Commonsense Justification for Action Explanation

Shaohua Yang, Qiaozi Gao, Sari Saba-Sadiya, Joyce Y. Chai
Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI 48824
{yangshao,gaoqiaoz,sadiyasa,jchai}@cse.msu.edu

Abstract

To enable collaboration and communication between humans and agents, this paper investigates learning to acquire commonsense evidence for action justification. In particular, we have developed an approach based on the generative Conditional Variational Autoencoder (`CVAE`) that models object relations/attributes of the world as latent variables and jointly learns a performer that predicts actions and an explainer that gathers commonsense evidence to justify the action. Our empirical results have shown that, compared to a typical attention-based model, `CVAE` achieves significantly higher performance in both action prediction and justification. A human subject study further shows that the commonsense evidence gathered by `CVAE` can be communicated to humans to achieve a significantly higher common ground between humans and agents.

1 Introduction

To make AI more accessible, transparent, and trustworthy, recent years have seen an increasing effort on Explainable AI (XAI) which develops explainable models that attempts to explain the agent's decision making behaviors while maintaining a high-level of performance. Two types of explanation have been explored by the research community: *introspective explanation* which addresses the process of decision making (e.g., how a decision is made) and *justification explanation* which gathers evidence to support a certain decision (Park et al., 2018; Biran and McKeown, 2017). In this paper we focus on justification explanation, particularly identifying ***commonsense evidence*** to justify the prediction of an action. The key question we are addressing is: when an AI agent makes a prediction about an action in the world, how can the system justify its prediction that makes sense to the human?

Humans have tremendous commonsense knowledge about actions in the world (e.g., key constituents of an action) which allows them to quickly recognize and infer actions in the environment from millions of available features (Rensink, 2000). As a first step in our investigation, we initiated a human study to observe the kind of commonsense reasoning used by humans to justify the prediction of an action. From this study, we identified several dimensions of commonsense evidence which is commonly used to explain an action. Motivated by this study, we frame our task as follows: given all the symbolic descriptions of the perceived physical world (e.g., object relations and attributes as a result of vision or other processing), the goal is to identify a small set of descriptions which can justify an action prediction in line with humans' commonsense knowledge about that action. The lack of commonsense knowledge is a major bottleneck in artificial agents which jeopardizes the common ground between humans and agents for successful communication. If artificial agents ever become partners with humans in joint tasks, the ability to learn and acquire commonsense evidence for action justification is crucial.

To address this problem, we developed an approach based on the generative Conditional Variational Autoencoder (`CVAE`). This approach models the perceived attributes/relations as latent variables and jointly learns a ***performer*** which predicts actions based on attributes/relations and a ***explainer*** which selects a subset of attributes/relations as commonsense evidence to justify the action prediction. Our empirical results on a subset of the Visual Genome data (Krishna et al., 2016) have shown that, compared to a typical attention-based model, `CVAE` has a significantly higher explanation ability in terms of identifying correct commonsense evidence to justify

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2627–2637
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

the predicted action. When adding the supervision of commonsense evidence during training, both the explainability and the performance (i.e., action prediction) are further improved.

As commonsense evidence is intuitive to humans, the agent's ability to select the right kind of commonsense evidence will allow the human and the agent to come to a common understanding of actions and their justifications, in other words, *common ground*. To evaluate the role of commonsense evidence in facilitating common ground, we conducted additional human subject studies. In these experiments, the agent is given a set of images and applies our models to predict actions and select commonsense evidence to justify the prediction. For each image, the agent communicates the selected commonsense evidence to the human. The human, who does not have access to the original image, makes a guess on the action only based on the communicated evidence. The agreement between the action guessed by the human and the action predicted by the agent is used to measure how well the selected commonsense evidence serves to bring the human and the agent to a common ground of perceived actions. Our experimental results have shown that the commonsense evidence selected by CVAE leads to a significantly higher common ground.

The contributions of this paper are three folds. First we identified several key dimensions of commonsense evidence, from a human's perspective, to justify concrete actions in the physical environment. These dimensions provide a basis for justification explanation that is aligned with human's commonsense knowledge about the action. Second we proposed a method using CVAE to jointly learn to predict actions and select commonsense evidence as action justification. CVAE naturally models the generation process of both actions and commonsense evidence. Inferring commonsense evidence is equivalent to the posterior inference of the CVAE model, which is flexible and powerful by incorporating actions as context. Our experimental results have shown a higher explainability of CVAE in action justification without sacrificing performance. Finally our dataset of commonsense evidence for action explanation is available to the community[1]. It can serve as a benchmark for future work on this topic.

2 Related Work

Advanced machine learning such as deep learning approaches have shown effectiveness in many applications, however, they often lack transparency and interpretability. This makes it difficult for humans to understand the agent's capabilities and limitations. To address this problem, there is a growing interest in Explainable AI. For example, previous work has applied high-precision rules to explain classifiers' decisions (Ribeiro et al., 2016, 2018). For Convolutional Neural Networks (CNNs), recent work attempts to explain model behaviors by mining semantic meanings of filters (Zhang et al., 2017a,b) or by generating language explanations (Hendricks et al., 2016; Park et al., 2018). An increasing amount of work on the Visual Question Answering (VQA) task (Antol et al., 2015; Lu et al., 2016) has also looked into more interpretable approaches, for example, by utilizing attention-based models (Fukui et al., 2016) or reasoning based on explicit evidence (Wang et al., 2017).

