Analogy-preserving Semantic Embedding for Visual Object Categorization

Sung Ju Hwang SJHWANG@CS.UTEXAS.EDU
Kristen Grauman GRAUMAN@CS.UTEXAS.EDU
Department of Computer Science, University of Texas, Austin, TX, 78701

Fei Sha FEISHA@USC.EDU
Department of Computer Science, University of Southern California, Los Angeles, CA, 90089

Abstract

In multi-class categorization tasks, knowledge about the classes' semantic relationships can provide valuable information beyond the class labels themselves. However, existing techniques focus on preserving the semantic distances between classes (e.g., according to a given object taxonomy for visual recognition), limiting the influence to pairwise structures. We propose to model *analogies* that reflect the relationships between multiple pairs of classes simultaneously, in the form "p is to q, as r is to s". We translate semantic analogies into higher-order geometric constraints called *analogical parallelograms*, and use them in a novel convex regularizer for a discriminatively learned label embedding. Furthermore, we show how to discover analogies from attribute-based class descriptions, and how to prioritize those likely to reduce inter-class confusion. Evaluating our Analogy-preserving Semantic Embedding (ASE) on two visual recognition datasets, we demonstrate clear improvements over existing approaches, both in terms of recognition accuracy and analogy completion.

1. Introduction

Discriminative approaches to object categorization have shown much success in recent years. However, as the community shifts its focus towards fine-grained and large-scale recognition problems, the traditional view of object categories as isolated low-level visual patterns is restrictive.

Therefore, researchers have begun to explore how semantic relationships between categories might inform a purely discriminative approach. By viewing the classes as interrelated entities in some latent semantic space, the goal is not only to improve the ultimate recognition accuracy, but also to yield models that make semantically reasonable errors. Recent work makes use of semantic knowledge that is proprietary (i.e., attribute-based) or inclusive (i.e., taxonomy-based) (Zweig & Weinshall, 2007; Lampert et al., 2009; Wang & Mori, 2010; Fergus et al., 2010; Zhao et al., 2011; Hwang et al., 2011b;a). For example, one approach is to require that semantically related categories select a common set of features (Zhao et al., 2011); another is to use mid-level semantic attributes to regularize object representations (Lampert et al., 2009; Wang & Mori, 2010; Hwang et al., 2011b). However, those methods focus on each *individual* class's relationships and properties;l thus they are limited to pairwise semantic structures.

Moving beyond per-class semantic relatedness, our goal is to exploit higher-order relationships jointly involving multiple classes. Specifically, we propose to model *analogies* between classes in the form "p is to q, as r is to s" (or, in shorthand, $p : q = r : s$). An analogy encodes the relational similarity between two pairs of semantic concepts. By augmenting labeled data instances with a set of semantic analogies during training, we aim to enrich the learned representation and thereby improve generalization. Analogies can be defined with almost arbitrary abstraction, ranging from "is-a" relationships (DOG : CANINE = CAT : FELINE), to contextual dependencies (FISH : WATER = BIRD : SKY). To examine analogies most likely to benefit visual learning, we restrict our focus to *analogical pro-*

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

portions (Miclet et al., 2008)—analogies between pairs of concrete objects in the same semantic universe and with similar abstraction level.

Before sketching our approach, we first motivate why this form of analogy should offer new information to a learning algorithm. As any standardized test-taker knows, analogies are used to gauge both vocabulary skills and reasoning ability. Notably, the pairs of entities involved in an analogy need not share properties. For example, in the analogy PLANET : SUN = ELECTRON : NUCLEUS, the PLANET and ELECTRON do not have anything in common; rather, the relational similarity (*orbiter* and *center*) is what makes us recognize the two pairs as parallel in meaning (Gentner, 1983). Furthermore, the common difference exhibited by the two pairs in an analogy may encapsulate a combination of multiple properties—and that combination need not have a succinct semantic name. For example, in the analogy LEOPARD : CAT = WOLF : DOG, the common difference relating the two pairs entails multiple low-level concepts; in both, the first class *lives in the wild*, *has fangs*, and *is more aggressive*, etc. Thus, to master analogies, one must not only estimate the similarity of words, but also infer the abstract relationships implied by their pairings.

Accordingly, we expect analogies to benefit a feature learning algorithm in ways that semantic distance constraints alone cannot. Whereas existing methods inject only "vocabulary skills" by requiring that semantically related instances be close and semantically unrelated ones be far, our method will also inject "reasoning ability" by requiring that the common differences implied by analogies be reflected in the learned semantic feature space. Often, the higher-order constraints may connect quite distant sets of categories. The analogies can thus facilitate a form of transfer from class pairs that are more easily discriminated in the original feature space to analogous class pairs that are not. For example, suppose LEOPARD and CAT are often confused in the visual space because the training set consists of only close-up images, whereas DOG and WOLF are easily separable due to their distinct backgrounds. Enforcing the analogy constraint LEOPARD : CAT = WOLF : DOG could make the separation in the first pair clearer, by aligning it with the same hypothetical semantic axis of differences (*wild/fanged/aggressive*) shared by the second (more distinctive) pair.

