Visual Relation Detection Using Hybrid Analogical Learning

Kezhen Chen and Ken Forbus
Northwestern University, Evanston, IL
kzchen@u.northwestern.edu, forbus@northwestern.edu

Abstract

Visual Relation Detection is currently one of the most popular problems for visual understanding. Many deep-learning models are designed for relation detection on images and have achieved impressive results. However, deep-learning models have several serious problems, including poor training-efficiency and lack of understandability. Psychologists have ample evidence that analogy is central in human learning and reasoning, including visual reasoning. This paper introduces a new hybrid system for visual relation detection combining deep-learning models and analogical generalization. Object bounding boxes and masks are detected using deep-learning models and analogical generalization over qualitative representations is used for visual relation detection between object pairs. Experiments on the Visual Relation Detection dataset indicates that our hybrid system gets comparable results on the task and is more training-efficient and explainable than pure deep-learning models.

Introduction

When performing visual understanding, people tend to encode relations between pairs of recognized objects. Visual relation detection is thus an important task for artificial intelligence and computer vision. Given an image, the goal of this task is to visually detect objects and predict relational predicates between them. Deep-learning models have achieved impressive results, mostly using two stages for this task: object detection followed by pairwise relation recognition. Most of the deep-learning models operate end-to-end, combining the two stages into one single neural network. They have broad coverage on the input data, especially on visual data. However, given the single neural network, models have some serious problems such as low training-efficiency and lack of understandability. For example, (Ciresan et al., 2011) uses data-augmentation to increase the training samples to learn a classification model with many epochs on MNIST dataset. Many researchers have argued that neural networks are hard to trust because they lack explainability (Samek, Wiegand, and Muller, 2017; Buhrmester, Munch, and Arens, 2019). This paper introduces a new hybrid system that combines deep learning models and analogical generalization to get the best features of both: the broad coverage of deep learning models and the high training-efficiency and explainability of probabilistic symbolic representations produced by analogical generalization.

Cognitive psychology provides evidence that analogy plays important roles in human vision (Sagi et al., 2012; Anderson et al., 2018). Moreover, computational models of analogy, combined with computational models of high-level vision using relational representations, have been successful in modeling a variety of psychological phenomena. For example, (Kandaswamy et al., 2014) showed that the Structure Mapping Engine (SME), an analogical matching model, can model learning forced-choice tasks with visual stimuli. Similarly, (Lovett and Forbus, 2013) showed that CogSketch [1](Forbus et al., 2011), a model of human visual perception that relies on analogical matching and relational structure, can solve mental rotation and paper folding tasks using SME, and performs better than most adult Americans on Raven's Progressive Matrices (Lovett and Forbus, 2017). These models have been used in performance-oriented systems as well: Analogical generalization has been used to perform visual tasks such as object recognition (Chen et al., 2019) and link plausibility (Liang & Forbus, 2015) to perform training-efficient learning with explainability.

Given an image, we use off-the-shelf deep-learning models to detect object categories, object bounding boxes and object masks. Object bounding boxes provide positional and spatial information. Object masks show the pose information of objects. For most object pairs, the visual semantic relations can be recognized using the positional information, pose information and category information. Therefore, we encode these qualitative representations for each object pair. The qualitative representations of object pairs are passed to analogical generalization to learn and classify the possible

[1] CogSketch can be downloaded at https://www.qrg.northwestern.edu/software/software_index.html.

This research was supported by the Office of Naval Research

visual relations between them. Our contributions are as follows. (1) We propose a new hybrid system that combines deep learning models and analogical generalization on the visual relation detection task. To our knowledge, this is the first system that combines deep learning and analogical learning. (2) We create a novel qualitative representation scheme for encoding pairwise information between pairs of objects. (3) Experiments and analysis show that our system has comparable results with several orders of magnitude fewer examples than pure deep-learning systems require, and the learned generalizations provide explainable models.

Related Work

Visual Relation Detection

In recent years, many approaches have been proposed for visual relation detection and scene graph parsing. The visual relations between two objects are written as a triple < Subject, Relation, Object>, where the detected objects can either be the Object or the Subject in the triple. Most systems use a two-stage pipeline that detect objects from the image first and then classifies the relations between objects. In the first stage, almost all methods use deep learning detectors, either off-the-shelf detectors (Lu et al., 2016; Zhuang et al., 2017; Dai et al., 2017) or fine-tuning with the relationship datasets (Li et al., 2017; Xu et al., 2017). In the second stage, the visual relation task is usually regarded as a classification task and a relation is predicted for each pair. Various architectures have been used for encoding and classification. For example, (Zellers et al., 2018) uses stacks of LSTMs to encode the features of object pairs. (Yang et al., 2018) uses an attentional graph convolutional network for encoding the context information of objects. (Tang et al., 2018) applied tree structure encoding and decoding for scene graph generation.

