Sherlock: Scalable Fact Learning in Images

Mohamed Elhoseiny,[1,2] **Scott Cohen,**[1] **Walter Chang,**[1] **Brian Price,**[1] **Ahmed Elgammal**[2]
[1] Adobe Research [2] Rutgers University, Computer Science Department

Abstract

The human visual system is capable of learning an unbounded number of facts from images including not only objects but also their attributes, actions and interactions. Such uniform understanding of visual facts has not received enough attention. Existing visual recognition systems are typically modeled differently for each fact type such as objects, actions, and interactions. We propose a setting where all these facts can be modeled simultaneously with a capacity to understand an unbounded number of facts in a structured way. The training data comes as structured facts in images, including (1) objects (e.g., <boy>), (2) attributes (e.g., <boy, tall>), (3) actions (e.g., <boy, playing>), and (4) interactions (e.g., <boy, riding, a horse >). Each fact has a language view (e.g., < boy, playing>) and a visual view (an image). We show that learning visual facts in a structured way enables not only a uniform but also generalizable visual understanding. We propose and investigate recent and strong approaches from the multiview learning literature and also introduce a structured embedding model. We applied the investigated methods on several datasets that we augmented with structured facts and a large scale dataset of > 202,000 facts and 814,000 images. Our results show the advantage of relating facts by the structure by the proposed model compared to the baselines.

Introduction

It is a capital mistake to theorize in advance of the facts. -Sherlock Holmes (The Adventure of the Second Stain)

Despite recent significant advances in Computer Vision, there is still a large gap between humans and machines in visual understanding. The human visual system is capable of efficiently acquiring visual knowledge by learning different types of facts in a never ending way from many or few examples, aided by the ability to generalize from other known facts with related structure. For example, a human can learn the concept of climbing and generalize it to rare interactions like "hippo climbing fence". To bridge this gap, we carefully focus on five desirable characteristics

- **Uniformity**: ability to handle objects ("dog"), attributes ("brown dog"), actions ("dog running") and interactions between objects ("dog chasing cat"); see Fig 1 for other examples.
- **Generalization**: ability to generalize to facts that have zero or few examples during training.
- **Scalability**: ability to learn an unbounded number of facts in one model.
- **Structure**: ability to provide a structured understanding of facts, for example that "baby" is the subject and has an attribute of "smiling".
- **Bi-directionality**: ability to retrieve a linguistic representation of an image, and the ability to retrieve images given language description of a fact.

Figure 1: Visual Facts in Images

Existing visual understanding systems may be categorized by two trends: (1) fact-level systems and (2) high-level systems. Fact level systems include object recognition, action recognition, attribute recognition, and interaction recognition (e.g., (Simonyan and Zisserman 2015), (Zhang et al. 2014), (Chen and Grauman 2014), (Zhou et al. 2014), (Gkioxari and Malik 2015), (Antol, Zitnick, and Parikh 2014a)). These systems are usually evaluated separately for each fact type (e.g., objects, actions, interactions, attributes, etc.) and are therefore not uniform. Typically, these systems have a fixed dictionary of facts (assuming that facts are found during training in at least tens of examples), and they treat facts independently. Such systems cannot generalize fact learning outside of the dictionary and will not scale to an unbounded number of facts, since model size scales with the number of facts. Furthermore, these recognition systems are typically uni-directional, only able to return the conditional probability of a fact given an image. In the zero/few-shot learning setting (e.g., (Romera-Paredes and Torr 2015; Lampert, Nickisch, and Harmeling 2009)), only a few or

even zero examples per fact are available; this is typically studied apart from the traditional recognition setting. We are not aware of a unified recognition/few shot learning system that learns an unbounded set of facts.

In the second trend, researchers study high-level tasks like image captioning (e.g., (Karpathy, Joulin, and Li 2014; Vinyals et al. 2015; Xu et al. 2015a; Mao et al. 2015)), image-caption similarity (e.g., (Karpathy, Joulin, and Li 2014; Kiros et al. 2015)), and visual question answering (e.g., (Antol et al. 2015; Malinowski, Rohrbach, and Fritz 2015; Ren, Kiros, and Zemel 2015)) with very promising results. These systems typically learn high-level tasks but their evaluation does not answer of whether the systems can relate captions or questions to images by fact-level understanding. Psychologists have also studied how people caption images, e.g., (Chaplin 2006; Kaufman et al. 2007) showed that captions usually reflect what a person sees as discriminative about a scene. This means that different captions for the same image may convey different facts and but are constrained to correlate to the same scene in existing image-caption similarity systems. In principle, captioning models can relate images to sentences and thus can mention any fact. However, we show in our experiments that image-caption similarity systems trained on MS COCO dataset (Lin et al. 2014) are confused when they are evaluated for fact level understanding even at a small scale; see Table 1. Also, (Devlin et al. 2015a; 2015b) reported that 60-70% of the captions generated by LSTM-based methods actually exist in the training data; the authors also show that nearest neighbor methods have competitive captioning performance. These results call into question both the core understanding and the generalization of the state-of-the-art caption-level systems.

High-level tasks requires attention to the fundamental problem of rich understanding of images at the scale of millions of unique facts, which is exactly what we work towards in this paper and may open the door to open-ended visual reasoning. Our goal in particular is a method that achieves a more sophisticated understanding of the actions, objects, attributes, and interactions between objects, and possesses the necessary properties of scalability, generalization, uniformity, bi-directionality, and structure. To achieve these properties, the key to our solution is to make the basic unit of understanding the structured fact, as shown in Fig. 1, and to have a structured embedding space in which different dimensions record information about the subject S, predicate P, and object O for a fact.

