Few-shot Linguistic Grounding of Visual Attributes and Relations using Gaussian Kernels

Daniel Koudouna and Kasim Terzić
School of Computer Science, University of St Andrews, U.K.

Keywords: Few-shot Learning, Learning Models, Attribute Learning, Relation Learning, Scene Understanding.

Abstract: Understanding complex visual scenes is one of fundamental problems in computer vision, but learning in this domain is challenging due to the inherent richness of the visual world and the vast number of possible scene configurations. Current state of the art approaches to scene understanding often employ deep networks which require large and densely annotated datasets. This goes against the seemingly intuitive learning abilities of humans and our ability to generalise from few examples to unseen situations. In this paper, we propose a unified framework for learning visual representation of words denoting attributes such as "blue" and relations such as "left of" based on Gaussian models operating in a simple, unified feature space. The strength of our model is that it only requires a small number of weak annotations and is able to generalize easily to unseen situations such as recognizing object relations in unusual configurations. We demonstrate the effectiveness of our model on the predicate detection task. Our model is able to outperform the state of the art on this task in both the normal and zero-shot scenarios, while training on a dataset an order of magnitude smaller.

1 INTRODUCTION

The task of scene understanding is typically understood to involve reasoning beyond the level of objects, and the ability to express complex attributes and relations between objects. This paves the way for an automatic, natural language description of a scene, or answering questions about it. In a world where we increasingly use language to communicate with artificial agents (such as cars, assistive robots and smart homes), it is essential for such agents to develop a rich understanding of everyday visual scenes in order to effectively communicate with users. Furthermore, it is crucial that such agents can update their representations easily in order to adapt to changing circumstances and environments. For example, an assistive robot helping a patient in their home is operating in an environment where meaning of words such as "fast" or "close to" can change over time, and will need to adapt to such changes based on few examples.

The past decade has seen huge improvements in terms of object detection performance, but the wider task of scene understanding remains difficult. Datasets such as GQA (Hudson and Manning, 2019) have attempted to further the field of visual reasoning, but the need for meticulously annotated datasets goes against the seemingly intuitive learning model used by children when acquiring language, with their ability to generalize learned concepts to unseen situations. Even with the emergence of large annotated datasets, there still exists the problem of the vast number of possible configurations of objects, attributes and relations in a natural scene. This leads to a combinatorial explosion for which it may be infeasible to collect sufficient examples for each sample. Novel algorithms and models are needed that can generalize, like humans do, from observed scenes to new unseen configurations and recent work has been shifting back to parsing language as a way to achieve this (Johnson et al., 2017).

Many recent approaches do not attempt to ground each word individually but focus on learning combinations of words, typically formulated as a joint feature learning problem on triples of the type (subject-object-predicate). This allows them to build on powerful feature extraction abilities of modern deep networks, but requires large annotated datasets and makes it harder to transfer knowledge between predicates (Peyre et al., 2017).

We propose a novel approach which parses statements in a formal language and attempts to sequentially ground each word separately onto a common feature space. This formalism allows us to express objects (such as "cat"), attributes (such as "red") and

Koudouna, D. and Terzić, K.
Few-shot Linguistic Grounding of Visual Attributes and Relations using Gaussian Kernels.
DOI: 10.5220/0010261301460156
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 146-156
ISBN: 978-989-758-488-6

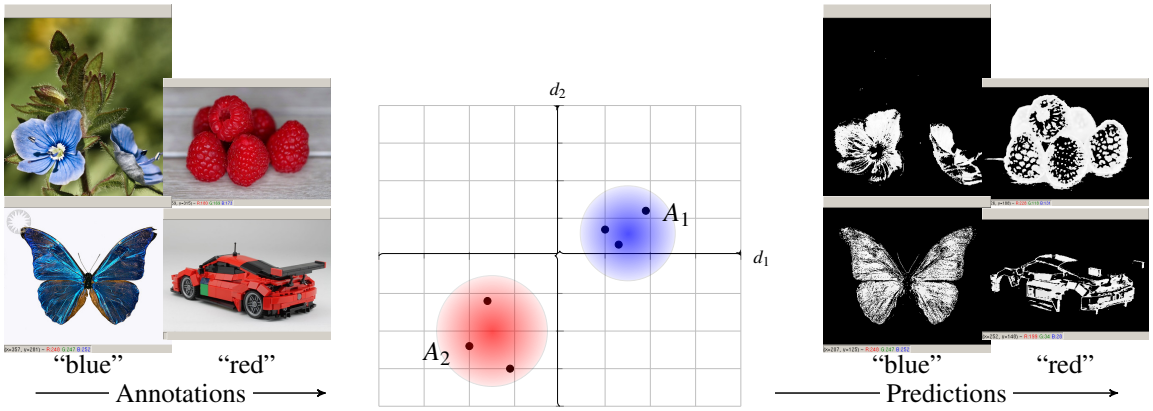

Figure 1: High-level overview of the model. Representing each pixel in the region of interest of the image as a feature vector, we estimate the probability distribution of the annotated attribute or relation in the embedding space . The model is then able to predict the degree of each attribute at each pixel.

relations (such as "left of") as functions operating on the same feature space. Although our model is far simpler than existing methods, we demonstrate that it can learn attributes (such as "red") and relations (such as "above") from a very small number of images and that they can easily transfer to unseen examples (zero-shot learning).