We propose an *Analogy-preserving Semantic Embedding* (ASE), which embeds features discriminatively with analogies-based structural regularization. Given a set of analogies involving various object categories,

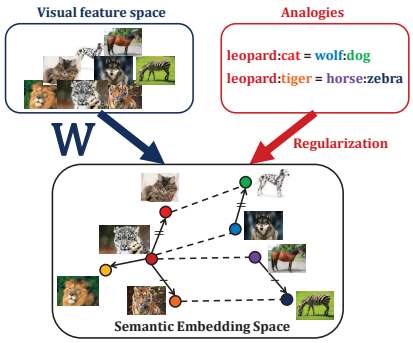

Figure 1. We introduce analogical parallelogram constraints to regularize a semantic embedding. By learning from both labeled instances and analogies, our method preserves structural similarities between category pairs.

we translate each one into a geometric constraint called an *analogical parallelogram*. This constraint states that the difference between the first pair of categories should be the same as the that between the second pair, where each category is represented by a (learned) prototype vector in some hypothetical semantic space. See Figure 1. We represent the constraints as a novel regularizer that augments a large-margin label embedding. Consequently, we obtain an embedding where examples with the same label are mutually close (and far from differently labeled points) and analogical parallelograms have nearly parallel sides.

Our learned embedding can be used for recognition, automatic analogy completion, visualization, and potentially other tasks. To use it for recognition, we project a novel image into the learned space, and predict its label based on the nearest category prototype. We further show how to automatically discover and prioritize useful analogies, which is valuable to concentrate on constraints that are influential for recognition.

Compared to traditional large-margin label embeddings (Weinberger & Chapelle, 2009; Bengio et al., 2010), our approach preserves a new form of relational similarity. While the prior methods also map to a space where semantic similarities are preserved, they risk learning spurious associations between features and labels. Our analogy-induced regularizer mitigates such adverse effects by constraining the hypothesis space with structural relations between category pairs, yielding robust models with better generalization. Even constraints not in the axes of visual properties can be helpful, as they shift the focus from brittle incidental correlations to higher-order semantic ties.

2. Related Work

Analogical logic and learning Several findings from cognitive science and AI provide background for

our approach. Gentner et al. (Gentner, 1983) study analogies in light of human cognition. They define an analogy as a relational similarity over two pairs of entities, and contrast it with the more superficial similarity defined by attributes. Based on this intuition, they suggest a conceptual structural mapping engine that enables analogical reasoning (Gentner & Markman, 1997). Recognizing that such generic analogies require high-level logical reasoning that may be problematic for an automated prediction system, Miclet et al. suggest focusing on the analogical dissimilarity between entities in the same semantic universe (Miclet et al., 2008). They exploit analogical dissimilarity to do direct logical inference when one of the entities is unknown. Our work focuses on similarly scoped analogies—the semantic universe of object categories. In contrast to their logical inference model, however, we propose geometric constraints to enforce analogical proportions in a learned embedding.

While our main idea is to use analogies in an embedding, we also show how to automatically discover categories that have analogical relationships using their attribute descriptions. In this respect, there is a connection to structural transfer learning work that discovers mappings between domains (Mihalkova et al., 2007; Wang & Yang, 2011). However, while that work aims to associate distinct source and target domains (e.g., computer viruses and human viruses), we aim to detect parallel associations within the same domain, and then use those pairings to constrain feature learning. In graphics, inferring the filter relating two input images allows the automatic creation of "image analogies" (Hertzmann et al., 2001); we deal with analogies on visual data, but our idea of using them to regularize the representation is different and original.

Semantics in recognition Recent research explores how external semantic knowledge can benefit visual recognition, e.g., (Zweig & Weinshall, 2007; Lampert et al., 2009; Wang & Mori, 2010; Fergus et al., 2010; Hwang et al., 2011b;a; Zhao et al., 2011). There, the semantics originate from taxonomies or attribute memberships, limiting what can be captured to proprietary or inclusive relations. To our knowledge, our work is the first to exploit analogical relations in learning an object recognition model, opening up the potential advantages discussed above.

Embedding and manifold learning Most existing embedding methods aim to preserve the distances between data points, either globally (Duda et al., 2001) or locally (Roweis & Saul, 2000; Weinberger & Saul, 2006). Label embeddings learned for object or document categorization also aim to preserve distances, but with further constraints to promote the discriminability of labeled classes (Weinberger & Chapelle, 2009). Recent embedding methods preserve not only the geometry of local neighborhoods, but also higher-order properties like category clusters (Shieh et al., 2011) or graph structure (Shaw & Jebara, 2009). We also aim to preserve more far-reaching structures. However, our method is distinct in that it enforces the *relative* distances between semantically related pairs of instances.

3. Approach

We assume a labeled dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{n=1}^{\mathsf{N}}$, where $\boldsymbol{x}_i \in \mathbb{R}^{\mathsf{D}}$ stands for the i-th D-dimensional feature vector and $y_i \in \mathcal{Y}$ the corresponding class label, which in our primary application of interest will correspond to an object category (PANDA, LEOPARD, etc.) We further assume that we have access to a set of analogies $\mathcal{A} = \{\alpha_1, \alpha_2, \cdots, \alpha_{\mathsf{A}}\}$. The analogies are derived either directly from human input or with an automatic discovery procedure we propose below.