Contributions: (1) We propose a problem setting to help study fact-level visual understanding of an unbounded number of facts while considering the aforementioned characteristics. (2) We design and investigate several baselines from multiview learning literature and apply them on this task. (3) We propose a learning representation model that relate different fact types using the structure exemplified in Fig 1. (4) We setup a large scale benchmark for this task that consists of more than $814,000$ examples and $202,000$ unique facts and we show the value of relating facts by structure using the proposed model in comparison to the designed baselines.

Related Work

Many approaches has been proposed to facilitate recognition in various settings, which may be categorized as follows

(A) Modeling Visual facts in Discrete Space: Recognition of objects or activities has been typically modeled as a mapping function $g : \mathcal{V} \rightarrow \mathcal{Y}$, where $\mathcal{Y}$ is discrete set of classes. The function g has recently been learned using deep learning (e.g.,VGGNet (Simonyan and Zisserman 2015; Szegedy et al. 2015)). Apart from being uni-driectional from $\mathcal{V}$ to $\mathcal{Y}$, different systems are typically built to recognize each fact type which requires maintaining and retraining different systems as new facts are added, and also does not allow learning the correlation among different fact types (e.g., "person riding wave" and "dog riding a wave") which limits its generalization. It is also not scalable since model size increases as new facts are added.

(B) zero/few shot fact learning with attributes: One of the most successful ideas for learning from few examples per class is by using semantic output codes like attributes as an intermediate layer between features and classes. Formally, g is a composition of two function $g = h(a(\cdot))$, where $a : \mathcal{V} \rightarrow \mathcal{A}$, and $h : \mathcal{A} \rightarrow \mathcal{Y}$ (Palatucci et al. 2009). The main idea is to learn an intermediate attribute layer, upon which classes are then represented to facilitate zero-shot/few-shot learning. (Chen and Grauman 2014) realized that attribute appearance is dependent on the class, as opposed to these earlier models (Palatucci et al. 2009; Lampert, Nickisch, and Harmeling 2009; Farhadi et al. 2009). However (Chen and Grauman 2014)'s approach is not scalable since it learns different classifiers for each category-attribute pair. More recently, Attribute Embedding (Akata et al. 2013) and ESZSL (Romera-Paredes and Torr 2015) have shown strong zero-shot performance by joint embedding images and attributes. However, their capability in a rich understanding setting was not explored.

(C) Object Recognition with Vision and Language: Recent works in language& vision involve using unannotated text to improve object recognition and to facilitate zero-shot learning. The following group of approaches model object recognition as a $g(v) = \arg\max_y s(v \in \mathcal{V}, y \in \mathcal{Y})$, where $s(\cdot, \cdot)$ is a similarity function between image v and class y represented by text. In (Frome et al. 2013), (Norouzi et al. 2014) and (Socher et al. 2013), word embedding language models (e.g., (Mikolov et al. 2013)) were adopted to represent class names as vectors. In their setting, the imageNet dataset has 1000 object facts with thousands of examples per class. In contrast, our setting assumes no limit for the facts and naturally has a long-tail distribution witrh two orders of magnitude more facts. Conversely, other works model the mapping of unstructured text descriptions for classes into a visual classifier (Elhoseiny, Saleh, and Elgammal 2013; Elhoseiny, Elgammal, and Saleh 2016; Ba et al. 2015). In contrast, we aim at extending the recognition task to an unbounded scale of facts, not only object recognition but also attributes, actions, and interactions in one model.

(D) Image/Video-Caption Similarity Methods: Several approaches have been proposed in this area (e.g., (Karpathy, Joulin, and Li 2014; Kiros et al. 2015; Vendrov et al. 2016)

Table 1: Image-Caption Models on Fact Level understanding (From "Image to Fact"(i2f) and from Fact to Image" (f2i)). Performance metrics are detailed in the experiments section.

		Caption-level (MS COCO)		**Fact-Level**	
Dataset	Method	**i2f**(Acc)	**f2i (mAP)**	**i2f** (Acc)	**f2i**(mAP)
Standord40	(Kiros et al. 2015)	33.73	26.29	60.86	51.9
(Yao et al. 2011)	(Vendrov et al. 2016)	32.32	29.3	66.36	67.78
Pascal 10 Actions	(Kiros et al. 2015)	46.050	40.712	60.27	50.58
(Everingham et al.)	(Vendrov et al. 2016)	33.108	36.247	66.102	69.884
6DS (186)	(Kiros et al. 2015)	15.71	9.37	26.13	26.17
see Table 2	(Vendrov et al. 2016)	12.56	10.21	31.57	35.23

for images and (Xu et al. 2015b; Elhoseiny et al. 2016b) for video). There is two important and interesting questions to explore. First, how image-caption similarity methods trained on caption level (the typical setting) performs on fact-level understanding; see Table 1 (caption-level columns). Second, these systems could be retrained in our setting by providing them with fact-level annotation, where every example is a phrase representing the fact and an image (e.g., "person riding horse" and an image with this fact); see Table 1 (fact-level columns). The table consistently shows a big gap between the two settings on three datasets varying in scale. These results motivated us to dig deeper and study more models to enable richer understanding.

Representation and Visual Modifiers

We deal with three groups of facts; see Fig. 1. First Order Facts <S,*,*> are object and scene categories (e.g., <baby,*,*>, <girl,*,*>, <beach,*,*>). Second Order Facts <S,P,*> are objects performing actions or attributed objects (e.g., <baby, smiling,*>, <baby, Asian,*>). Third Order Facts <S,P,O> are interactions and positional information (e.g. <baby, sitting_on, high_chair>, <person, riding, horse>). By allowing wild-cards in this structured representation (<baby,*,*>and <baby, smiling,*>), we can not only allow uniform representation of different fact types but also relate them by structure. We propose to model these facts by embedding them into a "structured" fact space that has three "continuous" hyper-dimensions ϕ_S, ϕ_P, and ϕ_O

$\phi_S \in \mathbb{R}^{d_S}$**:** The space of object categories or scenes S.