This paper presents the following contributions: 1. a process for mapping attributes and relations from language onto functions in a single, simple and interpretable feature space, using Gaussian models, 2. a formalism based on lambda calculus which ties both natural language descriptions into a series of such that are executed sequentially on a scene, and 3. a self-guided learning approach that can learn these mappings from very few examples.

We purposefully keep both image features and our model very simple to show that the power of the system does not come from deeply learned features or complex mathematical models, but from the modelling of each visual attribute and relation independently, and from the integration with language which allows for sequential parsing of the scene. Our feature space and functions used to represent concepts can both be extended in the future but we want to show that even at its simplest, this new model can match the state of the art on some specific tasks and thus presents a promising research direction. We demonstrate this on the predicate detection task on the VRD dataset, where we outperform the state of the art while using a vastly smaller number of training examples than other models. In addition, we present some qualitative results on learning attributes from random images queried from a search engine.

2 RELATED WORK

The relation between metric spaces and learned features has been explored in the context of **metric learning** and applied to the detection of attributes in the context of image classification (Zhang et al., 2019), object detection (Karlinsky et al., 2019; Sohn, 2016), and intra-class comparison (Kovashka et al., 2015). Our approach to attribute and relation learning is compatible with learned feature spaces but in this paper we use a simple hand-crafted feature space to demonstrate general applicability of our model.

Few-shot learning requires vision systems to learn novel categories from a small number of examples, inspired by the learning capabilities that humans, especially children, exhibit in language learning (Carey and Bartlett, 1978; Xu and Tenenbaum, 2000). Few-shot learning has been successful in the context of classification, which focuses on models trained on traditional dataset which can then learn new classes from one or few examples (Chen et al., 2019; He et al., 2019; Karlinsky et al., 2019; Zhang et al., 2019). We show that our model can learn generalizable attributes and relations with very few examples without requiring a pre-existing learned model, which more closely resembles human learning.

Relation learning is the task of learning and detecting the relation between two objects in an image. Some traditional AI approaches combined hand-modelled spatial relations with probabilistic scene models (Neumann and Terzić, 2010; Kreutzmann et al., 2009) or constraints (Hotz et al., 2007) but did not learn relations from annotated examples. Many recent approaches have used a combination of visual and language models to jointly learn triples in the

form (subject-object-predicate). However, due to the number of possible object classes and relations, there has been a recent focus on learning relations more independently of the participating objects, such as detecting unusual or previously unseen relations (Peyre et al., 2017; Peyre et al., 2019), by driving attention to regions of an image (Krishna et al., 2018; Luo and Shakhnarovich, 2017), or by the explicit modelling of executing functions (Andreas et al., 2016; Johnson et al., 2017). Our approach is aligned with this direction of research, but we show that many relations can be successfully learnt using a simple language model and without powerful deep networks.

Language grounding and **top-down processing** in visual tasks are long-standing topics in cognitive science. We adopt an approach similar to classical models such as BISHOP (Gorniak and Roy, 2004), where a formal grammar defines a series of functions executed on the scene. Unlike BISHOP, our model is able to adopt a growing lexicon of learned words representing attributes and relations using a grammar, extending the idea of scene description using a series of interpretable feature maps (Richter et al., 2014). This is in contrast with many modern approaches, which often rely on word vectors (Lu et al., 2016; Peyre et al., 2019) or generic embedding spaces (Krishna et al., 2018; Surís et al., 2019) to disambiguate language in relationship detection. Our work follows the recent interest in learning individual words or phrases from visual stimulus (Bisk et al., 2020; Jin et al., 2020). In addition, our model is able to learn in a continual manner, expanding its vocabulary as new words are introduced in a series of scenes.

3 MODEL

Our model operates on a common feature space for evaluating both attributes and relations. We define $D_1, D_2, ...D_n$ as the dimensions of features extracted from a scene. We represent each pixel in a $h \times w$ image as a vector corresponding to its value in each feature dimension D_i as $\boldsymbol{p}$. We also define a *field* F as a $h \times w$ matrix which stores a scalar value for each pixel in a scene. In our model, fields represent results of operations, similar to retinotopic maps in neural models.

3.1 Language

We adopt a language model in which each word of a sentence is mapped into a function which is executed on the feature space representing the image. We model the sentences as expressions in first order logic. Objects in the scene are modelled as terms in the logic. Object classes and attributes are modelled as unary predicates, e.g $\mathrm{Cat}(x)$ indicates object x is a cat. Combination of attributes is possible through conjunction of predicates. For example, *"there is a red cat"* is parsed as

$$\exists x \, \mathrm{Cat}(x) \wedge \mathrm{Red}(x) \tag{1}$$

Relations in the scene are modelled in terms of event semantics (Parsons, 1991). A relation is modelled as an existence of an event e with a landmark lm representing the object of the event, and a trajector tr representing the subject. We denote the relation and the landmark as a single predicate, e.g $\mathrm{Left\_of}(e,x)$ indicates a relation *"left of"* with object x as the landmark. We denote the trajector of a relation as $tr(e,y)$. For example, "a dog left of a cat" is parsed as

$$\begin{aligned} \exists e \exists x \exists y \, \mathrm{Left\_of}(e,x) \wedge \mathrm{Cat}(x) \\ \wedge tr(e,y) \wedge \mathrm{Dog}(y) \end{aligned} \tag{2}$$

First-order logic allows us to use conjunctions (*"and"*) and disjunctions (*"or"*) to parse arbitrarily complex sentences. This includes conjunction of attributes (*"the dog is big and black"*) and relations (*"the dog is next to the cat and the person"*). We omit the logical representations of these examples for brevity.