Specifically for action understanding, recent work explicitly models commonsense knowledge including causal relations (Gao et al., 2016; Forbes and Choi, 2017; Zellers and Choi, 2017; Gao et al., 2018) related to concrete actions, which can facilitate action explanation. Commonsense knowledge can be acquired from image annotations (Yatskar et al., 2016) or learned from visual abstraction (Vedantam et al., 2015). Different from the above work, our work here focuses on learning to acquire commonsense evidence for action justification.

3 A Study on Justification Explanation

While there is a rich literature on explanations in Psychology, Philosophy, and Linguistics, particularly for higher-level events and decision making (Thagard, 2000; Lombrozo, 2012; Dennett, 1987), explanations for recognition of lower-level concrete physical actions (e.g., drink, brush, cook, etc.) occurred in our daily life are rarely studied. One possible reason is that we humans are so intuitive in recognizing these actions, which are often taken for granted without the need for any further explanation. However, despite recent advances, the ability to recognize and understand actions in the real world is extremely challenging for artificial agents. Thus it becomes important for the agent to have an ability to explain and justify its

[1] The dataset is available at https://github.com/yangshao/Commonsense4Action

action prediction. What can be used to justify an action prediction, and more importantly, in a human understandable way? To address this question, we initiated a human study to examine what kind of evidence humans would gather in justifying their recognition of an action perceived from the physical world.

More specifically, we selected a set of 12 short video clips (each about 14 seconds) from the Microsoft Research Video to Text dataset (Xu et al., 2016). For each video clip, we asked human subjects to explain why they think a certain action is happening in the video. The answers were collected via an online interface. A total of about 140 responses from 67 Michigan State University engineering students were collected. From the data, we identified the following categories of evidence commonly used by the subjects in their justifications. Most responses contain multiple categories of explanation.

- **Transitive-Relation**. This kind of explanation does not directly focus on the structural relations between an action and its participants, but rather transits to the relation between the participant of the action and other related evidence. For example, using *a woman wears an apron* to justify the *cook* action. In the collected responses, 64% of them used transitive relations.
- **Subaction-Relation**. Lower-level subactions are used to justify a higher-level action. For example, the action is *cook* because there are sub-actions like *cutting* and *heating meat*. Almost 75% of the responses used subactions.
- **Spatial-Relation**. Spatial relations between the participants of the action play an important role. For example, *the knife is on the cutting board* is used to explain *cooking*; and *the water is in the bottle* to explain *drinking*. Around 15% of responses are in this category.
- **Effect-Attribute**. A change in the state of an object, in other words the effect state after the action, is often used as evidence. For example, *cucumber in small pieces* is used as the evidence for *chop*. Over 28% of the responses are in this category.
- **Associated-Attribute**. Other attributes associated with the participants of the action, but not the effect state of the participants as a result of the action (20%). While these attributes are not directly related to the action, they are linked to the action by association. For example, *banana is sliced* is used as evidence to justify *blend*.
- **Other**. Participants have also cited other commonsense such as the "definition" of the action (5%), or the manner associated with different sub actions(12%).

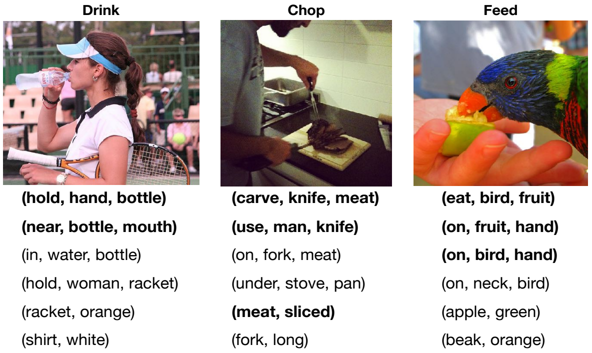


Figure 1: Examples of commonsense evidence selected by the crowd (in bold) from the list of relations and attributes.

Most of the above categories can be potentially perceived and represented through symbolic descriptions such as logic predicates to capture object attributes and relations between objects. This study has motivated us to collect additional data (Section 4) and formulate the task of commonsense justification as described in Section 5.

4 Data Collection

Motivated by the human study described above, we created a dataset based on the Visual Genome (VG) data (Krishna et al., 2016) for our investigation. Each image in the VG dataset is annotated with bounding boxes, attributes of the bounding boxes, and relations between the bounding boxes. The available annotation provides an ideal setup for us to focus on commonsense justification.

In this work, we are interested in the concrete physical actions that involve physical objects that can be perceived. We selected ten frequently occurred concrete actions: *feed, pull, ride, drink, chop, brush, fry, bake, blend, eat* and manually identified a set of images from the VG dataset depicting these actions. This has led to a dataset of 853 images with annotated ground-truth actions.