$\phi_P \in \mathbb{R}^{d_P}$**:** The space of actions, interactions, attributes, and positional relations.

$\phi_O \in \mathbb{R}^{d_O}$**:** The space of interacting objects, scenes that interact with S for SPO facts.

where d_S, d_P, and d_O are the dimensionalities corresponding to ϕ_S, ϕ_P, and ϕ_O, respectively. First order facts like <woman,*,*>, and <man,*,*> live in a hyper-plane in the $\phi_P \times \phi_O$ space. Second order facts (e.g., <man, walking,*>, <girl, walking,*>) live as a hyper-line that is parallel to ϕ_O axis. Finally, a third order fact like <man, walking, dog> is a point in the $\phi_S \times \phi_P \times \phi_O$ perceptual space. Inspired from the concept of language modifiers, the ϕ_S, ϕ_P, and ϕ_O could be viewed as what we call "visual modifiers". For example, the second order fact <baby, smiling,* > is a ϕ_P visual modifier for <baby,*,*>, and the third order fact <person, playing, flute> is the fact <person, *, *> visually modified on both ϕ_P and ϕ_O axes. By embedding

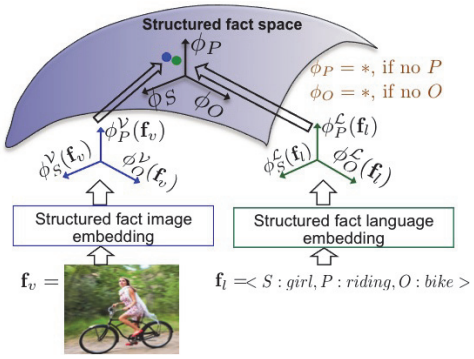

Figure 2: Structured Embedding

all language and images into this common space, our algorithm can scale efficiently. Further, this structured space can be used to retrieve a language view of an image as well as a visual view of a language description, making the model bi-directional.Modeling visual recognition based on this notion gives it a generalization capability. For example, if the model learned the facts <boy>, <girl>, <boy, petting, dog>, <girl, riding, horse>, we would aim at recognizing an unseen fact <boy, petting, horse>. We show these capabilities quantitatively in our experiments.

We model this setting as a problem with two views, one in the visual domain $\mathcal{V}$ and one in the language domain $\mathcal{L}$. Let $\mathbf{f}$ be a structured fact, $\mathbf{f}_v \in \mathcal{V}$ denoting the visual view of $\mathbf{f}$ and $\mathbf{f}_l \in \mathcal{L}$ denoting the language view of $\mathbf{f}$. For instance, an annotated fact with language view $\mathbf{f}_l$ =<S:girl, P:riding, O:bike> would have a corresponding visual view $\mathbf{f}_v$ as an image where this fact occurs; see Fig. 2.

We denote the embedding functions from a visual view to ϕ_S, ϕ_P, and ϕ_O as $\phi_S^{\mathcal{V}}(\cdot)$, $\phi_P^{\mathcal{V}}(\cdot)$, and $\phi_O^{\mathcal{V}}(\cdot)$, and the structured visual embeddings of a fact $\mathbf{f}_v$ by $\mathbf{v}_S = \phi_S^{\mathcal{V}}(\mathbf{f}_v)$, $\mathbf{v}_P = \phi_P^{\mathcal{V}}(\mathbf{f}_v)$, and $\mathbf{v}_O = \phi_O^{\mathcal{V}}(\mathbf{f}_v)$, respectively. Similarly, we denote the embedding functions from a language view to ϕ_S, ϕ_P, and ϕ_O as $\phi_S^{\mathcal{L}}(\cdot)$, $\phi_P^{\mathcal{L}}(\cdot)$, and $\phi_O^{\mathcal{L}}(\cdot)$, and the structured language embeddings of a fact $\mathbf{f}_l$ as $\mathbf{l}_S = \phi_S^{\mathcal{L}}(\mathbf{f}_l)$, $\mathbf{l}_P = \phi_P^{\mathcal{L}}(\mathbf{f}_l)$, and $\mathbf{l}_O = \phi_O^{\mathcal{L}}(\mathbf{f}_l)$. Let $\mathbf{v} = [\mathbf{v}_S, \mathbf{v}_P, \mathbf{v}_O]$ and $\mathbf{l} = [\mathbf{l}_S, \mathbf{l}_P, \mathbf{l}_O]$ (concatenation). Third-order facts <S,P,O> can be directly embedded in the structured fact space with $\mathbf{v} \in \mathbb{R}^{d_S} \times \mathbb{R}^{d_P} \times \mathbb{R}^{d_O}$ for the image view and $\mathbf{l} \in \mathbb{R}^{d_S} \times \mathbb{R}^{d_P} \times \mathbb{R}^{d_O}$ for the language view. Based on the "fact modifier" observation, we represent both second and first-order facts by wild cards "$*$", as illustrated in Eq. 2, 4. Set-

ting ϕ_P and ϕ_O to $*$ for first-order facts means that the P and O modifiers are not of interest for first-order facts, which is intuitive. Similarly, setting ϕ_O to $*$ for second-order facts indicates that the O modifier is not of interest for single-frame actions and attributed objects. If an image contains lower order fact such as <man>, then higher order facts such as <man, tall> or <man, walking, dog> may also be present. Hence, the wild cards (i.e. $*$) of the first- and second-order facts are not penalized during training.