However, these deep-learning models require multiple epochs for training and produce results that are hard to understand. Our approach combines the deep learning models and analogical learning, to improve understandability and training cost. Following the two-stage scheme, our approach uses analogical generalization to classify relations over symbolic qualitative representations for each pair of objects in the second stage. Thus, the learned generalizations for each type of relation can be easily explored and explained and in the second stage, our approach requires fewer examples for training, more like humans.

Analogical Learning

Analogy is one of the essential capabilities in human learning (Gentner 2003). Analogical learning involves three core processes: analogical matching, analogical retrieval, and analogical generalization. This *analogy stack*, described in more detail below, provides analogical learning capabilities that have been used in multiple tasks, including visual tasks (Chen et al., 2019; Chen and Forbus, 2018). However, the automatic visual encoding processes used in prior experiments lacked breadth: They handled simple images, as found in visual problem-solving tasks, or Kinect data. Here we use deep learning to do some of the initial encoding, in a way that takes advantage of its broad coverage, while maintaining the training efficiency and explainability of analogical learning.

Approach

Our system uses a two-stage pipeline of object detection followed by pairwise relation detection. The overview of the pipeline is depicted in Figure 1. Given an image, object detection uses deep learning models to detect bounding boxes, instance segmentations and their categories of objects.

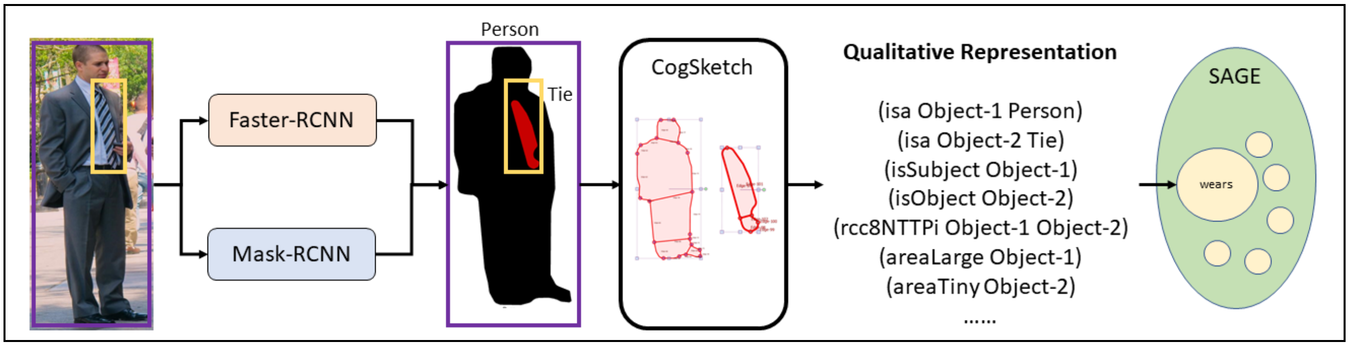


Figure 1: The pipeline overview to learn visual relation classification. Given an image, our system uses Faster-RCNN and Mask-RCNN to detect bounding boxes, instance segmentations, and categories of objects. Then, detection results are imported into CogSketch. CogSketch computes qualitative representations for each pair of objects based on a novel encoding scheme, including object encodings and pair encodings. Generated qualitative representations are added into targeted generalization pools in our analogical generalization model, SAGE. SAGE learns probabilistic relational generalizations, which are used for classification.

Object bounding boxes provide positional information, object masks give pose information and object categories provide the semantic category of objects. The second stage automatically computes qualitative visual representations based on our novel encoding scheme for pairs of objects. During training, analogical generalization is used to learn analogical models for the semantic relations, and during testing, analogical retrieval is used to classify the relationship for each pair of objects. We discuss each stage in turn next.