The goal of our learning algorithm is to embed both the data features and the class labels in a low-dimensional space $\mathbb{R}^{\mathsf{M}}$ with $\mathsf{M} < \mathsf{D}$, while minimizing misclassification errors. In what follows, we will denote the embedding of the feature vector $\boldsymbol{x}_i$ by $\boldsymbol{z}_i$, and the embedding of the class $c \in \mathcal{Y}$ by $\boldsymbol{u}_c$. In this "semantic space", we want to ensure instances from the same class stay close to each other *and* to their class label's location $\boldsymbol{u}_c$. Moreover, and most importantly, we would like the placement of the class labels to reflect their *analogy-based relationships*.

To this end, our approach addresses two crucial challenges: i) *how can we encode an analogy between class labels via their coordinates in the learned semantic space*? and ii) *how can we automatically discover analogy relationships among a large number of categories*?

3.1. Encoding analogies

For each class $c \in \mathcal{Y}$, $\boldsymbol{u}_c \in \mathbb{R}^{\mathsf{M}}$ denotes its coordinates in the M-dimensional semantic space. Each $\boldsymbol{u}_c$ can be thought of as a prototype for the category; we will explain how the prototypes are optimized jointly with the data projection matrix $\boldsymbol{W}$ in Sec. 3.3.

An analogy involves four categories, and we represent the relationship with an ordered quadruplet $(p, q, r, s) \in \mathcal{Y} \times \mathcal{Y} \times \mathcal{Y} \times \mathcal{Y}$. As we focus on *analogical proportions* (Miclet et al., 2008), the difference between p and q is equated with the difference between r and s. Moreover, the difference between p and r also is equated with the difference between q and s.

Analogical proportions naturally induce geometric

constraints among the embeddings of the four categories in the semantic space. In particular, the geometry is characterized by a parallelogram; we will show how to exploit this structure in our learning algorithm.

Analogy parallelogram We use the vector shift $(\boldsymbol{u}_q - \boldsymbol{u}_p)$ to represent the difference between the two categories q and p in the semantic space. Note that this difference is directed, that is, $\boldsymbol{u}_q - \boldsymbol{u}_p \neq \boldsymbol{u}_p - \boldsymbol{u}_q$. The analogical proportion implied by (p, q, r, s) is thus encoded by the following pair of equalities:

$$\boldsymbol{u}_q - \boldsymbol{u}_p = \boldsymbol{u}_s - \boldsymbol{u}_r, \quad \text{and} \quad \boldsymbol{u}_r - \boldsymbol{u}_p = \boldsymbol{u}_s - \boldsymbol{u}_q. \quad (1)$$

These constraints form a parallelogram in which each vertex is a category, as illustrated in Fig. 2.

Convex regularizer There are several ways of enforcing the analogical proportion constraints in eq. (1). A natural choice is to exploit the parallel property of opposing sides. Specifically, the normalized inner products between opposing sides are the cosine of their intersection degree, which should be 1 if perfectly parallel. Concretely, for an analogy $\alpha = (p, q, r, s)$, the resulting parallelogram "score" would be defined as

$$S(\alpha) = \frac{1}{2}\left(\frac{(\boldsymbol{u}_q - \boldsymbol{u}_p)^{\mathrm{T}}(\boldsymbol{u}_r - \boldsymbol{u}_s)}{\|\boldsymbol{u}_q - \boldsymbol{u}_p\| \cdot \|\boldsymbol{u}_r - \boldsymbol{u}_s\|} + \frac{(\boldsymbol{u}_r - \boldsymbol{u}_p)^{\mathrm{T}}(\boldsymbol{u}_s - \boldsymbol{u}_q)}{\|\boldsymbol{u}_r - \boldsymbol{u}_p\| \cdot \|\boldsymbol{u}_s - \boldsymbol{u}_q\|}\right) \quad (2)$$

While intuitive, maximizing the parallelogram score (or equivalently, minimizing its negative) is computationally inconvenient, since it is not convex in the embeddings $\boldsymbol{u}$. Thus, we use a relaxed version and compare the sides only in their *lengths*. Specifically, our regularizer is defined as

$$\begin{aligned} R(\alpha) =& 1/\sigma_1 \|(\boldsymbol{u}_q - \boldsymbol{u}_p) - (\boldsymbol{u}_r - \boldsymbol{u}_s)\|_2^2 \\ &+ 1/\sigma_2 \|(\boldsymbol{u}_r - \boldsymbol{u}_p) - (\boldsymbol{u}_s - \boldsymbol{u}_q)\|_2^2, \end{aligned} \quad (3)$$

where σ_1 and σ_2 are two scaling constants used to prevent either pair of sides from dominating the other. We simply estimate them as the mean distances between data instances from different classes.

$R(\alpha)$ is convex in the embedding coordinates. Moreover, it is straightforward to kernelize as it depends only on the distances (and thus inner products).