We conducted a crowd-source study to collect responses from the crowd in terms of common-

Table 1: The average number of available relations/attributes and the average number of annotated commonsense evidence relations/attributes across the corresponding images for each verb in the dataset.

	feed	**pull**	**ride**	**drink**	**chop**	**brush**	**fry**	**bake**	**blend**	**eat**
Rel#	15.49 ± 7.55	14.62 ± 9.36	12.42 ± 7.18	15.16 ± 9.89	12.00 ± 7.22	15.40 ± 8.93	14.02 ± 7.02	13.31 ± 7.27	14.37 ± 6.37	15.08 ± 6.87
Gold_Rel#	2.79 ± 1.28	1.86 ± 0.84	1.69 ± 0.83	2.41 ± 1.14	2.41 ± 1.66	2.26 ± 1.08	2.72 ± 2.06	2.25 ± 1.69	2.56 ± 1.84	2.52 ± 1.08
Att#	12.48 ± 7.11	13.60 ± 7.52	12.20 ± 7.13	10.86 ± 6.52	15.09 ± 6.82	12.31 ± 8.91	15.31 ± 7.16	13.44 ± 6.84	15.22 ± 7.18	11.98 ± 6.50
Gold_Att#	0.26 ± 0.48	0.20 ± 0.45	0.13 ± 0.40	0.30 ± 0.56	1.60 ± 1.33	0.22 ± 0.49	0.91 ± 1.26	0.93 ± 1.06	0.15 ± 0.40	0.41 ± 0.70

Table 2: Distributions of the categories of commonsense evidence relations/attributes for each verb.

	feed	**pull**	**ride**	**drink**	**chop**	**brush**	**fry**	**bake**	**blend**	**eat**
Transitive-Relation	0.10	0.14	0.15	0.11	0.11	0.13	0.12	0.18	0.15	0.09
Subaction-Relation	0.45	0.46	0.13	0.32	0.29	0.39	0.17	0.11	0.09	0.43
Spatial-Relation	0.45	0.40	0.72	0.57	0.60	0.48	0.71	0.71	0.76	0.48
Effect-Attribute	0.0	0.0	0.0	0.14	0.82	0.05	0.53	0.34	0.22	0.27
Associated-Attribute	1.0	1.0	1.0	0.86	0.18	0.95	0.47	0.66	0.78	0.73

sense evidence for action justification. As shown in Figure 1, for each image, we showed to the crowd (through Amazon Mechanical Turk) the image itself, the ground-truth action, and a list of relations/attributes. The workers were instructed to select the relations/attributes that were deemed to justify the corresponding action. We randomly assigned three workers to each image. The relations or attributes that were selected by the majority (two or more) workers were considered *gold commonsense evidence* for action justification.

Table 1 shows the average number of relations/attributes available (i.e., Rel# and Att#) for the corresponding images for each verb. It also shows the number of relations/attributes selected by the workers as commonsense evidence (i.e., Gold_Rel# and Gold_Att#). The average number of relations and attributes in each image for different actions varies slightly. However, only a small percentage of them are considered commonsense evidence. What's interesting is that the percentage of attributes considered good evidence is significantly less than the percentage of the relations. The sparsity of gold relations/attributes shows that it's a challenging task to learn an explainer for a target action.

We further inspected the selected gold commonsense relations and attributes. As shown in Table 2, they nicely fall into the categories of commonsense evidence discussed in Section 3. The ratios of Transitive-Relation are similar across different actions. The ratios of Subaction-Relation and Spatial-Relation vary for different verbs. For instance, *ride, bake, blend* tend to be justified by spatial relations more often than sub-actions. In addition, *feed, pull, ride* are rarely justified by Effect-Attribute while *chop* is mainly explained by the effect state of its direct object. These results will provide insight for generating justification explanations for a variety of verbs in the future.

5 Method

Before we formulate the problem, we will first give some formal definitions. The set of relations $\mathbf{R}$ is defined as $\{r_1, r_2, ..., r_m\}$ where each r_i is a tuple (r_i^p, r_i^s, r_i^o) corresponding to the *predicate*, *subject*, and *object*; and the set of attributes $\mathbf{E}$ is represented as $\{e_1, e_2, ..., e_n\}$ where each e_i is a tuple (e_i^o, e_i^p) corresponding to the *object* and *attribute*. We introduce $\mathbf{z}$ as a discrete vector $(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{m+n})$ where $\mathbf{z}_i \in \{0, 1\}$ represents the hidden explainable variable. $\mathbf{z}$ is interpreted as an evidence selector: $\mathbf{z}_i = 1$ means the corresponding relation/attribute justifies the target action $\mathbf{a}$. We define $\mathbf{A}$ as the vocabulary of target actions. Based on all these definitions, our goal is to jointly select evidence $\mathbf{z}$ and predict target action $\mathbf{a} \in \mathbf{A}$. In other words, to learn the probability $\mathbf{p}(\mathbf{a}, \mathbf{z}|\mathbf{R}, \mathbf{E})$.