Object Detection

In the first stage, we need to detect objects from images. To classify the visual relations in the second stage, the bounding box, mask, and category of each object are generated to provide enough information for encoding object pairs. We use Faster-RCNN (Ren et al., 2015) with VGG16 backbone to detect the object bounding boxes and categories. Faster-RCNN has two modules for object detection. The first module is a deep fully convolutional neural network that proposes regions, and the second module is the Faster-RCNN detector that uses regions for object detection. Given an image, Faster-RCNN generates a set of bounding boxes. Each box has an object prediction label and confidence score. The bounding boxes are filtered with a threshold to generate the final set. In our experiments, Faster-RCNN is pre-trained on COCO dataset (Lin et al., 2014) and trained on targeted visual relation datasets.

To detect object masks, we utilize Mask-RCNN model (He et al., 2018) with Resnet-50 as the backbone for instance segmentation. Mask-RCNN also has two modules. The first module is similar to Faster-RCNN, proposing regions for objects. The second module uses proposed regions to generate instance segmentation masks for objects. As most of the visual relation datasets do not have instance segmentation annotations, we directly detect object masks using a Mask-RCNN trained on COCO dataset.

During testing, Faster-RCNN detects the object bounding boxes and their categories. Mask-RCNN detects the object segmentations and the bounding boxes of corresponding segments. For each detected object from Faster-RCNN, we computed the intersection of union (IoU) between the bounding boxes from Faster-RCNN and Mask-RCNN. If the IoU is larger than 0.7, the corresponding instance segments from Mask-RCNN are assigned to the object. Otherwise, we regard the bounding boxes as the masks of the objects. In the next stage, we describe how the detection results are used for pairwise relation detection.

Pairwise Relation Detection

During training, we use the ground truth triples, i.e. <Object, Relation, Subject>, as training data for analogical generalization. Given the detected bounding boxes, categories, and segmentations of the Object and Subject of the triple, our system is trained to learn analogical models of the Relation, based on automatically constructed qualitative visual representation. During inference, analogical retrieval is used to identify the relation with highest probability for each pair of entities detected from detection stage. To model the information between two detected objects, we first describe our visual encoding scheme for generating the symbolic representations. Then we describe how analogical learning performs relation detection training. Finally, the inference process is presented.

Symbolic representations for object pairs: For each pair of objects, we build symbolic representations to represent the details of each object and spatial information between them. We use the off-the-shelf CogSketch system (Forbus et al. 2011) to help compute spatial information and relational predicates. CogSketch is a model of human visual perception that relies on analogical matching and relational structure. CogSketch uses NextKB[2], an off-the-shelf open-source broad coverage knowledge base. NextKB includes a large-scale lexicon that maps words to OpenCyc and FrameNet concepts. We use this lexicon to map labels produced by deep learning modules into concepts.

We divide the symbolic representations of object pairs into two parts: *Obj-reps and Spa-reps*. Obj-reps consists of facts for each object generated from their masks and categories. Spa-reps consist of the spatial relations between the pair of objects generated from their bounding boxes.

In Obj-reps, object category and pose are encoded. The semantic category is produced by using lexical lookup from the word produced as the category label to map that into the

Attribute	Description	Example
Cross-sectional curvature	Whether the edges of the shape are straight or curved.	(allCurved EC-1)
Edge concavity	Whether the shape has concave edges.	(hasConcavedEdge EC-1)
Shape Estimation	The simple geometric shape that is closest to the shape.	(ellipseSystemShape EC-1)
Rectangularity	How much the shape looks like a rectangle.	(highRectangularity EC-1)

Table 1: Detailed description of four attributes that describe geon shapes.

[2] NextKB can be downloaded at https://www.qrg.northwestern.edu/nextkb/index.html.

OpenCyc ontology. We use the predicate *isa* from OpenCyc to describe an object's category, for example,

(isa Object-1 Person)

indicates that the object *Object-1* is an instance of the category *Person*.