We explicitly model response maps of particular queries as partial applications of expressions in lambda calculus. Formally, existence of an object of a particular class is represented by an object-class function O:

$$\lambda \, Q \, \exists z \, O(z) \wedge Q(z) \tag{3}$$

In the case of a subject of a relation, we add the tr predicate as explained above:

$$\lambda \, Q \, \exists z \, \exists e \, O(z) \wedge tr(e,z) \wedge Q(z) \tag{4}$$

This allows response maps for attributes and relations to be modelled as lambda expressions. An attribute is represented by an attribute function A:

$$\lambda \, y \, A(y) \tag{5}$$

and a relation by a relation function R:

$$\lambda \, x \, \lambda \, y \, \exists \, e \, R(e,x) \tag{6}$$

where x and y represent objects, and e represents a particular event, which is used to group together the landmark and the trajector of a particular relation.

Each word in the vocabulary of the model will take the form of an expression in Equations 3, 4, 5 and 6. In this logic, we use terms to represent either objects or events. Therefore, the model only requires

Figure 2: Example of evaluating relation queries on an image. Left: An image from the COCO dataset. Centre: Evaluating *"left of the cat"* Right: Evaluating *"left of the dog"*. Response maps are overlaid with the original image for clarity.

definitions for all O, A and R predicates, which can be supplied either *a priori*, or learned by the model. For this paper, we consider O as supplied *a priori*, and focus on the learning of attributes and relations.

3.2 Objects

We model a term o in first-order logic by a point in our feature space $\boldsymbol{p}_o$. Conceptually, an object o, represented as a term in the logic, is obtained by a field in our model, where the response value for each pixel p indicates the confidence for the participation of the pixel to the particular object. Common representations for obtaining such fields are bounding boxes and segmentation masks obtained from an appropriate detector. This allows us to describe not only object classes obtained from an object detector, e.g *"cat"*, but arbitrarily complex noun phrases, i.e *"the red cat"*, *"the red cat under the clock"*.

In order to tie the object representation to language, we collapse an object to a single point $\boldsymbol{p}_o$, by considering the object field F_o. For each feature dimension D_i, the value of the object point is the mean value of D_i weighted by the object field. This represents that object regions of lower confidence should not contribute to the value, and non-object regions (i.e zero-valued pixels in F_o) should not contribute at all.

This operation is equivalent to taking the matrix inner product between F_o and D_i. Therefore, $\boldsymbol{p}_o$ is a vector containing the matrix inner product between F_o and each feature dimension:

$$\boldsymbol{p}_o = (< F_o, D_1 >, < F_o, D_2 >, \ldots < F_o, D_n >) \quad (7)$$

This product then represents the weighted mean feature of the object – mean position, mean colour, mean texture, etc. which form the basis for calculating attributes and relations as described previously. Here we note that this mean representation of an object is not used to detect or classify objects, but to provide a reference in our feature space to identify a particular object in the scene. This reference is then used as input to other functions operating on the space.

3.3 Attributes

An attribute is defined by an attribute function A which operates on a point in the feature space:

$$A : \mathbb{R}^n \rightarrow \mathbb{R}\,, \quad (8)$$

where n is the dimensionality of the feature space. As shown in Equation 5, this conveniently allows evaluation of an attribute on a single object represented by mean value $\boldsymbol{p}_o$. Note that we can also apply the attribute function to the entire $h \times w$ image at once, which will produce a field $F \in \mathbb{R}^{h \times w}$:

$$A : \mathbb{R}^{n \times h \times w} \rightarrow \mathbb{R}^{h \times w}\,. \quad (9)$$

Importantly, this means that evaluating an attribute on an image produces a field in the space. Figure 1 shows application of different attributes on entire images.

This formalism leads to, a flexible model which allows for attributes originating either from external knowledge, or from any employed learning technique. We describe a simple learning model for attributes in Section 4.3.

In addition, this formalism allows for two different types of evaluation of an attribute; the application of an attribute on an object $\boldsymbol{p}_o$, as above, and the partial evaluation of the attribute, without a participating object, on a particular scene. Extracting features from an image maps each pixel to a feature vector $\boldsymbol{p}$, which allows the evaluation of an attribute function on an image to create a response map, as shown in Figure 1.

The partial application of an attribute allows for top-down processing of a scene, enabling the model to drive its attention to a particular region of an image (e.g. *"dark"* or *"smooth"* regions). It also allows evaluating vague statements such as *"red left of green"* without having to detect objects and process them in turn, which is done in a bottom-up manner. In these cases, we convert the field representation of the attributes (i.e *"red"*) into an object as described in Section 3.2.

Finally, to determine whether a particular object has attribute A, we use the decision function:

$$t(\boldsymbol{p}_o;A) = \begin{cases} \text{true}, & \text{if} \quad A(\boldsymbol{p}_o) > T \\ \text{false}, & \text{otherwise.} \end{cases} \tag{10}$$

where T is a variable detection threshold.