3.2. Automatic discovery of analogies

Human knowledge is a natural source for harvesting analogy relationships among categories. However, it is likely expensive to completely rely on human assessment to acquire a sufficient number of analogies for training. To address this issue, we use *auxiliary* semantic knowledge to identify candidate analogies.

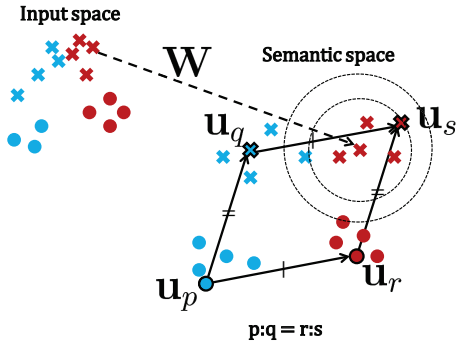

Figure 2. Geometry of ASE. **Analogy constraints for the semantic category embedding:** The analogy quadruplet (p, q, r, s) forms a parallelogram in the semantic embedding space, cf. eq. (1). **Data embedding $\boldsymbol{W}$:** At the same time, when projected onto the semantic space by $\boldsymbol{W}$, the data point $\boldsymbol{x}_i$ from class q should be closer to its semantic category embedding $\boldsymbol{u}_q$, compared to any other category embedding, by a large margin (see dotted circles).

In the context of visual object recognition, visual *attributes* are an appealing form of auxiliary semantic knowledge (Lampert et al., 2009). Attributes are binary predicates shared among certain visual categories—for example, the category PANDA has the "true" value for the SPOTTED attribute and the "false" value for the ORANGE attribute. Supposing we have access to attribute descriptions stating the typical attribute values for each category, we can automatically discover plausible analogies.

We next define two strategies to do so. The first is independent of the data instances, while the second exploits the instances to emphasize analogies more likely to lend discriminative information.

Attribute-based analogy discovery Our first strategy is to view attributes as a proxy to the embedding coordinates of the visual categories in the semantic space we are trying to learn. In the attribute space, each category is encoded with a binary vector, with bits set to one for attributes the class does possess, and bits set to zero for attributes the class does not possess. Note that this is a class-level description—we have one binary vector per object class.

Imagine that we enumerate all quadruplets of visual categories. For each quadruplet α, we compute its parallelogram score according to eq. (2), using the categories' attribute vectors as coordinates. We then select top-scoring quadruplets as our candidate analogies.

Pragmatically, we can only score a subset of all possible analogies for a large number of visual categories. Thus, to ensure good coverage, for each randomly selected pivot category p, we select at most K triplets of other categories, where K is far fewer than the to-

tal number of possible ones. We also remove equivalent analogies. For example, (p, q, r, s) is equivalent to (p, r, q, s) or other shift-invariant forms.

We will use the highest-scoring analogies to augment the class-labeled data when learning the embedding. We stress that while we discover analogies based on parallelogram scores computed in the space of attribute descriptions, we regularize the learned embedding according to parallelogram scores computed in the learned embedding coordinates (cf. Sec. 3.3). Thus, external semantics drive the "training" analogies, which in turn mold our learned semantic space.

Discriminative analogy discovery The process described thus far has two possible issues. First, it does not take the data instances into consideration. While our goal is to find a *joint* embedding space for both data instances and category labels, analogies inferred purely from attributes do not necessarily align the data and mid-level representations—they might even lead to conflicting embedding preferences! Secondly, being fully unsupervised, this procedure need not discover analogies directly useful to our classification task. In particular, the extracted candidate analogies are not indicative of whether two categories are easily distinguishable or confused.

We address both issues with an intuitive and empirically very effective heuristic. Mindful of our goal (described in the introduction) of improving discrimination for *confusable* categories by leveraging analogy relationships connecting those confusing categories to *easily distinguishable* categories, we first use baseline classifiers to estimate the pairwise confusability between categories. This step can be achieved easily with any off-the-shelf multi-way classifier and visual features computed from the training instances. The confusability between two categories p and q is defined in terms of the resulting misclassification error:

$$C_{pq} = 0.5\left[\epsilon_{p\to q} + \epsilon_{q\to p}\right],$$

where $\epsilon_{p\to q}$ is the rate of misclassifying instances from the category p as the category q, and likewise for $\epsilon_{q\to p}$.

Our next step is to refine the candidate analogies generated above by finding those with *unbalanced confusability*. Specifically, for each analogy $\alpha = (p, q, r, s)$, we compute its discrimination potential:

$$P(\alpha) = \left|\log(1 + C_{pq}) - \log(1 + C_{rs})\right|. \quad (4)$$

This score attains its maximum when C_{pq} and C_{rs} are drastically different—that is, if one is 0 and the other is 1. We use this score to re-rank the K candidate analogies generated for each category p. Intuitively, we seek the quadruplet where one pair of categories is easily distinguishable (based on the image data) while the other pair is difficult to differentiate. Precisely by enforcing their analogy relationship, we expect the easy pair to assist discrimination for the difficult one.