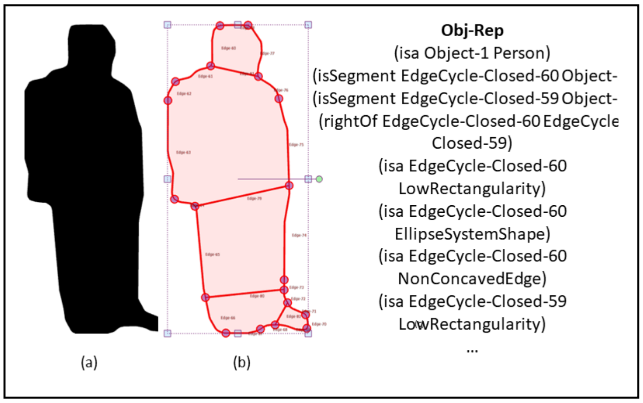
For object pose, we use the encoding approach from (Chen et al., 2019) to generate a geon representation. The encoding approach is inspired from recognition-by-component theory (Biederman, 1987), that people seem to encode visual input as a combination of simple shapes. In this encoding scheme, object masks imported into CogSketch are segmented into a set of *edge cycles* (Forbus et al., 2011), which are closed simple shapes. CogSketch computes the medial axis transform on the shape and generates *concave closures*, which are pairs of points on the object contour where the object is concave. A segmentation line is added to each concave closure to generate the set of segments represented as edge cycles. Each segment is described using four different attributes: *cross-sectional curvature, edge concavity, shape estimation, rectangularity*. Table 1 shows the details of each attribute. The connection relation and positional relation are described between each pair of segments. For each segment in an object, we use a predicate *isSegment* to indicate that the segment is part of the object, for example,

(isSegment Edge-Cycle-1 Object-1)

Figure 2 shows an object mask labeled as Person, the geon-segmentation and its corresponding Obj-reps. Obj-reps are computed for both Object and Subject of a triple.

For each pair of objects, each has a role in a relation triple, either Object or Subject. We use the unary predicates *isObject* and *isSubject* to indicate the roles for each object, for example,

(isObject Object-1)



Obj-Rep
(isa Object-1 Person)
(isSegment EdgeCycle-Closed-60 Object-
(isSegment EdgeCycle-Closed-59 Object-
(rightOf EdgeCycle-Closed-60 EdgeCycle
Closed-59)
(isa EdgeCycle-Closed-60
LowRectangularity)
(isa EdgeCycle-Closed-60
EllipseSystemShape)
(isa EdgeCycle-Closed-60
NonConcavedEdge)
(isa EdgeCycle-Closed-59
LowRectangularity)
...

Figure 2: (a) the mask of a Person object. (b) the geon-segmentation of the mask (c) Obj-reps for the object.

Spa-reps encode three types of spatial information between the pair of bounding boxes: RCC8 information (Cohn, 1996), positional information, and size information. RCC8 is widely used in qualitative spatial reasoning to describe topological relationships between regions. For example, if the bounding box of the tie is completely inside the bounding box for a person, this would be expressed as

(rcc8-NTPPi Tie Person)

We use six positional relations: *above, rightOf, enclosesHorizontally, enclosesVertically, centerAbove and centerRightOf.* The first four predicates describe the positional information on bounding boxes and the last two describe the positional information for their center points.

Size information is encoded with four predicates: *areaTiny, areaSmall, areaMedium* and *areaLarge*. The larger box in the pair is encoded as *areaLarge* and the smaller box is encoded based on the size relative to the larger box. If the area of smaller box is less than ¼ the area of the larger box, it is encoded as *areaTiny*. If the area of smaller box is between the ¼ and ½ the area of the larger box, it is encoded as *areaSmall*. Similarly, if it is between ½ and ¾ the area of larger box, it is encoded as *areaMedium* and above ¾ is encoded as *areaLarge*. The size predicates are unary, for example,

(areaLarge Object-1)

Combining the Obj-reps and Spa-reps, a relational representation is created for an object pair. Next, we describe how the relational representations of object pairs are used in analogical learning.

Analogical Learning: Our analogy stack uses three processes. Analogical matching is handled by the structure mapping engine (SME[3]) (Forbus et al., 2017) for analogical matching, analogical retrieval by MAC/FAC (Forbus et al., 1995), and generalization is performed by the Sequential Analogical Generalization Engine (SAGE) (McLure et al., 2015). We summarize each in turn.

SME is a computational model of analogical matching and similarity based on Structure Mapping Theory (Gentner, 1983). Given two cases consisting of structured, relational representations, called the *base* and *target*, SME computes a mapping between them. A mapping includes a set of correspondences that align entities and relations in the base and target, a similarity score that indicates how similar the base and the target are, and candidate inference, which are projections of unaligned structure from one case to the other, based on the correspondences. Here SME is used as a similarity metric and a means of combining cases into generalizations, as described below.