3.4 Relations

Following Equation 6, a relation naturally extends an attribute by producing a field given two points instead of one. In order to describe relations, we augment the feature space with respect to a fixed point $\boldsymbol{lm}$ representing the landmark of the relation. For each dimension D, we consider a additional dimensions. For this paper, we use:

$$\begin{aligned} D_{\Delta}(\boldsymbol{lm};D) &= D - \boldsymbol{lm}_D \\ D_{\Delta^2}(\boldsymbol{lm};D) &= (D - \boldsymbol{lm}_D)^2 \end{aligned} \tag{11}$$

Formally, for a additional dimensions for each dimension, a relation can be represented as:

$$R : \mathbb{R}^{a \cdot n} \times \mathbb{R}^n \rightarrow \mathbb{R} \tag{12}$$

A property shared with attributes is the ability to partially apply relations. Provided two objects, the relation function can provide a value representing the intensity of the relation. With only a single object, the relation can be applied to give a value for each pixel in the scene, creating response maps such as in Figures 2 and 4.

This concept can be illustrated in Figure 2. If one of the feature dimensions is the x coordinate of the object represented by $\boldsymbol{p}_o$, then the two relative dimensions in Eq.11 will be the horizontal offset of $\boldsymbol{p}_o$ from the landmark and the square of the horizontal offset, respectively. The horizontal offset makes it trivial to learn the relation "left-of" with respect to the dog and the cat as landmarks. Adding the y coordinate to the features would also add the corresponding vertical offset and its square, allowing us to easily learn "above", "below", as well as distance-based concepts such as "near" or "far".

3.5 Learning Approach

In this paper, we model each attribute and relation function as a Gaussian distribution in a shared feature space. This follows the idea of representing concepts as convex regions in a conceptual space (Gärdenfors, 2000). We use this model instead of a deep network to demonstrate that the power of our model lies in the formalism used to sequentially parse the scene, and not the complexity of the architecture. However, there is nothing preventing the use of powerful learned features in the future. We use a set of domains, each of which containing a number of dimensions. The set of domains used in this paper are as follows:

- the pixel colour in the CIELAB colour space,
- horizontal and vertical offset (in pixels) from image centre,
- the magnitude and angle of optical flow, and
- the responses of oriented Gabor filters covering a range of sizes and orientations.

In addition, we use the difference and square difference from a particular landmark as augmented features when considering relations, as described in Section 3.4.

The features we use are well-established in the field of computer vision and all map reasonably well to semantic concepts describing colours, spatial relations, motion, and texture. When operating on still images, the dimensions of optical flow are fixed at zero. Our learning approach described below will learn to ignore dimensions that are not useful for representing specific attributes and relations.

In all cases, we model both the word W and environment E as two multivariate normal distributions over the feature space:

$$\begin{aligned} W &\sim \mathcal{N}(\mu_w, \Sigma_w) \\ E &\sim \mathcal{N}(\mu_e, \Sigma_e) \end{aligned} \tag{13}$$

In this paper we assume independence of the features in the space, meaning Σ is diagonal, but this is not required by the model.

For a training example, we split the image into two sets of points by considering the participating object. For example, objects obtained by a set of pixels B from a bounding box produce a field using the function:

$$F(p;B) = \begin{cases} 1, & \text{if} \quad p \in B \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

We assign the points of the object to W and the remainder to E. The environment samples are taken from any annotation that does not define the word and allows us to learn without annotating negative examples.

We then construct a field function F which accepts a single point $\boldsymbol{p}$:

$$F(\boldsymbol{p};W,E) = \frac{1}{k}\, exp\left(\left[\sum_i \frac{|\boldsymbol{p}_i - \boldsymbol{\mu}_i|}{\boldsymbol{\sigma}_i} D_{KL}(W_i, E_i)\right]^2\right) \tag{15}$$

where D_{KL} represents the KL-divergence between two normal distributions W_i and E_i representing the distributions for feature dimension i as:

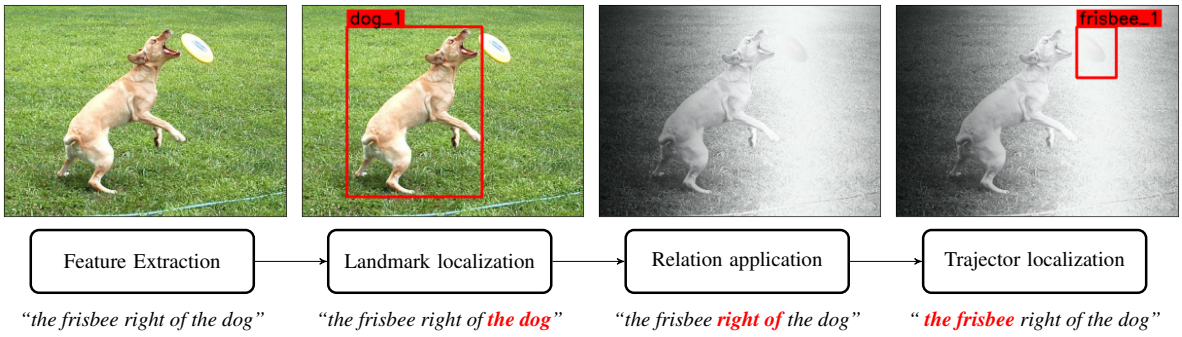

Figure 3: Example of the evaluation process for a single relation *""the frisbee right of the dog"*. First, the model detects the appropriate landmark *"dog"*, and generates the response map for the relation *"right of"* using the detected object as the landmark. Finally, the model detects the trajector *"frisbee"* and calculates the mean value of the response map of the bounding box.