$$\textbf{Second-Order <S, P,*>:}\quad \mathbf{v} = [\mathbf{v}_S, \mathbf{v}_P, \mathbf{v}_O = *] \tag{1}$$

$$\mathbf{l} = [\mathbf{l}_S, \mathbf{l}_P, \mathbf{l}_O = *] \tag{2}$$

$$\textbf{First-Order <S,*,*>:}\quad \mathbf{v} = [\mathbf{v}_S, \mathbf{v}_P = *, \mathbf{v}_O = *] \tag{3}$$

$$\mathbf{l} = [\mathbf{l}_S, \mathbf{l}_P = *, \mathbf{l}_O = *] \tag{4}$$

Structured Embedding Model

We propose a structured fact embedding model while considering the five properties discussed in the introduction. Satisfying the first four properties can be achieved by using a generative model $p(\mathbf{f}_v, \mathbf{f}_l)$ that connects the visual and the language views of $\mathbf{f}$, where more importantly $\mathbf{f}_v$ and $\mathbf{f}_l$ inhabit a continuous space. We model $p(\mathbf{f}_v, \mathbf{f}_l) \propto s(\mathbf{v}, \mathbf{l})$, where $s(\cdot, \cdot)$ is a similarity function defined over the structured fact space. We satisfy the fifth property by building our models over the aforementioned structured wild card representation. Our objective is that two views of the same fact should be embedded so that they are close to each other; see Fig 2. The question now is how to model and train $\phi^{\mathcal{V}}(\cdot)$ visual functions $(\phi^{\mathcal{V}}_S(\cdot), \phi^{\mathcal{V}}_P(\cdot), \phi^{\mathcal{V}}_O(\cdot))$ and $\phi^{\mathcal{L}}(\cdot)$ language functions $(\phi^{\mathcal{L}}_S(\cdot), \phi^{\mathcal{L}}_P(\cdot), \phi^{\mathcal{L}}_O(\cdot))$. We model $\phi^{\mathcal{V}}(\cdot)$ as a CNN encoder (e.g., (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015)), and $\phi^{\mathcal{L}}(\cdot)$ as word2vec encoder (e.g., (Mikolov et al. 2013; Pennington, Socher, and Manning 2014)) due to their recent success as encoders for images and words, respectively.

We start by defining an activation operator $\psi(\theta, a)$, where a is an input, and θ is a series of one or more neural network layers (may include different layer types, e.g., convolution, pooling, then another convolution and pooling). The operator $\psi(\theta, a)$ applies θ parameters layer by layer to compute the final activation of a using θ subnetwork.

Structured fact CNN image encoder: We use different convolutional layers for S than for P and O, inspired by the idea that P and O are modifiers to S (Fig. 3(a)). Starting from $\mathbf{f}_v$, there is a common set of convolutional layers, denoted by θ^0_c, then the network splits into two branches, producing two sets of convolutional layers θ^S_c and θ^{PO}_c, followed by two sets of fully connected layers θ^S_u and θ^{PO}_u. Finally $\phi^{\mathcal{V}}_S(\mathbf{f}_v)$,$\phi^{\mathcal{V}}_P(\mathbf{f}_v)$, and $\phi^{\mathcal{V}}_O(\mathbf{f}_v)$ are computed by applying W^S, W^P, and W^O transformation matrices for subject, predicate, and object respectively. If we define the output of the common S,P,O layers as $d = \psi(\theta^0_c, \mathbf{f}_v)$ and the output of the P,O column as $e = \psi(\theta^{PO}_u, \psi(\theta^{PO}_c, d))$, then

$$\mathbf{v}_S = W^S\ \psi(\theta^S_u, \psi(\theta^S_c, d)),\ \mathbf{v}_P = W^P\ e,\ \ \mathbf{v}_O = W^O\ e. \tag{5}$$

Structured fact language encoder: The structured fact language view is encoded using word embedding vectors for S, P and, O separately. Hence

$$\mathbf{l}_S = \text{vec}_{\theta_l}(\mathbf{f}^S_l), \mathbf{l}_P = \text{vec}_{\theta_l}(\mathbf{f}^P_l), \mathbf{l}_O = \text{vec}_{\theta_l}(\mathbf{f}^O_l) \tag{6}$$

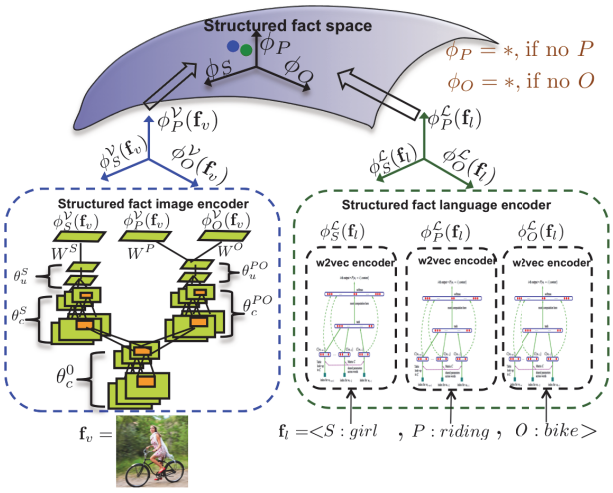

Figure 3: Structured Embedding Model (SEM). See Fig. 2 for the full picture.

where $\mathbf{f}^S_l$, $\mathbf{f}^P_l$, and $\mathbf{f}^O_l$ are the Subject, Predicate, and Object parts of $\mathbf{f}_l \in \mathcal{L}$. For each of them, the literals are dropped. In our experiments, θ_l is fixed to a pre-trained word vector embedding model (e.g. (Mikolov et al. 2013; Pennington, Socher, and Manning 2014)) for $\mathbf{f}^S_l$, $\mathbf{f}^P_l$, and $\mathbf{f}^O_l$; see Fig 3(b).