The MAC/FAC algorithm models analogical retrieval. Given a *probe* (a case) and a library of cases, MAC/FAC retrieves a highly similar case to the probe from that library.

[3] SME version 4 can be downloaded at https://www.qrg.northwestern.edu/software/sme4/index.html.

When cases are added to the library, a *content vector* is automatically constructed from the case, where each dimension represents the number of occurrences of a predicate in that case. The dot product of two content vectors provides a rough estimate of what SME would compute for a similarity score for the corresponding structured cases. Thus, the first stage, MAC, is a map/reduce operation, with content vector dot product of the probe with all the cases followed by accumulating the best *N* results. (In these experiments, $N = 5$.) The FAC stage is also map/reduce but using SME to compare the probe with the cases returned by MAC, returning the most similar case as a reminding. MAC provides scalability, while FAC provides the sensitivity to structure that human remindings exhibit. MAC/FAC is used for retrieval during both training and testing.

SAGE models analogical generalization. Each concept to be learned is represented by a *generalization pool* (aka *gpool*), which, given an incremental stream of examples, constructs a set of probabilistic generalizations and outliers that constitute an *analogical model* of that concept. Each item in a gpool is a disjunct in the model. There are two basic operations: adding an example and classifying an example.

When adding a training example to a gpool, MAC/FAC is used to retrieve the most similar item, treating the gpool as a case library. An *assimilation threshold* is used to determine whether an example is sufficiently similar to be merged. If the similarity is below this threshold, the new example is added to the gpool as an outlier. Otherwise, if the reminding is another example, then a new generalization is formed. This involves replacing non-identical aligned entities with new unique symbols (i.e. *skolems*) and taking the union of the statements involved. A probability is calculated for each statement, 1.0 if it is aligned in the match, and 0.5 otherwise. If the reminding is a generalization, it is updated by adding new statements, and perhaps new skolems, and updating the probability for each statement. Thus, a statement's probability reflects the frequency with which the examples assimilated into it contained an expression that mapped to that statement. Statements whose probability gets too low are eventually deleted. Since SAGE can accumulate multiple generalizations and outliers, it is like k-means with outliers, except that there is no a priori determination of how many clusters are needed: SAGE automatically derives that from the data.

For visual relation detection, the model needs to classify a relation category for a pair of objects. As the geon representations have many facts when object contours are complicated and each gpool has many cases, we use two-step process for relation classification to speed up the retrieval and scope relations to improve performance. In the first step, only object categories in Obj-reps and Spa-reps are used to provide a rough estimation for relations. We call this the *rough case* for an example. In the second step, full Obj-reps and Spa-reps are combined to predict the final relation from the estimated relations in first step. We call this the *full case* for an example. For every relation, there are two gpools, one for rough cases and one for full cases. During training, each relation triple has its rough and full cases computed, which are added to the appropriate gpools.

Relation Detection: Given a pair of objects detected in an image by the first stage, our algorithm builds its rough case and full case, as per above. The rough case is used as a probe to MAC/FAC, with all gpools for rough cases serving as the case library. The relations corresponding to the top five retrievals are used as filters for retrieval over full cases. That is, the full case is used as a probe with MAC/FAC over the union of all full-case gpools whose relations were retrieved in the prior step. The relation associated with the highest similarity score retrieved by MAC/FAC is assigned to the pair of entities as its classification. This process is run over every pair of detected objects in the image.

Experiments

We evaluate our hybrid system on the Visual Relationship Dataset (VRD) (Lu et al., 2016). We show that our model can get competitive results with lower training cost. Detecting a visual relational tuple involves classifying both object entities, the predicate between them, and the bounding boxes of both entities. Consequently, the performance of models relies on both the accuracy of object entity detectors and the visual relationship classifiers.

Following (Lu et al., 2016), we measure two conditions to evaluate the performance of our model. The first condition is predicate detection (PREDT). In PREDT, the input is an image, a set of localized objects in the image and their labels. The task is to predict a set of possible predicates between pairs of objects. This condition shows how relation detection via analogical learning performs on ground truth

Models	Recall@50	Recall@100
VRD (Lu et al., 2016)	47.87	47.87
VTransE (Zhang et al., 2017)	44.76	44.76
Zoom-Net (Yin et al., 2018)	50.69	50.69
LK (Yu et al., 2017)	55.16	55.16
CAI+SCA-M (Yin et al., 2018)	55.98	55.98
MMLFM-LC (Ma et al., 2019)	56.65	56.65
Ours	**52.38**	**52.38**

Table 2: Results of PREDT condition on VRD dataset.