To summarize, our automatic discovery of analogies is a two-phase strategy. We first use an auxiliary semantic space to identify a set of candidate analogies where the four categories are highly likely to form a parallelogram. Then, we analyze misclassification error patterns of these categories and use the scoring function in eq. (4) to determine the potential of each analogy in improving classification performance. We describe next how to use the highest-scoring analogies to learn the joint embedding of both features and categories.

3.3. Discriminative learning of the ASE

Next we explain how we regularize a discriminative embedding to account for the analogies.

Large margin-based discrimination We aim to learn a projection matrix $\boldsymbol{W} \in \mathbb{R}^{\mathsf{M}\times\mathsf{D}}$ to map each data instance (image example) $\boldsymbol{x}_i$ into the semantic space, giving its M-dimensional coordinates $\boldsymbol{z}_i = \boldsymbol{W}\boldsymbol{x}_i$.[1] The ideal projection matrix $\boldsymbol{W}$ should make $\boldsymbol{z}_i$ close to its corresponding label's embedding $\boldsymbol{u}_{y_i}$ and distant to all other labels' embeddings (Weinberger & Chapelle, 2009). Specifically, we enforce the large margin constraint for every training instance,

$$\|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{u}_{y_i}\|_2^2 + 1 \le \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{u}_c\|_2^2 + \xi_{ic}, \ \ \forall\ c \neq y_i \quad (5)$$

where $\xi_{ic} \ge 0$ is a slack variable for satisfying the separation by the margin of 1.

Regularization To jointly embed both features and class labels, we regularize so that the class labels in the analogy set $\mathcal{A}$ form parallelograms as much as possible. The regularizer is given by

$$R_{total}(\mathcal{A}) = \sum_a \omega_a R(\alpha_a), \quad (6)$$

which is the weighted sum of the regularization defined in eq. (3) for each analogy α_a. If using the "raw" attribute-based analogies, the weight $\omega_a = S(\alpha_a)$, thus enforcing stricter regularization for category quadruplets whose structure is closer to a "perfect" analogy. If using discriminatively discovered analogies, the weight is instead $\omega_a = P(\alpha_a)$, thus prioritizing those that are more discriminative.

Additionally, we also constrain the parameters $\boldsymbol{W}$ and all $\boldsymbol{u}_c$ with their Frobenius norms: $\|\boldsymbol{W}\|_F^2$ and

[1] Nonlinear embeddings are possible via kernelization.

$R(\boldsymbol{u}) = \sum_c \|\boldsymbol{u}_c - \boldsymbol{u}_c^{\text{PRIOR}}\|_2^2$. In particular, for the class label embeddings, we constrain them to be close to our prior knowledge on their locations $\boldsymbol{u}_c^{\text{PRIOR}}$. The prior knowledge could be null such that we set $\boldsymbol{u}_c^{\text{PRIOR}}$ to zeroes. Or, the class label embeddings could be computed from auxiliary information, for example, the multi-dimensional embedding of class labels where the dissimilarities between labels are measured with tree distances from a taxonomy (Weinberger & Chapelle, 2009) or attributes. We consider both in the results.

3.4. Numerical optimization

Our learning problem is thus cast as the following optimization problem:

$$\min_{\boldsymbol{W},\{\boldsymbol{u}_c\}} \quad \sum_{ic} \xi_{ic} + \lambda R_{total}(\mathcal{A}) + \mu \|\boldsymbol{W}\|_F + \tau R(\boldsymbol{u}) \quad (7)$$

subject to both the large margin constraints in eq. (5) and non-negativity constraints on the slack variables ξ_{ic}. The regularization coefficients λ, μ, and τ are determined via cross-validation.

The optimization is nonconvex due to the quadratically-formed large margin constraints. We have developed two methods for solving it. Our first method uses stochastic (sub)gradient descent, where we update $\boldsymbol{W}$ and $\boldsymbol{u}_c$ according to their subgradients computed on a subset of instances. Despite its simplicity, this method works well in practice and scales better to problems with many categories.

We also consider a convex relaxation analogous to the procedure in (Weinberger & Chapelle, 2009). Briefly, in eq. (7), we hold $\{\boldsymbol{u}_c\}$ fixed first and solve $\boldsymbol{W}$ in closed-form, $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{Q}$ where the matrix $\boldsymbol{U}$ is composed of $\{\boldsymbol{u}_c\}$ as column vectors. The matrix $\boldsymbol{Q}$ depends only on $\boldsymbol{x}_i$ and is constant with respect to $\boldsymbol{U}$ or $\boldsymbol{W}$. Substituting the solution of $\boldsymbol{W}$ into both the objective function eq. (7) and the large margin constraints eq. (5), we can reformulate the optimization in terms of $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$. In particular, the original non-convex large margin constraints in $\boldsymbol{U}$ can be relaxed into convex if we reparameterize $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}$ as a positive semidefinite matrix $\boldsymbol{V}$. We then solve $\boldsymbol{V}$ and recover the solutions $\boldsymbol{U}$ and $\boldsymbol{W}$, respectively. For cases where D is much larger than the number of categories, we expect this variant to optimize faster.