5.1 Conditional Variational Autoencoder

The varational autoencoder(VAE) (Kingma and Welling, 2013) is proposed as a generative model to combine the power of both directed continuous

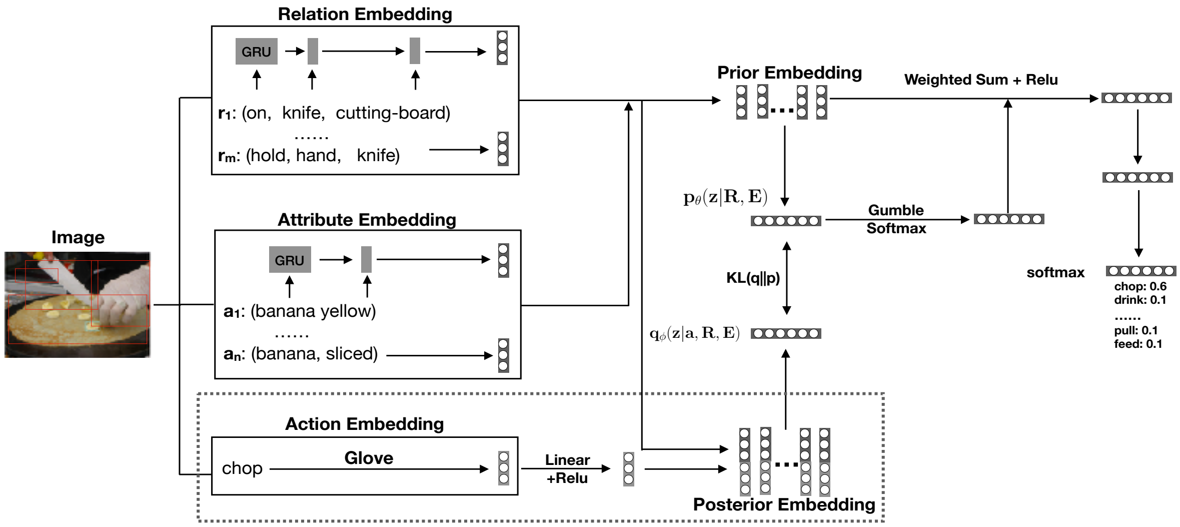


Figure 2: System architecture for the CVAE model. The dotted region is only used during the model training process.

or discrete graphical models and neural network with latent variables. The VAE models the generative process of a random variable $\mathbf{x}$ as following: first the latent variable $\mathbf{z}$ is generated from a prior probability distribution $\mathbf{p}(\mathbf{z})$, then a data sample $\mathbf{x}$ is generated from a conditional probability distribution $\mathbf{p}(\mathbf{x}|\mathbf{z})$. The CVAE (Zhao et al., 2017) is a natural extension of VAE: Both the prior distribution and conditional distribution now are conditioned on an additional context $\mathbf{c}$: $\mathbf{p}(\mathbf{z}|\mathbf{c})$ and $\mathbf{p}(\mathbf{x}|\mathbf{z},\mathbf{c})$.

In our task, we decompose the inference problem $\mathbf{p}(\mathbf{a},\mathbf{z}|\mathbf{R},\mathbf{E})$ into two smaller problems. The first sub-problem is to infer $\mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E})$, which is a **performer**. The second problem is to infer $\mathbf{p}(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})$ which is an **explainer**. These two problems are closely coupled, hence we model them jointly. The probability distribution $\mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E})$ can be written as :

$$\mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E}) = \sum_{\mathbf{z}} \mathbf{p}_\theta(\mathbf{a}|\mathbf{z},\mathbf{R},\mathbf{E})\mathbf{p}(\mathbf{z}|\mathbf{R},\mathbf{E})$$

Directly optimizing this conditional probability is not feasible. Usually the Evidence Lower Bound (ELBO) (Sohn et al., 2015) is optimized, which can be derived as the following:

$$\begin{aligned}
&\mathrm{ELBO}(\mathbf{a},\mathbf{R},\mathbf{E};\theta,\phi) \\
&= -\mathrm{KL}(\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})||\mathbf{p}_\theta(\mathbf{z}|\mathbf{R},\mathbf{E})) \\
&+ \mathbf{E}_{\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E})}[\log p_\theta(\mathbf{a}|\mathbf{z},\mathbf{R},\mathbf{E})] \\
&\leq \log \mathbf{p}(\mathbf{a}|\mathbf{R},\mathbf{E})
\end{aligned} \tag{1}$$

The first KL divergence term is to minimize the distance between the posterior distribution and the prior distribution. The second term is to maximize the expectation of the target action based on the posterior latent distribution.

In most previous work using VAE, there is no explicit meaning for the hidden representation $\mathbf{z}$, thus it's hard for humans to interpret. For example, $\mathbf{z}$ is simply assumed as a Gaussian distribution or a categorical distribution. In order to have a more explicit representation for the purpose of explanation, our latent discrete variable $\mathbf{z}$ is used to indicate whether the corresponding relation or attribute can be used for justifying the action.