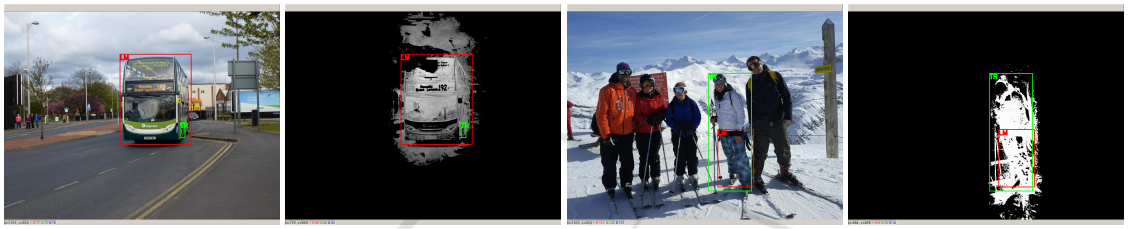

Figure 4: Examples of attention maps from the VRD dataset. Left: *"on the bus"* . Right: *"wears the pants"*. The maps show the partial application of the relation, i.e *"on"* to the landmark, i.e *"the bus"* on the left. Calculation of the relation uses the mean response of the bounding box of the trajector, i.e *"wheel"*.

$$D_{KL}(P,Q) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \quad (16)$$

and k normalizes the value in the range $[0,1)$.

3.5.1 Learning Attributes

A training example for an attribute consists of a set of points obtained either from a bounding box of an object or a segmentation mask, and a label representing the name of the attribute. The model then updates its knowledge base and calculates the new Gaussian A for the attribute, as discussed in Section 3.3. The process is shown in Algorithm 1. Line 1 extracts the image features to assign a feature vector to each pixel in the image, collectively defined as P. Lines 2 and 3 split these vectors into the feature vectors of the object and the environment respectively. Finally, lines 4 to 7 add the data to the model and retrain any words with the updated data.

3.5.2 Learning Relations

For a relation, two sets of points are needed, for the landmark and the trajector. The model then calculates the augmented features relative to the landmark, as discussed in Section 3.4, and updates its knowledge base, as shown in Algorithm 2. In this case, the dimensionality of the feature space used is $a \times n$, where a is the number of feature augmentations and n is the size of the original feature space. As shown, this method allows for the insertion of an arbitrary numbers of words in the lexicon, without affecting the existing words. In addition, it allows attributes and relations to be stored with different dimensionalities, while being derived from the same base feature space.

Algorithm 1: Training an attribute.

Input: Labelled image I with object o
Input: Word w (attribute name)
1 $P \leftarrow$ extract_features(I)
2 $P_o \leftarrow$ get_object_points(P, o)
3 $P_e \leftarrow P - P_o$
4 add_samples(w, P_o)
5 add_environment_samples(P_e)
6 **for** w *in lexicon* **do**
7 recalculate_distribution(w)

4 EVALUATION

We evaluate our model on the predicate detection task of the Visual Relationship Dataset (VRD) (Lu et al., 2016). Contrary to most modern approaches which

Algorithm 2: Training a relation.

Input: Labelled image I with landmark lm and trajector tr
Input: Word w (relation name)
$P \leftarrow$ extract_features(I, lm)
// P contains augmented features
$P_{tr} \leftarrow$ get_object_points(P, tr)
$P_e \leftarrow P - P_{tr}$
add_samples(w, P_{tr})
add_environment_samples(P_e)
for w *in lexicon* **do**
    recalculate_distribution(w)

use the entire training set, we randomly select a number of examples for each predicate. We demonstrate that our model can still outperform the state-of-the-art in the predicate detection task. In addition, we are able to also outperform in the zero-shot portion of the test set, which means a relation was observed before, but the subject-predicate-object triple it is being tested on was not. We also show qualitative querying on the COCO dataset to demonstrate the generalizability of the model. Finally, we demonstrate learning of simple attributes by querying images from a search engine.

4.1 Relationship Detection on VRD

The VRD predicate detection task tries to learn and predict the predicate between two annotated objects in a scene (represented by ground truth bounding boxes). We focus on this task as our model is not tied to a particular object detection model, as explained in Section 3.2. Most approaches use a train/test split of 4000/1000 images, which roughly translates to 30000 triples in the training set, and 7000 in the test set. We compare our model to a number of methods: 1. LP introduces the dataset, and uses a language module utilizing word embeddings along with visual features. LP-V is a diagnostic using only visual features (Lu et al., 2016) 2. DR-Net utilizes spatial configurations, as well as a probabilistic model for each triplet (Dai et al., 2017) 3. LK uses external linguistic knowledge to better predict probabilities of each predicate given the subject and object (Yu et al., 2017). 4. SR fuses all features and uses structural ranking loss (Liang et al., 2017). 5. NL uses a bi-directional RNN and relative spatial features of the subject-object pair to predict predicates (Liao et al., 2019).