Loss function: One way to model $p(\mathbf{f}_v, \mathbf{f}_l)$ is to assume that $p(\mathbf{f}_v, \mathbf{f}_l) \propto= \exp(-\text{loss}_w(\mathbf{f}_v, \mathbf{f}_l))$ and minimize the distance $\text{loss}_w(\mathbf{f}_v, \mathbf{f}_l)$ defined as

$$\text{loss}_w(\mathbf{f}_v, \mathbf{f}_l) = w^{\mathbf{f}}_S \cdot D(\mathbf{v}_S, \mathbf{l}_S) + w^{\mathbf{f}}_P \cdot D(\mathbf{v}_P, \mathbf{l}_P) + w^{\mathbf{f}}_O \cdot D(\mathbf{v}_O, \mathbf{l}_O).$$

where $D(\cdot, \cdot)$ is a distance function. Thus we minimize the distance between the embeddings of the visual view and the language view. Our solution to penalize wild-card facts is to ignore their wild-card modifiers in the loss. Here $w^{\mathbf{f}}_S = 1$, $w^{\mathbf{f}}_P = 1$, $w^{\mathbf{f}}_O = 1$ for <S,P,O> facts , $w^{\mathbf{f}}_S = 1$, $w^{\mathbf{f}}_P = 1$, $w^{\mathbf{f}}_O = 0$ for <S,P> facts, and $w^{\mathbf{f}}_S = 1$, $w^{\mathbf{f}}_P = 0$, $w^{\mathbf{f}}_O = 0$ for <S> facts. Hence loss_w does not penalize the O modifier for second-order facts or the P and O modifiers for first-order facts, which follows our definition of wild-cards. In this paper, we used $D(\cdot, \cdot)$ as the standard Euclidean distance.

Testing (Two-view retrieval): After training, we embed all the testing $\mathbf{f}_v$s (images) by the learnt models, and similarly embed all the test $\mathbf{f}_l$s as shown in Eq 6. For language view retrieval (retrieve relevant facts in language given an image), we compute the distance between the structured embedding of an image $\mathbf{v}$ and all the facts structured language embeddings $\mathbf{l}$s, which indicates relevance for each fact $\mathbf{f}_l$ for the given image. For visual view retrieval (retrieve relevant images given fact in language form), we compute the distance between the structured embedding of the given fact $\mathbf{l}$ and all structured visual embedding of images $\mathbf{v}$s in the test set. For first and second order facts, the wild-card part is ignored while computing the distance.

Experiments

Datasets and Large Scale Benchmark

We began our data collection by augmenting existing datasets with fact language view labels $\mathbf{f}_l$: PPMI (Yao and Fei-Fei 2010), Stanford40 (Yao et al. 2011), Pascal Actions (Everingham et al.), Sports (Gupta 2009), Visual Phrases (Sadeghi and Farhadi 2011), INTERACT (Antol, Zitnick, and Parikh 2014b) datasets. The union of these 6 datasets resulted in 186 facts with 28,624 images as broken out in Table 2. We added to this data, structured facts from the Scene Graph dataset (Johnson et al. 2015) with 5000 manually annotated images in a graph structure from which first-, second-, and third-order relationships can be extracted. We extracted 110,000 second-order facts and 112,000 third-order facts. The majority of these are positional relationships.We also added to the aforementioned data, 380,000 second and third order fact images from(Elhoseiny et al. 2016a). We further augmented these data with 2000 images for each MS COCO object (80 classes) as first-order facts. We also used object annotations in the Scene Graph dataset as first-order fact annotations with a maximum of 2000 images per object. Table 3 shows the unique facts of the resultant large scale dataset with $>$ 202K facts and 814,000 images. For fast retrieval, we used the FLANN library (Muja and Lowe 2009) to compute the (approximate) 100 nearest neighbors for $\mathbf{f}_l$ given $\mathbf{f}_v$, and vice-versa.

Setup of of the investigated Models

Our Structured Embedding Model (SEM-A and SEM-B): We used is the GloVE840B model (Pennington, Socher, and Manning 2014) to encode structured facts in the language view (i.e., θ_l). For the visual encoder, The shared layers θ_c^0 match the architecture of the convolutional layers and pooling layer in VGG-16 named `conv_1_1` until `pool3`, and have seven convolution layers. The subject layers θ_c^S and predicate-object layers θ_c^{PO} are two branches of convolution and pooling layers with the same architecture as VGG-16 layers named `conv_4_1` until `pool5` layer, which makes six convolution-pooling layers in each branch. Finally, θ_u^S and θ_u^{PO} are two instances of `fc6` and `fc7` layers in VGG-16 network. W^S, W^P, and W^O are initialized randomly and the rest are initialized from VGG-16 trained on ImageNet.

Table 2: Our fact augmentation of six datasets

	Unique language views $\mathbf{f}_l$				Number of ($\mathbf{f}_v$, $\mathbf{f}_l$) pairs			
	S .	SP.	SPO .	total	S	SP	SPO	total images
INTERACT	0	0	60	60	0	0	3171	3171
VisualPhrases	11	4	17	32	3594	372	1745	5711
Stanford40	0	11	29	40	0	2886	6646	9532
PPMI	0	0	24	24	0	0	4209	4209
SPORT	14	0	6	20	398	0	300	698
Pascal Actions	0	5	5	10	0	2640	2663	5303
Union	**25**	**20**	**141**	**186**	**3992**	**5898**	**18734**	**28624**

Table 3: Large Scale Dataset (202K facts, 814K images)

	S	SP	SPO	Total
Training unique facts	6116	57681	107472	171269
Testing unique facts	2733	22237	33447	**58417**
Train/Test unique Intersection	1923	13043	11774	26740
Test unique unseen facts	**810**	**9194**	**21673**	**31677**

In order to show the value of branching some convolutional layers, we performed experiment on a variant of our SEM model where $\theta_c^S = \theta_c^{PO} = \theta_c$ and $\theta_u^S = \theta_u^{PO} = \theta_u$ (all layers shared except W^S, W^P, and W^O). We denote this Model as SEM-B while the original SEM model as SEM-A.