The whole system architecture is shown in Figure 2. From an image, we first extract a candidate relation set $\mathbf{R}$ and an attribute set $\mathbf{E}$. Every relation $\mathbf{r}$ and attribute $\mathbf{e}$ are embedded using a Gated Recurrent Neural Network (Chung et al., 2014).

$$\mathbf{r}^{emb} = \mathrm{GRU}([r^p, r^s, r^o])$$

$$\mathbf{e}^{emb} = \mathrm{GRU}([e^o, e^p])$$

The action $\mathbf{a}$ is represented by a GloVe embedding (Pennington et al., 2014), followed by another non-linear layer:

$$\mathbf{a}^{emb} = \mathrm{ReLU}(\mathbf{W}_i\mathbf{a}^{glove} + \mathbf{b}_i)$$

where $\mathbf{a}^{glove} \in \mathbb{R}^k$ is the pre-trained GloVe embedding. Then the latent variable $\mathbf{z}$ can be calculated as:

$$\mathbf{q}_\phi(\mathbf{z}|\mathbf{a},\mathbf{R},\mathbf{E}) = \mathrm{softmax}(\mathbf{W}_\mathbf{z}[\mathbf{U};\mathbf{a}^{emb}] + \mathbf{b}_\mathbf{z})$$

where $\mathbf{U} = [\mathbf{r}_1^{emb}, ..., \mathbf{r}_m^{emb}, \mathbf{e}_1^{emb}, ..., \mathbf{e}_n^{emb}]$ and $[\mathbf{U}, \mathbf{a}^{emb}]$ means the concatenation of $\mathbf{U}$ and $\mathbf{a}^{emb}$. and $W_{\mathbf{z}} \in \mathbb{R}^{2\times 2k}$ as we assume each $\mathbf{z}_i$ belongs to one of the two classes $\{0, 1\}$.

The prior distribution can be calculated as:

$$\mathbf{p}_\theta(\mathbf{z}|\mathbf{R}, \mathbf{E}) = \text{softmax}(\mathbf{W}_{\mathbf{z}}^{'}\mathbf{U} + \mathbf{b}_{\mathbf{z}}^{'})$$

The KL divergence between the prior random variable $\mathbf{z}_{prior}$ from $\mathbf{p}_\theta(\mathbf{z}|\mathbf{R}, \mathbf{E})$ and the posterior random variable $\mathbf{z}_{posterior}$ from $\mathbf{q}_\phi(\mathbf{z}|\mathbf{a}, \mathbf{R}, \mathbf{E})$ is:

$$\text{KL}(\mathbf{z}_{prior}, \mathbf{z}_{posterior}) = -p_i \log \frac{p_i}{p_i^{'}} - (1 - p_i) \log \frac{1 - p_i}{1 - p_i^{'}}$$

here $\mathbf{z}_{prior} \sim \text{Bern}\,(p_i)$, $\mathbf{z}_{posterior} \sim \text{Bern}\left(p_i^{'}\right)$.

Another challenge is that $\mathbf{z}$ is a discrete variable which blocks the gradient and makes the end-to-end training infeasible. Gumbel-Softmax (Jang et al., 2016) is a re-parameterization trick to deal with the discrete variables in the neural network. We use this trick to sample discrete $\mathbf{z}$. Then we do a weighted sum pooling between discretized $\mathbf{z}$ and $\mathbf{U}$:

$$\mathbf{h}_z = \text{ReLU}(\sum_i \mathbf{z}_i * \mathbf{U}_i)$$

$$\mathbf{h} = \text{ReLU}(\mathbf{W}_h \mathbf{h}_z + \mathbf{b}_h)$$

$$\mathbf{p}_\theta(\mathbf{a}|\mathbf{z}, \mathbf{R}, \mathbf{E}) = \text{softmax}(\mathbf{W}\mathbf{h} + \mathbf{b})$$

During training, we also add a sparsity regularization on the latent variable $\mathbf{z}$ besides the ELBO. So our final training objective is

$$\begin{aligned} \mathcal{L}_{CVAE} = &- \text{ELBO}(\mathbf{a}, \mathbf{R}, \mathbf{E}; \theta, \phi) \\ &+ \beta\, \text{KL}(\mathbf{q}_\phi(\mathbf{z}|\mathbf{a}, \mathbf{R}, \mathbf{E}) || \,\text{Bern}(0)) \end{aligned} \quad (2)$$

During testing, we have two objectives. First we want to infer the target action $\mathbf{a}$, which can be computed through sampling:

$$\begin{aligned} \mathbf{p}(\mathbf{a}|\mathbf{R}, \mathbf{E}) &= \sum_{\mathbf{z}} \mathbf{p}_\theta(\mathbf{z}|\mathbf{R}, \mathbf{E}) \mathbf{p}_\theta(\mathbf{a}|\mathbf{z}, \mathbf{R}, \mathbf{E}) \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \mathbf{p}_\theta(\mathbf{a}|\mathbf{z}_s, \mathbf{R}, \mathbf{E}) \end{aligned} \quad (3)$$

where $\mathbf{z}_s \sim \mathbf{p}(\mathbf{z}|\mathbf{R}, \mathbf{E})$ and S is the number of samples. After obtaining the predicted action $\hat{\mathbf{a}}$, the posterior explanation is inferred as $\mathbf{q}_\phi(\mathbf{z}|\hat{\mathbf{a}}, \mathbf{R}, \mathbf{E})$.