Following standard procedure for the predicate detection task, we use the ground-truth bounding boxes of the subject and object without the use of an object detector.

Table 1 shows the results. We significantly outperform traditional approaches using a training set an order of magnitude smaller ($n = 8$ examples for each relation, with some relations having fewer annotated examples in the training set). In addition, we are able to achieve state-of-the-art on the subset of unseen triples in the VRD test set. This demonstrates that our model generalizes well, since we learn on very few examples yet achieve excellent results in the zero-shot subset.

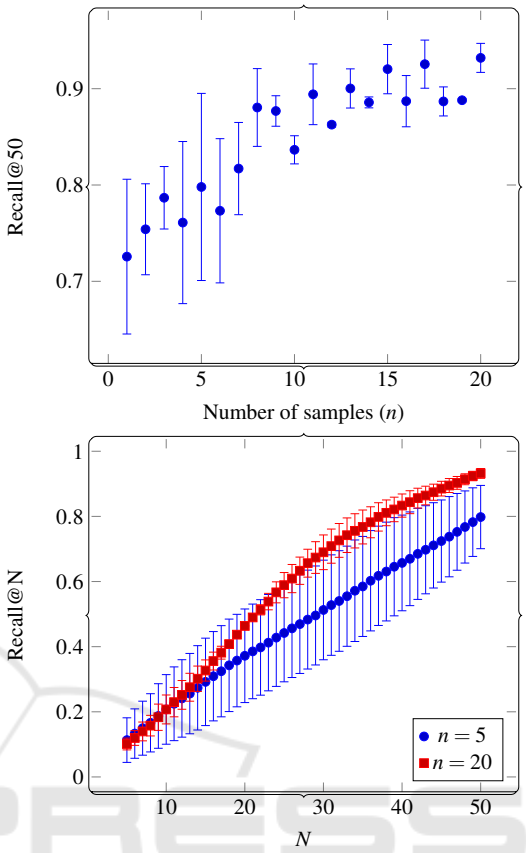

Figure 5: Left: Graph of increasing sample size n against mean Recall@50 on the VRD test set, with error bars. Right: Graph of the effect of increasing sample size n against mean Recall@N, with error bars.

Figure 5 shows the effect of increasing sample size for each relation on the VRD test set. Notably, the number of samples represents the maximum number, as some relations have ≤ 20 training examples. As expected, lower sample sizes show very high variance due to the random nature of the selection, as well as the fact that the model might not be able to learn some relations from a single example, such as symmetric relations. However, the variance is quickly reduced at higher sample sizes. In addition, Recall@N is shown for sample sizes $n = 5$ and $n = 20$. The model exhibits much higher variance at a lower number of samples, with a significant increase in performance as n increases.

We attribute the success of the model to two fac-

Table 1: Results on predicate and relationship detection in the VRD (Lu et al., 2016). We compare our model with similar results discussed in Section 4.1. Results are compared, where applicable, to the zero-shot learning scenario.

Comparison	Predicate Detection		
	Entire set R@50	Zero-shot R@50	Training size
LP-V (Lu et al., 2016)	7.11	-	30000
LP (Lu et al., 2016)	47.87	8.45	30000
DR-Net (Dai et al., 2017)	80.87	-	30000
LK (Yu et al., 2017)	85.64	54.20	30000
SR (Liang et al., 2018)	86.01	60.90	30000
NL (Liao et al., 2019)	84.39	80.75	30000
Ours	**93.67**	**89.55**	**560**

tors. First, the model only considers objects as points in the feature space, as described in Section 3.2, and so is able to easily generalize spatial relations between triples, either seen or unseen. As described, we use the feature vector of the object $\boldsymbol{p}_o$ to effectively translate the feature space. This allows us to easily adapt to symmetric and anti-symmetric relations, as we learn the average translation of the space by considering the objects in our knowledge base. This also helps to explain why our model achieves a very high recall at only a small number of examples, as shown in 5.

Secondly, we are able to keep the predicates independent of each other, since we learn a different distribution for each. This allows us to extend the model to an arbitrary number of predicates without affecting the performance of other predicates, which is particularly helpful when considering learning with very few examples.

4.2 Image Retrieval

In order to further demonstrate the generalizability of our model, we use the COCO dataset to carry out content-based image retrieval. We use the same model trained on the VRD dataset as described in Section 4.1. In addition, we use an off-the-shelf pretrained YOLO object detector (Redmon and Farhadi, 2018) to provide bounding boxes. We are able to query the COCO dataset for images containing a particular relation, as shown in Figure 6. As shown, the model then ranks images in the dataset according to the confidence in the existence of the relation. We demonstrate the strength of the model to handle complex language by combining two queries with *"and"*. As discussed in Section 3.1, we use a simple grammar to parse a sentence in natural language and convert it into a series of operations on the feature space, as shown in Figure 3.

4.3 Online Attribute Learning

In order to demonstrate the learning of attributes, we use the Flickr API to obtain a small set of weakly-annotated training images for a variety of attributes. Some examples are shown in Figures 1 and 7. We treat the centre of the image as an object possessing the desired property, by treating the whole image as an object o using a Gaussian centered on the image as a response field. This bypasses the need for an object detector, and leverages the fact that, on average, an attribute will be most represented in the centre of the image.