Models	Recall@50	Recall@100
VTransE (Zhang et al., 2017)	14.07	15.20
SA-Full (Peyre et al., 2017)	15.80	17.10
Zoom-Net (Yin et al., 2018)	21.37	27.30
CAI+SCA-M (Yin et al., 2018)	22.34	28.52
KL (Yu et al., 2017)	22.68	31.89
Large-Scale VLU (Zhang et al., 2019)	26.98	32.63
Ours	16.12	18.41

Table 3: Results of RELDT condition on VRD dataset.

inputs. The second condition is relationship detection (RELDT). In RELDT, the input is only an image. The task is to output a set of triples <Object, Relation, Subject> and localize both entities in the image having at least 0.5 overlap with their ground truth boxes simultaneously. This condition evaluates how the whole pipeline performs on visual relation detection. We use the same evaluation metrices Recall@50 and Recall@100 as in (Zhang et al., 2019). The details of VRD dataset and implementations are introduced below.

Visual Relationship Detection Dataset

This dataset contains 5,000 images with 100 object categories and 70 predicates. There are 37,993 triple combinations total. We follow the popular train/test split, using 4,000 images for training and the other 1,000 images for testing. We trained the Faster-RCNN model using the method from (Zhang et al., 2019). For Mask-RCNN, since the VRD dataset lacks instance segmentation annotations, we directly use the checkpoint pre-trained on COCO dataset[4]. In SAGE, we use 0.8 for the assimilation threshold and 0.2 for the cut-off threshold (which eliminates low-probability facts from a generalization). Table 2 shows the results on PREDT task comparing with (Lu et al., 2016). Table 3 shows the results comparing with other state-of-art models on RELDT dataset. Besides the results in the two tables, we also compute Recall@1 of our model. Our system achieves 32.26 for Recall@1 in PREDT task and 8.91 in RELDT task.

Discussion

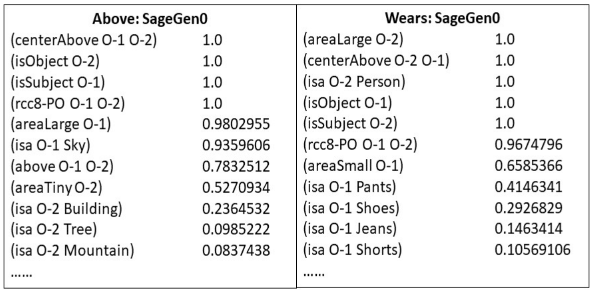
In Table 2, we compare our system with several existing models. From the results, our system outperforms three baseline models. In PREDT task, the ground truth object bounding boxes and categories are passed into model. The results show that our hybrid system has good performance when the object detection results are good. In Table 3, our results are better than VTransE and SA-Full. In RELDT task, detected object bounding boxes and categories from object detection model are used. Thus, the relation predication results depend on both the performance of object detections and relation classification. The results prove that our hybrid system has reasonable adaptation on noisy object detection results.

Indeed, although our model does not outperform all baselines, we use much less training cost to achieve this performance. Firstly, all baselines use the pre-trained object detector models in their first stage, which has similar cost as our model. However, using analogical learning, our model learns the generalization pools in the second stage with only one epoch on the whole training dataset. All other deep learning baselines require 7 to 30 epochs on the whole training dataset to converge. Thus, our model uses less training time to achieve the results. Also, analogical learning does not require to use expensive hardware resources such as GPUs, but all deep learning baselines need to use GPUs to speed up the training process. Besides less training cost, analogical learning is easier to understand than deep learning models. In the next section, we discuss the explainability of our model.

Explainability

The use of analogical learning for relation detection provides strong explainability. The contents of SAGE generalization pools consist of schema-like descriptions which can be easily understood by people. For example, Figure 3 shows the descriptions of the largest generalizations in Above and Wears gpools. In the generalization of Above, (centerAbove O1 O2) has probability score 1.0. O1 and O2 are the skolems for two different objects. This fact shows that, in many cases for this relation, the center of one object bounding boxes is above the center of the other object. Also, this generalization reveals information about common object types, e.g. that objects of type Sky have the relation Above with objects of types Building, Tree, Mountain, etc.