4. Experimental Results

We validate three aspects: i) the effectiveness of our analogy discovery approach; ii) recognition accuracy when incorporating discovered analogies in learning embeddings; and iii) "fill in the blank"—a Graduate Record Examination (GRE)-style prediction task of filling in the category that would form a valid analogy.

Datasets and implementation details We use three datasets created from two public image datasets: Animals with Attributes (AWA), which contains 50 animal classes (Lampert et al., 2009) and ImageNet, which contains general object categories (Deng et al., 2009). They were chosen due to their available attribute descriptions and their challenging diverse content. From AWA, we create two datasets: **AWA-10** of 6,180 images from 10 classes (Lampert et al., 2009), and the complete 50-class **AWA-50** of 30,475 images. From ImageNet, we use the 50-class **ImageNet-50** with annotated attributes (Russakovsky & Fei-Fei, 2010), totaling 70,380 images.

We use the features provided by the authors, which consist of SIFT and other texture and color descriptors. We use PCA to reduce the feature dimensionality to $\mathsf{D} = 150$ for efficient computation. Additionally, we augment ImageNet-50 with attribute labels for colors, material, habitat, and behaviors (e.g., *big*, *round*, *feline*), yielding 39 and 85 binary attributes for ImageNet and AWA, respectively. We fix $\mathsf{K} = 10,000$. We use the convex relaxation, since the dimensionality is much greater than the number of classes; accordingly, the semantic space dimensionality M equals the number of categories (10 or 50).

4.1. Automatic discovery of analogies

In real-world settings, acquiring all analogies from manual input may be costly and impractical. Thus, we first examine the analogies discovered by our method (Sec. 3.2), which assumes only that attribute-labeled object classes are available.

Figure 3 displays several examples for AWA-50 and ImageNet-50. Most analogies are intuitive to understand. For example, in the second row of COLLIE:DALMATIAN = LION:LEOPARD, the categories COLLIE and LION are both furry and brown, while the categories DALMATIAN and LEOPARD are both spotted and lean. We also see that the analogies can be largely visual (e.g., the third row), an upshot of the many visually relevant attributes offered with the datasets.

4.2. Visual recognition with ASE

We compare the classification performance of our Analogy-preserving Semantic Embedding (ASE) to the following baselines, all of which lack analogies:

(1) **SVM-RBF**: Multiclass SVM with RBF kernel.

(2) **Large margin embedding (LME)**: The existing

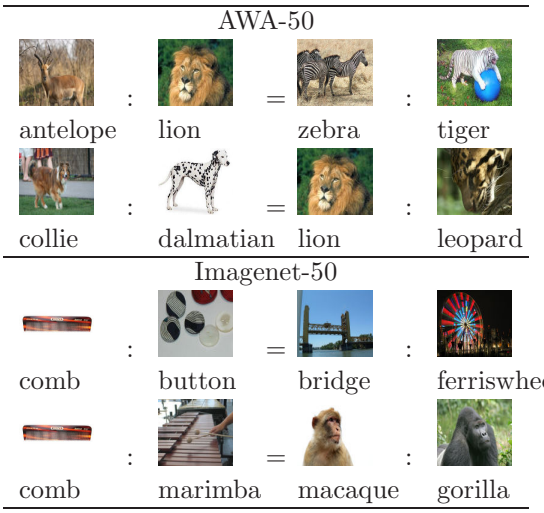

Figure 3. Example analogies discovered from attributes.

technique of (Weinberger & Chapelle, 2009), which is a special case of our approach where the effect of the analogy regularizer is disabled by setting $\tau = 0$ and $\lambda = 0$ in eq. (7). For this baseline, the embeddings of the class labels are constrained only to satisfy the large margin separation criterion of eq. (5);

(3) **Large margin embedding with attributes prior (LME$^{\text{prior}}$)**: This baseline adds the prior regularizer to LME, where we adjust τ for eq. (7) via cross-validation. In particular, we use the multi-dimensional scaling (MDS) embedding of class labels where the pairwise dissimilarity is the Euclidean distance between the *attribute* vectors of two classes. The contrast between LME and LME$^{\text{PRIOR}}$ reveals how useful attributes as auxiliary semantic information are in yielding discriminative embeddings, separating out the impact of attributes from the impact of analogies.[2]

All embedding methods classify novel images according to the nearest category $\boldsymbol{u}_c$ in the embedding space.

For our method, we include two variants, differentiated only by how the analogies are discovered, cf. Sec. 3.2. In ASE-A, the analogies are derived solely from attributes, aiming to preserve parallelograms as much as possible. In ASE-C, the analogies are derived from the discrimination-based discovery, aiming to use distinct categories to assist confusable categories. The confusability among categories is measured using the baseline LME classifier on the validation set.

In our experiments, all hyperparameters (regularization coefficients, kernel function parameters) are tuned via cross-validation. We use 30 examples per class for both training and testing, and use another 30 images as a validation set to learn the parameters. We report the average results over 5 such random splits.

[2]We also tested LME using WordNet object distances as a prior, but found it inferior to the attribute prior.

Table 1. Multiclass classification accuracy. The numbers denote mean and the standard error over 5 runs.