Figure 8 shows the rate of convergence of our model, by considering the shift in the centre of the attribute distribution that each additional sample causes. As expected, our model shows the highest change and variance at the first additional sample, while converging at a very low sample number ($n > 10$).

We also experimented with learning the same attributes by scaling down original images before training. Figure 8 shows the average resolution of the training data. We scale the input data by a factor of 4 and 16. As shown, the model is able to converge just as fast with orders of magnitude fewer data points, suggesting that the model learns not from an abundance of similar points, but by slowly differentiating the features of the centre of the images possessing the attribute, from the rest of the environment, as discussed in Section 3.5.

Figure 6: Examples of content-based image retrieval on the COCO dataset. The model trained on VRD is presented a textual query, and ranks the images in the dataset by the confidence of the statement being true in the image. The top 5 results for each query are shown, sorted by confidence from left to right.

Figure 7: Examples of attention maps from the Flickr API. Top: *"purple"* Bottom: *"green"*.

5 DISCUSSION

5.1 Learning approach Advantages

One of the main advantages of our model is its flexibility in the vocabulary employed. Since the model does not have a fixed lexicon, it can learn either from a large annotated dataset or in an online manner. Whenever it encounters a word with no model (i.e. no Gaussian function to link it to the feature space), it can look for annotated examples which include that word. When using a dataset, unlike neural network approaches, our model does not iterate over the data in training epochs, but obtains new vocabulary or refines its existing knowledge base by considering each data point in the training set in sequence.

Another major benefit of this approach is the lack of explicit feature selection. Traditional neural network models rely on back-propagation to emphasize the importance of a particular feature. In contrast, our model will naturally learn that only a few of the feature dimensions will be important for a particular attribute (e.g colour for 'green', but position for 'left') by considering the variance of the strength of the feature in the training data, as discussed in Section 3.5. This property will exhibit a lower variance (i.e higher concentration of data points) in the salient feature dimensions, and a higher variance in the non-salient features, relative to the environment.

Our approach can also distinguish between symmetric and anti-symmetric relations. Most visual relations between objects can be described as symmetric *"A close to B"* $\implies$ *"B close to A"* or anti-symmetric *"A left of B"* $\implies$ $\neg$*"B left of A"*. For example, the symmetric relation "close to" exhibits a spread of values across the regular horizontal and vertical dimensions, as objects may be close on both sides, but a higher concentration of values in the respective squared dimensions, since objects close to each other are usually a fixed distance apart.

5.2 Introspection and Meta-querying

Due to the statistical nature of our model, and the fact that our feature dimensions are relatively simple, we are able to answer questions about learned linguistic terms. As mentioned in Section 3.3, we use the KL-Divergence in order to select the features in our space which exhibit the most information gain relative to the environment. This, combined with the fact that each predicate is learned independently, naturally allows the model to describe which dimension is most salient for each predicate, as shown in Figure 9. This

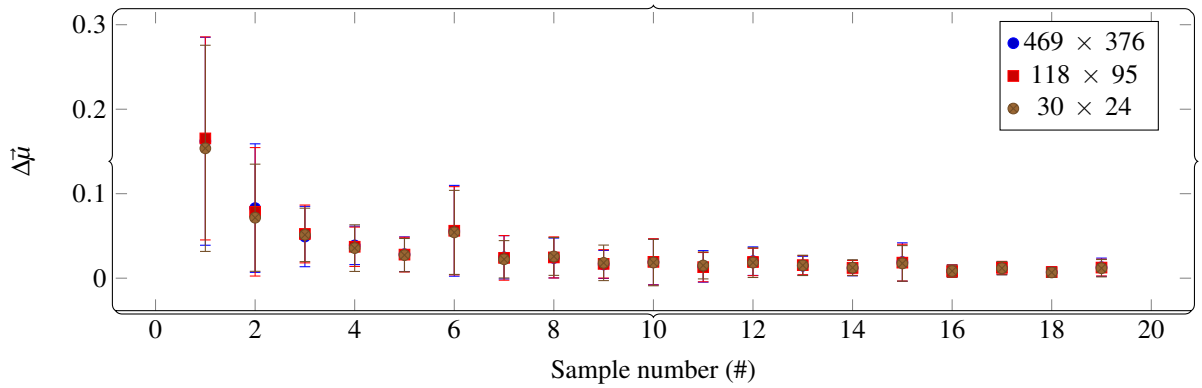

Figure 8: Graph of average deviation in the attribute distribution mean $\vec{\mu}$ against sample number for $n = 5$ attributes, for various image resolutions. Attributes are learned as described in Section 4.3. The graph shows the convergence of the (averaged) mean of the distributions as more data is added to the model's knowledge base.

```
red is related to 'dim_2' :
{'dim_1': 0.117, 'dim_2': 0.861,
 'dim_3': 0.378}
purple is related to 'dim_3' :
{'dim_1': 0.0679, 'dim_2': 0.365,
 'dim_3': 0.534}
yellow is related to 'dim_3' :
{'dim_1': 0.335, 'dim_2': 0.387,
 'dim_3': 1.008}
```

Figure 9: Example output from learned concepts from images queried from a search engine, discussed in Section 4.3. Dimensions 1 to 3 represent the components of LAB space, in order.

is particularly powerful when feature dimensions correlate well with semantic concepts as they do in our case: e.g. the L dimension of the CIELAB space corresponds to intensity/lightness. This allows the evaluation of simple queries related to learned attributes such as 'what is dark?' with generated answers such as 'dark is related to lightness'. This can be evaluated at any time, using the accumulated knowledge of the model. This can either be a collection of images, when applied to a general task such as a dataset, or a changing environment over time, when applied to a real-world scenario such as a surveillance camera.