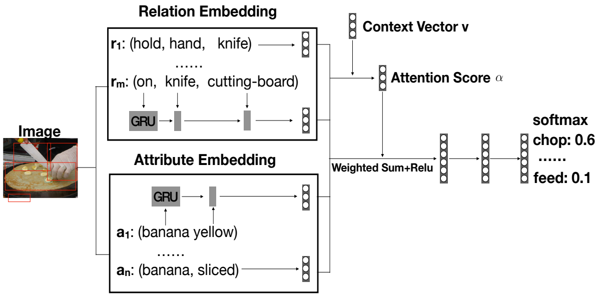


Figure 3: The system architecture for attention-based method.

5.2 Conditional Variational Autoencoder with Supervision (CVAE+SV)

In this setting, we assume we have the supervision for the discrete latent variable z, which is more like a multi-task setting. We optimize both the action prediction loss and the evidence selection loss. The final loss function is defined as:

$$\mathcal{L}_{SV} = \lambda \mathcal{L}_{CVAE} + (1 - \lambda)\mathcal{L}_{evidence}$$

where

$$\mathcal{L}_{evidence} = -\sum_k (\mathbf{z}_k \log \mathbf{p}(\hat{\mathbf{z}}_k) + (1 - \mathbf{z}_k) \log(1 - \mathbf{p}(\hat{\mathbf{z}}_k)))$$

in which $\mathbf{z}_k \in \{0, 1\}$ is the ground truth label, $\hat{\mathbf{z}}_k$ is the predicted label and λ is a hyper-parameter.

6 Evaluation on Action Explanation

To evaluate our model, we randomly split our dataset (853 images) into 60% for training, 20% for validation, and 20% for test. For all the models we use the Adam optimizer (Kingma and Ba, 2014) with a starting learning rate 1e-4. All other hyperparameters are tuned on the validation set.

6.1 Baseline: Attention Model

We use an attention-based model as a baseline, which is similar to the model originally proposed for document classification (Yang et al., 2016). The architecture is shown in Figure 3. Different from the CVAE-based method, this model directly learns a context parameter instead of learning from the posterior action context. The attention is calculated as:

$$\alpha_i = \frac{\exp(\mathbf{u}_i^T \mathbf{v})}{\sum_j \exp(\mathbf{u}_i^T \mathbf{v})}$$

where $\mathbf{v}$ is the context parameter, and $\mathbf{u}_i$ is the GRU embedding of the corresponding relation/attribute. The learned attention weights are used for the selection of commonsense evidence.

Table 3: Action prediction accuracy and evidence selection MAP.

	Action Accuracy	Evidence MAP
Attention	0.789	0.442
CVAE	0.835	0.572
CVAE+SV	0.871	0.690
Upper Bound	0.918	1.0

6.2 Evaluation Metrics and Comparison

Our evaluation compares the performance from the following models:

- `Baseline`. The attention model presented in Section 6.1.
- `CVAE`. The conditional variational autoencoder model presented in Section 5.1.
- `CVAE+SV`. The `CVAE` model with supervision as presented in Section 5.2.
- `Upper Bound`. We also calculate the upper bound of the `CVAE` model using the human annotated gold evidence.

For each of the above model, evaluate model performance on both action prediction (i.e., performer) and action justification (i.e., explainer)

- **Performer**: Accuracy is used to measure the percentage of actions that are correctly predicted by the model.
- **Explainer**: As discussed in Section 5, the binary random variable **z** is used to capture commonsense evidence. The probability of each **z** represents the model's belief that the corresponding evidence supports the action decision. As we hope to rank the gold evidence higher, the Mean Average Precision (MAP) metric is calculated for evaluating evidence selection.

6.3 Evaluation Results

The results are shown in Table 3. Since the `Upper Bound` method directly uses the human annotated gold evidence, its MAP for selecting evidence is always 1.0.

The `CVAE` model outperforms the attention-based model in both action prediction and evidence selection. This indicates that the `CVAE` model can incorporate a better guidance for evidence selection during the training process. One possible explanation is that the CVAE model incorporates the target action as the context during learning instead of directly learning a context parameter. Furthermore, after adding the evidence supervision, the `CVAE+SV` model gives even better performance in both action prediction and evidence selection. We notice that for the `CVAE+SV` model, its action prediction accuracy is approaching the upper bound 91.8%, however the evidence selection MAP is still far from the upper bound even with supervision.