Multiview CCA IJCV14 (Gong et al. 2014) (MV CCA) and ESZSL ICML15 Baseline (Romera-Paredes and Torr 2015) : Both methods expects features from both views. For visual view features, we used VGG16 (FC6). For the language view features, we used GloVE. Since MV CCA does not support wild-cards, we fill the wild-card parts of $\Phi^{\mathcal{L}}(\mathbf{f}_l)$ with zeros for First Order and Second order facts.

Image-Sentence Similarity (TACL15 (Kiros et al. 2015)) (MS COCO pretrained) and (retrianed): We used the theano implementation of this method published by the authors.In order to test these models in our setting, we provide them with the image and a phrase constructed from the fact language representation. For example <person, riding, horse $>$ is converted to "person riding horse". We evaluated two instances of this model (One trained on captions level on MSCOCO dataset) and another instance retrained on image-fact training pairs where facts are converted to phrases.

It is not hard to observe the multiview nature of the baselines make them applicable to bidirectional retrieval tasks that we evaluate against our method.

Evaluation Metrics For visual view retrieval (To retrieve images given a fact in language view like <S: person, P: riding, O: horse>), we measure the performance by mAP (mean Average Precision) and a variant of it mAP10 which is restricted to the top 100 images respectively. For recognizing facts given an image (language view retrieval), we use top 1, top 5, top 10 accuracy.

Datasets Setup We performed experiments on several datasets ranging in scale: Stanford40 (Yao et al. 2011), Pascal Actions (Yao and Fei-Fei 2010), Visual Phrases (Sadeghi and Farhadi 2011), the union of six datasets described earlier in Table 2. In all these training/testing splits, each fact language view $\mathbf{f}_l$ has corresponding tens of visual views $\mathbf{f}_v$ (i.e., images) split into training and test sets. So, each test image belongs to a fact that was seen by other images in the training set. We also performed comparison between the designed methods on the Large Scale Benchmark in Table 3. In this dataset, we randomly split all the annotations into an 80%-20% split, constructing sets of 647,746 $(\mathbf{f}_v, \mathbf{f}_l)$ training pairs (with 171,269 unique fact language views $\mathbf{f}_l$) and 168,691 $(\mathbf{f}_v, \mathbf{f}_l)$ testing pairs (with 58,417 unique $\mathbf{f}_l$), for a total of $(\mathbf{f}_v, \mathbf{f}_l)$ 816,436 pairs, 202,946 unique $\mathbf{f}_l$. Table 3 shows the coverage of different types of facts. There are 31,677 language view test facts that were unseen in the training set (851 <S>, 9,194 <S,P>, 21,673 <S,P,O>). The majority of the facts have only one example; see the supplementary.

Small and Mid-Scale Results Table 4 shows the performance of our SEM-B, SEM-A, and the designed baselines on these four datasets for both view retrieval tasks. We note that SEM-A works relatively better than SEM-B as the scale size increases as shown here when comparing results on Pascal dataset to larger datasets like Stanford40, and 6DS; see

Table 4: Bi-directional Retrieval Experiments

Dataset	Method	**i2f**	**f2i**
		Top1	**mAP**
Standord40	**SEM-A**	**74.46**	73
(40 facts)	**SEM-B**	71.22	**74.57**
(11 SP, 29 SPO)	**MV CCA IJCV14**	67.74	66.00
Chance = 2.5%	**ESZSL ICML15**	40.89	50.9
	TACL15 (COCO pretrained)	33.73	26.29
	TACL15 (retrained)	60.86	51.9
Pascal Actions	**SEM-A**	**74.760**	**80.950**
(10 facts)	**SEM-B**	74.080	80.530
(5 SP, 5 SPO)	**MV CCA IJCV14**	59.82	33.45
Chance = 10%	**ESZSL ICML15**	44.846	54.274
	TACL15 (COCO pretrained)	46.050	40.712
	TACL15 (retrained)	60.27	50.58
		Top1	**mAP/mAP100**
6DS	**SEM-A**	**69.63**	**34.86/ 50.68**
(186 facts)	**SEM-B**	68.94	34.64 / 47.87
(25 S, 20 SP, 141 SPO)	**MV CCA IJCV14**	29.84	23.93 / 36.44
Chance = 0.54%	**ESZSL ICML15**	27.53	30.7 / 47.58
	TACL15 (COCO pretrained)	15.71	9.37 / 15.88
	TACL15 (retrained)	26.13	26.17 / 40.4
		Top1/5/10	**mAP100**
Large Scale Dataset	**SEM-A**	**16.39 / 17.62 / 18.41**	0.96
Chance = 0.0017%	**SEM-B**	13.27 / 14.19 / 14.80	0.73
	MV CCA IJCV14	12.28 / 12.84 / 13.15	**1.0**
	ESZSL ICML15	5.80 / 5.84 / 5.86	0.4
	TACL15 (COCO pretrained)	3.48 / 3.48 / 3.5	0.021
	TACL15 (retrained)	5.87 / 6.06 / 6.15	0.29

Fig 4. In the next subsection, we show also that SEM-A is clearly better than SEM-B in the large scale setting. Our intuition behind this result is that SEM-A learns a different set of convolutional filters in the PO branch to understand action/attributes and interactions which is different from the filter bank learned to discriminate between different subjects for the S branch. In contrast, SEM-B is trained by optimizing one bank of filters for SPO altogether, which might conflict to optimize for both S and PO together; see Fig 3.

