure 10: (cont) Images used in Section 5.4, along with their predicted AMRs and generated captions. Refer to Section 5.4 for more details.

Images used in this section (and the rest of the paper) are under a Creative Commons Attribution License 2.0. They are available at (by the order of their appearance in this section):

- http://farm6.staticflickr.com/5299/5465041730_3fe1246cae_z.jpg and http://cocodataset.org/#explore?id=505440
- http://farm6.staticflickr.com/5294/5461489420_1e4141517b_z.jpg and http://cocodataset.org/#explore?id=332654
- http://farm4.staticflickr.com/3719/9115013219_344a42ce47_z.jpg and http://cocodataset.org/#explore?id=329486
- http://farm4.staticflickr.com/3091/3187069218_162b55b720_z.jpg and http://cocodataset.org/#explore?id=569839
- http://farm3.staticflickr.com/2020/1932016761_934411ac16_z.jpg and http://cocodataset.org/#explore?id=5754
- http://farm4.staticflickr.com/3703/10047186866_e6b43fbd32_z.jpg and http://cocodataset.org/#explore?id=298443
- http://farm8.staticflickr.com/7170/6795850593_435a36bcd9_z.jpg and http://cocodataset.org/#explore?id=239235
- http://farm4.staticflickr.com/3786/9676804086_dbb624af5c_z.jpg and http://cocodataset.org/#explore?id=386559