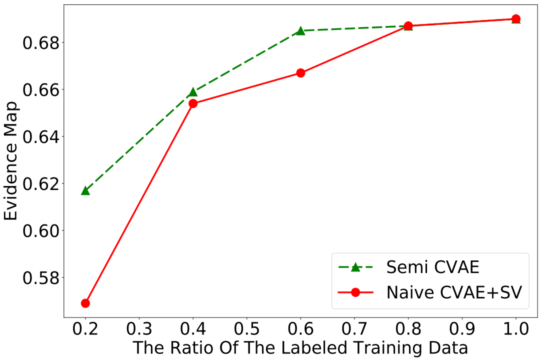


Figure 4: Evidence selection MAP for semi-supervised learning.

6.4 Semi-Supervised Learning

Although we have shown that adding supervision on the latent variable **z** improves the model performance, collecting this label information through human annotation is usually time consuming and expensive. In this section, we explore how semi-supervised learning can help to alleviate this difficulty.

As a generative model, `VAE` has shown its advantage on semi-supervised learning (Kingma et al., 2014). Following the method in (Kingma et al., 2014), our semi-supervised learning loss function is defined as:

$$\mathcal{L} = \sum_{(\mathbf{a},\mathbf{R},\mathbf{E},\mathbf{z})\sim\mathbf{p}_l} \mathcal{L}_{SV} + \sum_{(\mathbf{a},\mathbf{R},\mathbf{E})\sim\mathbf{p}_u} \mathcal{L}_{CVAE}$$

where $\mathcal{L}_{SV}$ is defined in section 5.2 and $\mathcal{L}_{CVAE}$ is detailed in section 5.1. In other words, the data sample with evidence label is fed to $\mathcal{L}_{SV}$, otherwise is fed to $\mathcal{L}_{CVAE}$.

The results are shown in Figure 4 where the x-axis shows the ratio of labeled examples. The incremental `Naive CVAE+SV` model only uses the labeled evidence examples while the `Semi CVAE` model also uses unlabeled evidence examples. The figure shows that the `Semi CVAE`

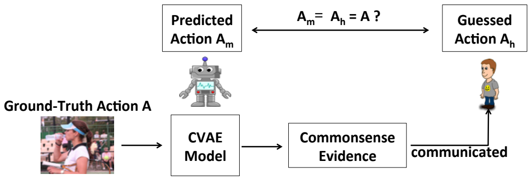


Figure 5: The experimental setup for the human subject study examining the role of commonsense justification towards common ground.

model outperforms the `Naive CVAE+SV` model. This indicates that the semi-supervised method can improve the evidence selection by making use of unlabeled examples.

7 Commonsense Justification towards Common Ground

In human-agent communication, the success of communication is largely dependent on common ground which captures shared knowledge, beliefs, or past experience (Clark, 1996). As commonsense evidence what humans use to justify actions, To validate this hypothesis, we conducted a human-subject experiment to examine the role of commonsense justification in facilitating common ground.

7.1 Experiment Setup

Figure 5 shows the setup of our experiment. The agent is provided with an image and applies various models (e.g., `CVAE`) to jointly predict the action and identify commonsense evidence. The human is provided with a list of six action choices and does not have access to the image. The agent communicates to the human only the identified commonsense evidence and the human makes a guess on the action from the candidate list purely based on the communicated evidence. The idea is that, if the human and the agent share the same beliefs about evidence to justify an action, then the action guessed by the human should be the same as the action predicted by the agent.

Generating Distracting Verbs. For each image, the human is provided with a list of six action/verb candidates. To generate this list, we mix four distracting verbs with the ground-truth action verb plus a default `Other`. Most of the distracting verbs come from the concrete action verbs made available by (Gao et al., 2018). We first manually filtered out the verbs which have the same meaning with the ground-truth verb. We then selected two groups of distracting verbs: an *easy* group (where the distracting verbs have larger distance from the ground-truth verb in the embedding space, with an average similarity of 0.284) and a *hard* group (more close to the ground-truth verbs with an average similarity of 0.479). The temperature based softmax distribution (Chorowski and Jaitly, 2016) was used to sample the easy and the hard distracting verbs based on the pre-trained GloVe (Pennington et al., 2014) embedding cosine similarity.

Process. A total of 170 images were used in this experiment, and 24 workers from AMT participated in our study. For each image, we applied three different models: `Attention` baseline, `CVAE`, and `CVAE+SV` to generate the commonsense evidence. An upper bound based on gold commonsense evidence was also measured. Note that, the agent has no knowledge of the human's action choices when generating the commonsense evidence. Theory of mind is an important aspect in human-agent communication. Incorporating human's action choices in justifying action is an interesting however a different problem which requires different solutions. In this paper, we only focus on the situation where the mind of the human is opaque to the agent.

For each model and each image under the easy or hard configurations, the top five predicted commonsense evidence (associated with the predicted action) were shown to a worker. The the worker was requested to select the most probable action from the distracting list only based on these five pieces of evidence. We randomly assigned three workers to each image. The majority of three selections was considered as the final answer. If all three selections disagreed, one worker's choice was randomly selected as the final answer.