[4] We use the pre-trained model from https://github.com/facebookresearch/detectron2.



Above: SageGen0		Wears: SageGen0	
(centerAbove O-1 O-2)	1.0	(areaLarge O-2)	1.0
(isObject O-2)	1.0	(centerAbove O-2 O-1)	1.0
(isSubject O-1)	1.0	(isa O-2 Person)	1.0
(rcc8-PO O-1 O-2)	1.0	(isObject O-1)	1.0
(areaLarge O-1)	0.9802955	(isSubject O-2)	1.0
(isa O-1 Sky)	0.9359606	(rcc8-PO O-1 O-2)	0.9674796
(above O-1 O-2)	0.7832512	(areaSmall O-1)	0.6585366
(areaTiny O-2)	0.5270934	(isa O-1 Pants)	0.4146341
(isa O-2 Building)	0.2364532	(isa O-1 Shoes)	0.2926829
(isa O-2 Tree)	0.0985222	(isa O-1 Jeans)	0.1463414
(isa O-2 Mountain)	0.0837438	(isa O-1 Shorts)	0.10569106
……		……	

Figure 3: Scheme-like descriptions and corresponding probabilities in largest generalizations of Above and Wears gpools.

Similarly, objects of type Sky tend to have (areaLarge O-1) and all other objects have (areaTiny O-2), which means Sky is much larger than other objects. In the generalization of Wears, one of the objects is Person and the other object is clothes for lower body, such as Pants, Jeans or Shorts. Therefore, Person has large area and cloth objects have small area. The RCC8 relation PO has high score in this generalization, which means the two objects have intersection with each other. These probabilistic generalizations provide new insights, including possibly into dataset bias. Moreover, one interesting possibility is tuning learned knowledge, via trainers manually editing facts, something which is difficult for deep learning models.

Conclusion

We present a hybrid system combining deep learning models and analogical learning on visual relation detection using object information and spatial information between objects. Results on the PREDT task indicate that given accurate object detections, analogical learning is a promising approach to detect relations in images. Furthermore, analogical learning is more efficient for training than deep learning and has better explainability. People can easily explore the learned generalizations to understand the high-probability generalizations and outliers, which provides a more solid foundation for building reliable and human-like visual systems. Results on RELDT also shows that analogical learning is flexible enough to combine with other methods.

We see two important lines of future work. The first is to examine performance with other deep-learning modules in the first encoding stage. The second is to explore doing richer encoding of object shapes, moving beyond bounding boxes and including part information about objects.

References

Anderson, E. M.; Chang, Y.; Hespos, S.; and Gentner, D. 2018. Comparison within pairs promotes analogical abstraction in three-month-olds. *Cognition*, pages 176:74-86.

Buhrmester, V.; Munch, D.; and Arens, M. 2019. Analysis of explainers of black box deep neural networks for computer vision: a survey. *arXiv preprint arXiv:1911.12116.*

Cohn, A. Calculi for qualitative spatial reasoning. In *Artificial Intelligence and Symbolic Mathematical Computation, LNCS 1138*, (pp. 124-143). New York: Springer-Verlag.

Ciresan, D.C.; Meier, U.; Gambardella, L. M.; and Schmidhuber, J. 2011. Convolutional neural network committees for handwritten character classification. *Document Analysis and Recognition (ICDAR),* Page. 1135-1139.

Chen, K.; Rabkina, I.; McLure, M. D.; and Forbus, K. D. 2019. Human-like Sketch Object Recognition via Analogical Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 1336-1343.

Chen, K.; and Forbus, K.D. 2018. Action Recognition from Skeleton Data via Analogical Generalization over Qualitative Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

Crouse, M.; McFate, C.; and Forbus, K.D. 2018. Learning from Unannotated QA Pairs to Analogically Disambiguate and Answer Questions. In *Proceedings of AAAI Conference on Artificial Intelligence*.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. *IEEE conference on computer vision and pattern recognition*.

Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; and Li F. F. 2009. Imagenet: A large-scale hierarchical image database. *IEEE conference on computer vision and pattern recognition*.

Forbus, K. D. 1995. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 131-205.

Forbus, K.D.; Usher, J.; Lovett, A.; Lockwood, K.; and Wetzel, J. 2011. CogSketch: Sketch understanding for cognitive science research and for education. Cognitive Science, 3, no. 4, 648-666.