Dataset	AWA-10	AWA-50	ImageNet-50
#. analogies	5	50	50
SVM-RBF	43.00 ± 1.94	19.32±0.57	15.37±0.93
LME	44.40±2.83	19.65±0.90	16.52±1.10
LME$^{\text{PRIOR}}$	44.93 ± 3.57	20.12±1.03	16.59±0.39
ASE-A (ours)	45.47±3.10	20.60±0.93	17.08±0.36
ASE-C (ours)	**45.93±2.90**	**21.05±0.82**	**17.24±0.62**

How do analogies affect recognition accuracy? We first validate our method on multiclass classification. Since the analogies help preserve the intrinsic semantic structure among objects, we expect the learned space to show better generalization power, and hence improved object categorization.

Table 1 shows the results.[3] We report the optimal number of analogies selected from preliminary experiments, though the results were in general insensitive to the number of analogies. On all three datasets, we observe clear improvement using our analogy-preserving embedding variants over both LME variants.

We see that the difference in accuracy for LME and LME$^{\text{PRIOR}}$ is in general smaller than the improvement from LME to ASE. This suggests that using attribute distances *alone* as a prior to constrain embeddings (as LME$^{\text{PRIOR}}$ does) is not sufficient. In contrast, in ASE, the prior and the analogy constraints work together, leading to a noticeable improvement.

Which types of analogies should we use? We also observe that our ASE-C variant outperforms ASE-A. This coincides with our intuition that the analogies would be much more helpful for discrimination if a pair of easily confusable categories can leverage a pair of easily distinguishable categories.

Detailed analysis supports this intuition even more strongly. Figure 4 compares the amount of reduction in confusability among the 10 classes of AWA-10, from LME$^{\text{PRIOR}}$ to either ASE-A (left) or ASE-C (right). We observe that for ASE-A, the improvement is made on pairs that are not included in the analogies; in contrast, for ASE-C, the improvements are mostly made on pairs that *are* included in analogies. This noticeable correlation between the category pairs selected for analogies, and the pairs whose confusion is reduced (for ASE-C) suggests that our consideration of the pairwise confusion is indeed the reason ASE-

[3]Attribute-based categorization (Lampert et al., 2009) underperforms all baselines (AWA-10: 28.80, AWA-50: 17.80, ImageNet-50: 11.14).

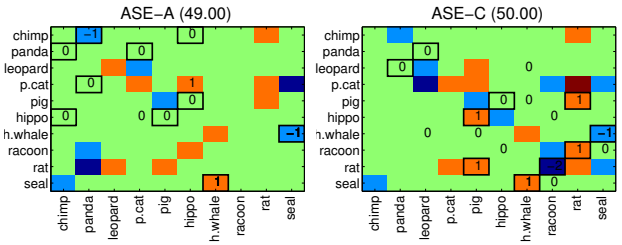

Figure 4. Confusion reduction. **Left:** $C_{\text{LME}} - C_{\text{ASE-A}}$, **Right:** $C_{\text{LME}} - C_{\text{ASE-C}}$. The numbers and colors at each entry show the reduction in confusion (red:↑, blue:↓). Outlined entries are pairs that appear in the training analogies. Positive off-diagonal entries indicate reduced confusion. ASE-C focuses on initially confused classes.

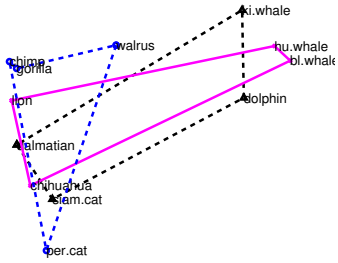

Figure 5. AWA-50 categories projected to 2D using ASE-C. We show only three analogies for ease of viewing: 1) DALMATIAN:SIAMESE CAT=KILLER WHALE:DOLPHIN, 2) LION:CHIHUAHUA = HUMPBACKWHALE:BLUEWHALE, and 3) CHIMP:GORILLA = PERSIAN CAT:WALRUS.

C outperforms ASE-A, whose analogies do not care about the data distribution.

Figure 5 shows projections of AWA-50 categories to a $2D$ space using ASE-C. We see that the quadrilaterals formed by the four categories involved in each analogy do indeed show distinct parallelogram shape.

4.3. Completing a visual analogy

Finally, we subject our method to a GRE test. Given $p : q = r :?$, how well can our method fill in the blank, based on its representation of the three other classes? In this analogical reasoning task, which is performed by virtually every graduate school applicant, the learning algorithm is given a set of complete analogies $\mathcal{A}^{\text{TRAIN}}$. Then it is given a disjoint test set of analogies $\mathcal{A}^{\text{TEST}}$, each of which has its fourth category missing. No analogies overlap in (p, q, r) between the two sets. To fill in the blank with ASE or LME, we simply rank each category according to its parallelogram score when its $\boldsymbol{u}_c$ is used as the fourth category. The more parallelogram-like, the more it appears to be the right answer. The ground truth answer is the class maximizing the parallelogram score according to the auxiliary attribute ground truth.

Our hypothesis is that by learning to discriminate categories *in conjunction with* preserving the analogy constraints in $\mathcal{A}^{\text{TRAIN}}$, the learned semantic embedding will generalize well to complete the novel analogies, without resorting to auxiliary information.