Compared to other methods on language view retrieval, we can see that both SEM-B and SEM-A perform significantly better than TACL15 (Kiros et al. 2015) even when retraining on our fact-level setting, especially on PASCAL10, Stanford40, and 6DS datasets which are dominated by SP and SPO facts; see Table 2. For visual view retrieval, performance is competitive in some of the datasets. We think the reason is due to the structure that makes our models relate all fact types by the visual modifiers notion. Although ESZSL is applicable in our setting, it is among the worst performing methods in Table 4. This could be because ESZSL is mainly designed for Zero-Shot Learning, but each fact has some training examples in these experiments. Interestingly, MV CCA is among the best.

Large Scale Benchmark Results Qualitative results of the large scale setting are shown in Fig. 5, 6 (with many more in the supplementary). In Fig. 5, our model's ability to generalize can be seen in the red facts. For example, for the leftmost image our model was able to correctly identify the image as <dog, riding, wave> despite that fact never being seen in our training data. The left images in Fig. 6 show the variety of images we can retrieve for the query <airplane, flying>. In the right images in Fig. 6, note how our model learns to visually distinguish gender ("man" versus "girl"), and group versus single. It can also correctly retrieve images for facts that were never seen in the training set (<girl, using, racket>). Highlighting the harshness of the performance metric, Fig. 6 also shows that <airplane, flying> has zero AP10 value giving us zero credit since the top images were just annotated as just an < airplane>.

The large scale results in Table 4 indicate that SEM-A is better than SEM-B for retrieval from both views, which is consistent with our medium scale results and our intuition. SEM-A is also multiple orders of magnitude better than chance and is also significantly better than the competing methods. To test the value of structure, we ran an experiment where we averaged the S, P, and O parts of the visual and language embedding vectors instead of keeping the structure. Removing the structure leads to a noticeable decrease in performance from 16.39% to 8.1% for the K1 metric; see Table 4 (Large Scale Dataset). The other datasets in Table 4 are orders of magnitudes smaller and also less challenging since all facts were seen during training. Figure 4 shows the effect of the scale on the Top1 performance for language view retrieval task (denoted K1). There is an observable increase on the improvement of SEM-A compared to the baselines in the large scale setting and performance of the image-caption similarity methods degrade substantially. We think this is due to both the large scale of the facts and

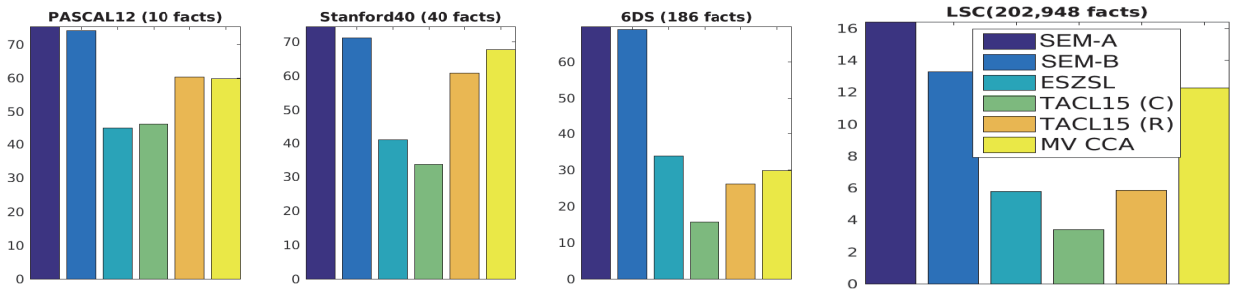

Figure 4: K1 Performance Across Different Datasets. These graphs show the advantage of the proposed models as the scale increases from left to right. (R) for TACL15 means the retrained version, (C) means COCO pretrained model

that the majority of the facts have zero or very few training examples. Interestingly, MV CCA is among the best performing methods in the large scale setting. However, SEM-A and SEM-B outperform MV CCA on both Top1 and Top 5 metrics; see Table 4. On the language view retrieval, SEM-A and MV CCA performs similarly.

Generalization It is desirable for a method to be able to generalize to understand an SPO interaction from training examples involving its components, even when there are zero or very few training examples for the exact SPO with all its parts S,P and O. Table 5 shows the K10 performance for SPOs where the number of training examples is ≤ 5. For example, the column **SP≥15, O≥15** means ≤ 5 examples of an SPO that has at least 15 examples for the SP part and for the O part. An example of this case is when we see zero or very few examples of <person, petting, horse>, but we see at least 15 examples of <person, petting, something=dog/cat/etc (not horse)> and at least 15 examples of something interacting with a horse <*,*, horse>. SEM-A performs the best in all the listed generalization cases in Table 4 and also in additional generalization cases in the supplementary. Figure 7 analyzes the Top10 large scale knowledge view retrieval (K10) results reported in Table 4 broken out by the number of images per fact (for SPO facts). These results show that SEM-A generally behaves better with compared other models with the increase of fact examples. Figures for other fact types are in the supp.