6 CONCLUSION

We have presented a framework for grounding of attributes and visual relations and a model built on this framework. Our model allows the learning of attributes and relations in the same way, using Gaussian distributions in a feature space of common image features. Importantly, we use a single model for learning both attributes and relations. The novelty of our contribution does not come from the use of Gaussian models, but from our approach to parsing queries and sequentially processing the scene. We use the same formalism for grounding objects, attributes, and relations, in contrast to most current work on relation and attribute learning.

Our system outperforms the state-of-the-art on the task of predicate detection on the VRD dataset, and generalizes well in a zero-shot learning task. In addition, we show that our approach is able to learn attributes in an online manner, by using images retrieved by a search engine. We feel that these are promising early results for our formalism and plan to explore deep learning of conceptual spaces to improve performance in the future.

ACKNOWLEDGEMENTS

The Titan Xp GPU used for this research was donated by the NVIDIA Corporation.

REFERENCES

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. (2016). Learning to compose neural networks for question answering. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., and Turian, J. P. (2020). Experience grounds language. *ArXiv*, abs/2004.10151.

Carey, S. and Bartlett, E. (1978). Acquiring a single new word.

Chen, Z., Fu, Y., Zhang, Y., Jiang, Y., Xue, X., and Sigal, L. (2019). Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing*, pages 1–1.

Dai, B., Zhang, Y., and Lin, D. (2017). Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gorniak, P. and Roy, D. (2004). Grounded semantic composition for visual scenes. *The Journal of Artificial Intelligence Research*, 21:429–470.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press, Cambridge, MA, USA.

He, X., Qiao, P., Dou, Y., and Niu, X. (2019). Spatial attention network for few-shot learning. In *ICANN 2019: Deep Learning*, pages 567–578, Cham. Springer International Publishing.

Hotz, L., Neumann, B., Terzić, K., and Šochman, J. (2007). Feedback between low-level and high-level image processing. Technical Report Report FBI-HH-B-278/07, University of Hamburg, Hamburg.

Hudson, D. A. and Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jin, X., Du, J., Sadhu, A., Nevatia, R., and Ren, X. (2020). Visually grounded continual learning of compositional phrases.

Johnson, J., Hariharan, B., Van Der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2017). Inferring and executing programs for visual reasoning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3008–3017.

Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., and Bronstein, A. M. (2019). Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kovashka, A., Parikh, D., and Grauman, K. (2015). Whittlesearch: Interactive image search with relative attribute feedback. In *International Journal of Computer Vision (IJCV)*.

Kreutzmann, A., Terzić, K., and Neumann, B. (2009). Context-aware classification for incremental scene interpretation. In *Workshop on Use of Context in Vision Processing*, Boston.

Krishna, R., Chami, I., Bernstein, M., and Fei-Fei, L. (2018). Referring relationships. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Liang, K., Guo, Y., Chang, H., and Chen, X. (2018). Visual relationship detection with deep structural ranking. In *AAAI*.

Liang, X., Lee, L., and Xing, E. P. (2017). Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, pages 4408–4417.

Liao, W., Rosenhahn, B., Shuai, L., and Ying Yang, M. (2019). Natural language guided visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Lu, C., Krishna, R., Bernstein, M., and Fei-Fei, L. (2016). Visual relationship detection with language priors. In *ECCV*.

Luo, R. and Shakhnarovich, G. (2017). Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Neumann, B. and Terzić, K. (2010). Context-based probabilistic scene interpretation. In *Proc. Third IFIP Int. Conf. on Artificial Intelligence in Theory and Practice*, pages 155–164, Brisbane.

Parsons, T. (1991). *Events in the Semantics of English*.

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

Peyre, J., Laptev, I., Schmid, C., and Sivic, J. (2019). Detecting unseen visual relations using analogies. In *The IEEE International Conference on Computer Vision (ICCV)*.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv*.

Richter, M., Lins, J., Schneegans, S., and Schöner, G. (2014). A neural dynamic architecture resolves phrases about spatial relations in visual scenes. In *Artificial Neural Networks and Machine Learning – ICANN 2014*, pages 201–208.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*.

Surís, D., Epstein, D., Ji, H., Chang, S.-F., and Vondrick, C. (2019). Learning to learn words from visual scenes. *arXiv preprint arXiv:1911.11237*.

Xu, F. and Tenenbaum, J. B. (2000). Word learning as bayesian inference. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 517–522. Erlbaum.

Yu, R., Li, A., Morariu, V. I., and Davis, L. S. (2017). Visual relationship detection with internal and external linguistic knowledge distillation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1068–1076.

Zhang, J., Zhao, C., Ni, B., Xu, M., and Yang, X. (2019). Variational few-shot learning. In *The IEEE International Conference on Computer Vision (ICCV)*.