Dataset	AWA-10		AWA-50		Imagenet-50	
k	1	3	1	3	1	3
Chance	14.29	42.86	2.13	6.38	2.13	6.38
LME	36.00	52.00	4.80	12.40	1.60	7.20
$\text{LME}^{\text{PRIOR}}$	52.00	68.00	5.60	14.40	0.80	6.80
ASE-A	**64.00**	**88.00**	**8.40**	**20.80**	2.80	6.40
ASE-C	60.00	80.00	5.20	15.60	**3.20**	**8.80**

Table 2. Top-k class prediction accuracy, given an analogy with an unknown class in the form *p:q=r:?*

Analogy question	LME	$\text{LME}^{\text{PRIOR}}$	ASE-A
AWA-50			
leopard:lion = dalmatian:?	bobcat	s. monkey	fox
horse:g.shepherd = sheep:?	weasel	antelope	collie
skunk:mouse = killerwhale:?	fox	bluewhale	dolphine
Imagenet-50			
badger:skunk = button:?	g.spider	bathtub	buckle
marimba:rule = baboon:?	kitfox	orangutan	patas
b.ball:bathtub = r.coaster:?	jaguar	pooltable	bridge

Table 3. Sample analogy completion results

Table 2 strongly supports our hypothesis. We report the prediction accuracy averaged over 5 random trials, where we take the classes with the top k parallelogram scores as guesses. ASE-A achieves the best accuracy, followed by ASE-C. They both outperform the LME methods, which lack analogical constraints. On AWA-10, we predict the right completion in the first guess ($k = 1$) 64% of the time. There is clearly room for improvement, though, as accuracy decreases substantially for all methods on the larger 50-class datasets. Table 3 shows example completed analogies for AWA-50. Compared to LME, ASE selects more intuitive classes to fill in the missing values.

5. Conclusion

Our work introduces a semantic embedding for visual data that preserves structural similarities in the form of analogies. In addition to formulating a novel regularizer suitable for our goal, we also explore ways to systematically discover plausible analogies from auxiliary attribute information. Our method improves recognition accuracy over an existing "distance-only" embedding approach, thanks to its ability to preserve higher-order structures and facilitate transfer between easier and harder pairs of objects. Beyond benefitting recognition, we show it also allows analogy completion—a high-level reasoning task. We next plan to explore more general forms of analogies, such as pairs of subgraphs containing multiple categories.

Acknowledgements Research is supported in part by NSF IIS-1065390 (KG) and NSF IIS-1065243 (FS).

References

Bengio, S, Weston, J, and Grangier, D. Label Embedding Trees for Large Multi-Class Task. In *NIPS*, 2010.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

Duda, R., Hart, P., and Stork, D. *Pattern Classification*, chapter 10. John Wiley and Sons, Inc., New York, 2 edition, 2001.

Fergus, R., Bernal, H., Weiss, Y., and Torralba, A. Semantic label sharing for learning with many categories. In *ECCV*, 2010.

Gentner, D. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155 – 170, 1983.

Gentner, D. and Markman, A. B. Structure mapping in analogy and similarity. *American Psychologist*, 52:45–56, 1997.

Hertzmann, A., Jacobs, C., Oliver, N., Curless, B., and Salesin, D. Image analogies. In *SIGGRAPH*, 2001.

Hwang, S. J., Grauman, K., and Sha, F. Learning a tree of metrics with disjoint visual features. In *NIPS*, 2011a.

Hwang, S. J., Sha, F., and Grauman, K. Sharing features between objects and their attributes. In *CVPR*, 2011b.

Lampert, C., Nickisch, H., and Harmeling, S. Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009.

Miclet, L., Bayoudh, S., and Delhay, A. Analogical dissimilarity. *JAIR*, 32(1):793–824, 2008. ISSN 1076-9757.

Mihalkova, L., Huynh, T., and Mooney, R. Mapping and revising markov logic networks for transfer learning. In *AAAI*, 2007.

Roweis, S. and Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Russakovsky, O. and Fei-Fei, L. Attribute learning in large-scale datasets. In *ECCV*, 2010.

Shaw, B. and Jebara, T. Structure preserving embedding. In *ICML*, 2009.

Shieh, Albert, Hashimoto, Tatsunori, and Airoldi, Edo. Tree preserving embedding. In *ICML*, 2011.

Wang, H. and Yang, Q. Transfer learning by structural analogy. In *AAAI*, 2011.

Wang, Y. and Mori, G. A discriminative latent model of object classes and attributes. In *ECCV*, 2010.

Weinberger, K. Q. and Chapelle, O. Large margin taxonomy embedding for document categorization. In *NIPS*, 2009.

Weinberger, K. Q. and Saul, L. K. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*, 2006.

Zhao, B., Fei, L. Fei, and Xing, E. P. Large-scale category structure aware image categorization. In *NIPS*, 2011.

Zweig, A. and Weinshall, D. Exploiting Object Hierarchy: Combining Models from Different Category Levels. In *ICCV*, 2007.