Figure 5: Language View Retrieval examples (i2f) (red means unseen facts during training)

Figure 6: Visual View Retrieval Examples (f2i))

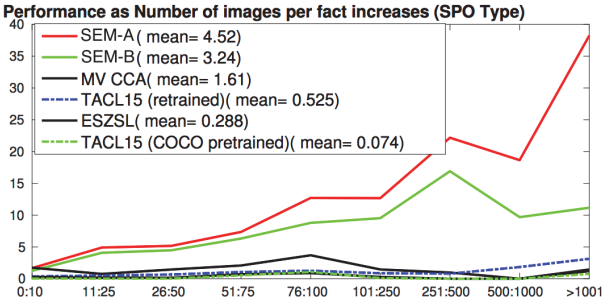

Figure 7: K10 Performance (y-axis) versus the number of images per fact (x-axis) for interactions (SPO) facts (other fact types are attached in the supplementary)

Conclusion

We introduce a problem setting for learning unbounded number of facts in images, which facilitates gaining visual knowledge. While studying this task, we consider Uniformity, Generalization, Scalability, Bi-directionality, and Structure. We investigated several baselines from multi-view learning literature, adapted to our setting. We proposed a structured embedding model that outperform the designed baselines mainly by the advantage of relating facts by structure. Our comparison to the baselines further show that the structure, we introduced in the convolutional lay-

Table 5: Generalization: SPO Facts of less than or equal 5 examples (K10 metric)

Cases	SP≥15, O≥15	SO≥15, P≥15	SO≥15, PO≥15	SO≥15, SP≥15
Number of Facts	**10605**	**4842**	**1755**	**3133**
SEM-A	**2.063**	**3.022**	**3.092**	**2.962**
SEM-B	1.751	1.961	1.645	2.097
ESZSL	0.149	0.098	0.041	0.038
TACL15 (COCO pretrained)	0.013	0.025	0.000	0.013
TACL15 (retrained)	0.367	0.473	0.543	0.586
MV CCA	1.221	1.462	1.786	1.109

ers (i.e.,branching in Fig 3(a)), enabled our model to learn in a data efficient way relatively and to generalize better on unseen/rarely seen facts (Table 5,Fig 7). A future interesting direction is to adopt dependency based word embedding like (Levy and Goldberg 2014) to improve the language representation and to incorporate information from ontologies like DOLCE (Gangemi et al. 2002) to enable reasoning about facts.

Acknowledgment. This work was also partially supported by an gift from Adobe Research and NSF-USA award # 1409683.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 819–826.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*.

Antol, S.; Zitnick, C. L.; and Parikh, D. 2014a. Zero-Shot Learning via Visual Abstraction. In *ECCV*.

Antol, S.; Zitnick, C. L.; and Parikh, D. 2014b. Zero-shot learning via visual abstraction. In *ECCV*.

Ba, J.; Swersky, K.; Fidler, S.; and Salakhutdinov, R. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*.

Chaplin, E. 2006. The convention of captioning: Wg sebald and the release of the captive image. *Visual Studies* 21(01):42–53.

Chen, C.-Y., and Grauman, K. 2014. Inferring analogous attributes. In *CVPR*.

Devlin, J.; Cheng, H.; Fang, H.; Gupta, S.; Deng, L.; He, X.; Zweig, G.; and Mitchell, M. 2015a. Language models for image captioning: The quirks and what works. In *ACL*.

Devlin, J.; Gupta, S.; Girshick, R.; Mitchell, M.; and Zitnick, C. L. 2015b. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.

Elhoseiny, M.; Cohen, S.; Chang, W.; Price, B.; and Elgammal, A. 2016a. Automatic annotation of structured facts in images. In *Vision and Language Workshop at ACL*.

Elhoseiny, M.; Liu, J.; Cheng, H.; Sawhney, H.; and Elgammal, A. 2016b. Zero-shot event detection by multimodal distributional semantic embedding of videos. In *AAAI*.

Elhoseiny, M.; Elgammal, A.; and Saleh, B. 2016. Write a classifier: Predicting visual classifiers from unstructured text descriptions. *TPAMI*.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *CVPR*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Gangemi, A.; Guarino, N.; Masolo, C.; Oltramari, A.; and Schneider, L. 2002. Sweetening ontologies with dolce. In *International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*.

Gkioxari, G., and Malik, J. 2015. Finding action tubes. In *CVPR*.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*.

Gupta, A. 2009. Sports Dataset. http://www.cs.cmu.edu/~abhinavg/Downloads.html. [Online; accessed 15-July-2015].

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *CVPR*.

Karpathy, A.; Joulin, A.; and Li, F. F. F. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*.

Kaufman, J. C.; Lee, J.; Baer, J.; and Lee, S. 2007. Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity* 2(2):96–106.

Kiros, R.; Salakhutdinov, R.; Zemel, R. S.; and et al. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.

Levy, O., and Goldberg, Y. 2014. Dependency-based word embeddings. In *ACL*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer.

Malinowski, M.; Rohrbach, M.; and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; and Yuille, A. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Muja, M., and Lowe, D. 2009. Flann-fast library for approximate nearest neighbors user manual. *Computer Science Department, University of British Columbia, Vancouver, BC, Canada*.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G. S.; and Dean, J. 2014. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *NIPS*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. *EMNLP*.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *NIPS*.

Romera-Paredes, B., and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *Proceedings of The 32nd International Conference on Machine Learning*, 2152–2161.

Sadeghi, M. A., and Farhadi, A. 2011. Recognition using visual phrases. In *CVPR*.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Socher, R.; Ganjoo, M.; Sridhar, H.; Bastani, O.; Manning, C. D.; and Ng, A. Y. 2013. Zero shot learning through cross-modal transfer. In *NIPS*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions.

Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2016. Order-embeddings of images and language. *ICLR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator.

Xu, K.; Ba, J.; Kiros, R.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015a. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Xu, R.; Xiong, C.; Chen, W.; and Corso, J. J. 2015b. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*.

Yao, B., and Fei-Fei, L. 2010. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*.

Yao, B.; Jiang, X.; Khosla, A.; Lin, A. L.; Guibas, L.; and Fei-Fei, L. 2011. Human action recognition by learning bases of action attributes and parts. In *ICCV*.

Zhang, N.; Paluri, M.; Ranzato, M.; Darrell, T.; and Bourdev, L. 2014. Panda: Pose aligned networks for deep attribute modeling. In *CVPR*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *NIPS*.