Forbus, K. D.; Ferguson, R. W.; Lovett, A.; and Gentner, D. 2017. Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, 1152-1201.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170

Gentner, D. 2003. Why we're so smart. In *Language in Mind: Advances in the study of language and thought* (pp. 195-235). MIT Press.

He, K.; Gkioxari, G.; Dollar, Piotr.; and Girshick, R. 2017. Mask R-CNN. *arXiv preprint arXiv:1703.06870*

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; and Ferrari. V. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint*, 1881.00982.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D. A.; Bernstein, M. S. and Li, F. F. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123.1.

Kandaswamy, S.; Forbus, K.D.; and Gentner, D. 2014. Modeling Learning via Progressive Alignment using

Interim Generalizations. *Cognitive Science Society*, Vol. 36, No. 36.

Lovett, A; and Forbus, K. D. 2013. Modeling spatial ability in mental rotation and paper-folding. *Cognitive Science Society*, Vol. 25.

Lovett, A.; and Forbus, K. D. 2017. Modeling visual problem solving as analogical reasoning. *Psychological review*, 124.1:60.

Liang, C.; and Forbus, K. 2015. Learning Plausible Inferences from Semantic Web Knowledge by Combining Analogical Generalization with Structured Logistic Regression. *Proceedings of AAAI15.*

Lu, C.; Krishna, R.; Bernstein, M.; Li F.F. 2016. Visual Relationship Detection with Language Priors. *European Conference on Computer Vision.*, *Springer, Cham.*

Lin, T. Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollar, P. 2014. Microsoft COCO: Common Objects in Context. *Springer, Cham.*

McLure, M.; Friedman, S.; and Forbus, K. D. 2015. Extending Analogical Generalization with Near-Misses. *In Proceedings of AAAI conference of Artificial Intelligence*, 565-571.

Ma, X.; Bao, B.; and Yao, L; Xu, C. 2019. Multimodel Latent Factor Model with Language Constraint for Predicate Detection. *In 2019 IEEE International Conference on Image Processing (ICIP).*

Peyre, J.; Laptev, I.; Schmid, C.; and Sivic, J. 2017. Weakly-supervised learning of visual relations. *In ICCV.*

Ren, S.; He, K.; Girshick, R.; Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*.

Sagi, E.; Gentner, D.; and Lovett, A. What difference reveals about similarity. *Cognitive Science*, 36(6), 1019-1050.

Samek, W.; Wiegand, T.; and Muller, K. 2017. Explainable artificial intelligence: understanding, visualizing, and interpreting deep learning models. *arXiv preprint arXiv:1708.08296.*

Xu, D.; Zhu, Y.; Choy, C. B.; Li, F. F. 2017. Scene graph generation by iterative message passing. *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Page. 5410-5419.

Yin, G.; Sheng, L.; Liu, B.; Yu, N.; Wang, X.; Shao, J.; and Change Loy, C. 2018. Zoom-net: Mining deep feature interactions for visual relationship recognition. *In The European Conference on Computer Vision (ECCV).*

Yu, R.; Li, A.l Morariu, V. I.; and Davis, L. S. 2017. Visual relationship detection with internal and external linguistic knowledge distillation. *In the IEEE International Conference on Computer Vision (ICCV)*

Zhuang, B.; Liu, L.; Shen, C.; and Reid, I. 2017. Towards context-aware interaction recognition for visual relationship detection. *In proceedings of the IEEE International Conference on Computer Vision*, Page. 589-598.

Zhang, H.; Kyaw, Z.; Chang, S. F.; Chua, T. S. 2017. Visual Translation Embedding Network for Visual Relation Detection. *arXiv preprint arXiv:1802.08319.*

Zhang, J.; Shih, K. J.; Elgammal, A.; Tao, A.; and Catanzaro, B. 2019. Graphical Contrastive Losses for Scene Graph Parsing. *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Page. 11535-11543.

Zhang, J.; Kalantidis, Y.; Rohrbach, M.; Paluri, M.; Elgammal, A.; and Elhoseiny, M. 2019. Large-Scale Visual Relationship Understanding. *In AAAI 2019.*

Zhang, H,. Kyaw, Z., Chang, S.F. and Chua, T.S. 2017. Visual Translation Embedding Network for Visual Relation Detection. *In CVPR 2017.*

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural Motifs: Scene Graph Parsing with Global Context. *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Page. 5831-5840.