Metrics for Common Ground. We use the agreement between the action guessed by the human and the action predicted by the agent to measure how well the selected commonsense evidence serves to bring the human and the agent to a common ground of perceived actions. More formally, as shown in Figure 5, given an image, suppose its ground-truth action is A, the action predicted by the agent/machine is A_m, and the action guessed by the human is A_h, the *Common Ground* is defined as: $A_m = A_h = A$. Here we also enforce that the predicted action should be the same as the ground-truth action. The percentage of trials based on different models that have led to a com-

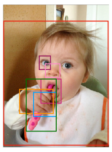
	Attention	CVAE	CVAE+SV	Gold
Gold Action: Bake	A_m: Eat A_h: Bake	A_m: Bake A_h: Bake	A_m: Bake A_h: Bake	A_m: Bake A_h: Bake
	• The bread is next to the bread. • The bread is on the rack. • The bread is on the pan. • The man has keys. • The man has the band.	• The bread is on the rack. • The bread is on the pan. • The bread is on the tray. • The bread is next to the bread. • The bread is baked.	• The bread is baked. • The bread is next to the bread. • The person is pushing the tray. • The bread is on the pan. • The bread is on the rack.	• The bread is on the tray. • The person is pushing the tray. • The bread is baked.
Gold Action: Brush	A_m: Brush A_h: Skin	A_m: Brush A_h: Brush	A_m: Brush A_h: Brush	A_m: Brush A_h: Brush
	• The baby has a mouth. • The baby has a hand. • The baby has eyeballs. • The baby has fingers. • The baby has a nose.	• The hand holds the toothbrush. • The toothbrush is in the mouth. • The baby has a mouth. • The baby has fingers. • The baby has a nose.	• The hand holds the toothbrush. • The toothbrush is in the mouth. • The baby has eyeballs. • The baby has a mouth. • The baby has a hand.	• The toothbrush is in the mouth. • The hand holds the toothbrush.

Figure 6: Two examples of the common ground study based on different models. In each example, a ranked list of commonsense evidence generated by different models is shown. A_m captures the action predicted by the agent. A_h captures the action guessed by the human based on the selected commonsense evidence.

Table 4: Results from the human subject study on common ground.

	Attenton	CVAE	CVAE+SV	Gold
Easy	0.665	0.776	0.818	0.888
Hard	0.576	0.718	0.788	0.841

mon ground is measured and compared.

7.2 Experimental Results

Table 4 shows the comparison results among various models and the upper bound where the gold commonsense evidence provided to the human. It's not surprising that performance on common ground is worse in the *hard* configuration as the distracting verbs are more similar to the target action. The CVAE-based method is better than the attention-based method in facilitating common ground.

Figure 6 shows two examples of the top five predicted evidence under different models. For each model, it also shows the agent predicted action (A_m) and the human guessed action (A_h). In both examples, all models were able to establish a common ground except for the attention-based model. The evidence selected by the CVAE+SV model is clearly more accurate than the CVAE model and is more close to the ground-truth evidence. The second example shows that although the attention-based model predicts a correct target action, it fails to convey correct commonsense evidence to establish a common ground with the human.

8 Conclusion

This paper describes an approach for action justification using commonsense evidence. As demonstrated in our experiments, commonsense evidence is selected to align with humans' justification of an action and is therefore critical in establishing a common ground between humans and agents.

For all experiments in this paper, we use the annotated relations/attributes from the original Visual Genome data. As the state-of-the-art recall@50 on the relation detection with a limited vocabulary is only around 20% (Liang et al., 2018). Using annotated relations and attributes allows us to focus on the study of commonsense evidence and its role in action justification and common ground. Nevertheless, our proposed method has the potential to handle the erroneous relations/entities, e.g., as a result of vision processing, for example, by avoiding to select erroneous relations as they do not correlate with actions and other indicative relations/attributes. Our future work will extend the model and findings from this work to vision processing that will not only identify commonsense evidence but also explain where and how in the perceived environment the evidence is gathered.

Acknowledgement

This work was supported by IIS-1617682 from the National Science Foundation and the DARPA XAI program under a subcontract from UCLA (N66001-17-2-4029).

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Or Biran and Kathleen McKeown. 2017. Human-centric justification of machine learning predictions. *IJCAI, Melbourne, Australia*.

Jan Chorowski and Navdeep Jaitly. 2016. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Herbert H Clark. 1996. Using language. 1996. *Cambridge University Press: Cambridge*, 952:274–296.

D. Dennett. 1987. *The intentional Stance*. MIT Press.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. *arXiv preprint arXiv:1706.03799*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Chai. 2016. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1814–1824.

Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. 2018. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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.

Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. 2018. Visual relationship detection with deep structural ranking. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.

Tania Lombrozo. 2012. Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1802.08129*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ronald A Rensink. 2000. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491.

Thagard. 2000. Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly*.

Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Learning common sense through visual abstraction. In *Proceedings of the IEEE international conference on computer vision*, pages 2542–2550.

Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. 2017. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *Proc. CVPR*.

J. Xu, T. Mei, T. Yao, and Y. Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198.

Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. *arXiv preprint arXiv:1707.09468*.

Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. 2017a. Interpreting cnn knowledge via an explanatory graph. *arXiv preprint arXiv:1708.01785*.

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2017b. Interpretable convolutional neural networks. *arXiv preprint arXiv:1710.00935*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.