

A Unified Semantic Embedding: Relating Taxonomies and Attributes

Sung Ju Hwang

Ulsan National Institute of Science and Technology
Ulsan, Korea

Leonid Sigal

Disney Research
Pittsburgh, PA

Abstract

We propose a method that learns a discriminative yet semantic space for object categorization, where we also embed auxiliary semantic entities such as supercategories and attributes. Contrary to prior work which only utilized them as side information, we explicitly embed the semantic entities into the same space where we embed categories, which enables us to represent a category as their linear combination. By exploiting such a unified model for semantics, we enforce each category to be represented by a supercategory + sparse combination of attributes, with an additional exclusive regularization to learn discriminative composition.

Semantic approaches have gained a lot of attention recently for object categorization, as object categorization problems became more focused on large-scale and fine-grained recognition tasks and datasets. Attributes (Lampert, Nickisch, and Harmeling 2009; Farhadi et al. 2009; Hwang, Sha, and Grauman 2011; Akata et al. 2013) and semantic taxonomies (Marszalek and Schmid 2008; Griffin and Perona 2008; Weinberger and Chapelle 2009; Gao and Koller 2011) are two of the popular semantic sources which impose certain relations between the category models. While many techniques have been introduced to utilize each of the individual semantic sources for object categorization, no unified model has been proposed to relate them.

We propose a unified semantic model where we can learn to place categories, super categories, and attributes as points (or vectors) in a hypothetical common semantic space. Further, we propose a discriminative learning framework based on dictionary learning and large margin embedding, to learn each of these semantic entities to be well separated and pseudo-orthogonal, such that we can use them to improve visual recognition tasks, e.g., category/attribute recognition.

However, having semantic entities embedded into a common space is not enough to utilize the vast number of relations that exist among them. Thus, we impose a graph-based regularization between the semantic embeddings, such that each semantic embedding is regularized by sparse combination of auxiliary semantic embeddings.

The observation we make to draw the relation between the categories and attributes, is that a category can be represented as the sum of its super category + the category-specific modifier, which in many cases can be represented by a combination of attributes. Further, we want the representation to be compact. Instead of describing a dalmatian as a domestic animal with a lean body, four legs, a long tail, and spots, it is more efficient to say it is a spotted dog (Figure 1). It is also more exact since the higher-level category dog contains all general properties of different dog breeds, including indescribable dog-specific properties, such as the shape of the head, and its posture. This exemplifies how a human would describe an object, to efficiently communicate and understand the concept.

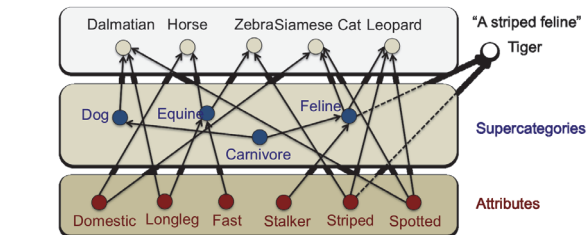


Figure 1: **Concept:** We regularize each category to be represented by its supercategory + a sparse combination of attributes, where the combinations are learned. The resulting embedding model improves the generalization, and is also able to compactly represent a novel category. For example, our model can describe a *tiger* as a *striped feline*. Such decomposition can hold for categories at any level. For example, supercategory *feline* can be described as a *stalking carnivore*.

resented as the sum of its super category + the category-specific modifier, which in many cases can be represented by a combination of attributes. Further, we want the representation to be compact. Instead of describing a dalmatian as a domestic animal with a lean body, four legs, a long tail, and spots, it is more efficient to say it is a spotted dog (Figure 1). It is also more exact since the higher-level category dog contains all general properties of different dog breeds, including indescribable dog-specific properties, such as the shape of the head, and its posture. This exemplifies how a human would describe an object, to efficiently communicate and understand the concept.

This additional requirement imposed on the discriminative learning model would guide the learning such that we obtain not just the optimal model for class discrimination, but to learn a semantically plausible model which has a potential to be more robust and human-interpretable; we call this model Unified Semantic Embedding (USE).

Learning a unified semantic embedding space

Suppose we have d -dimensional image descriptors and m -dimensional label vectors, including category labels, at different semantic granularities, and attributes. Our goal is to embed both images and labels into a single unified semantic space. To formally state the problem, given a training set \mathcal{D}

that has N labeled examples, i.e. $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes image descriptors and $y_i \in \{1, \dots, m\}$ are their labels associated with m unique concepts, we want to embed each \mathbf{x}_i as \mathbf{z}_i , and each label y_i as \mathbf{u}_{y_i} in the d_e -dimensional space, such that the similarity between \mathbf{z}_i and \mathbf{u}_{y_i} , $S(\mathbf{z}_i, \mathbf{u}_{y_i})$, is maximized. Assuming linear embedding with matrix \mathbf{W} , $\mathbf{z}_i = \mathbf{W}\mathbf{x}_i$.

To ensure that the projected instances have higher similarity to its own category embedding than to others, we add discriminate constraints, which are large-margin constraints on distance: $\|\mathbf{W}\mathbf{x}_i - \mathbf{u}_{y_i}\|_2 + 1 \leq \|\mathbf{W}\mathbf{x}_i - \mathbf{u}_c\|_2 + \xi_{ic}$, $y_i \neq c$. This translates to the following discriminative loss:

$$\mathcal{L}_C(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i) = \sum_c [1 + \|\mathbf{W}\mathbf{x}_i - \mathbf{u}_{y_i}\|_2^2 - \|\mathbf{W}\mathbf{x}_i - \mathbf{u}_c\|_2^2]_+$$

where \mathbf{U} is the columnwise concatenation label embedding vectors, such that \mathbf{u}_j denotes j_{th} column of \mathbf{U} . After replacing the generative loss in the ridge regression formula with the discriminative loss, we get the following discriminative learning problem:

$$\min_{\mathbf{W}, \mathbf{U}} \sum_i^N \mathcal{L}_C(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i) + \lambda \|\mathbf{W}\|_F^2 + \lambda \|\mathbf{U}\|_F^2,$$

where λ regularizes \mathbf{W} and \mathbf{U} from going to infinity. This is one of the most common objectives used for learning discriminative category embeddings for multi-class classification (Bengio, Weston, and Grangier 2010; Weinberger and Chapelle 2009).

Supercategories. While our objective is to better categorize *entry* level categories, categories in general can appear in different semantic granularities. For example, a *zebra* could be both an *equus*, and an *odd-toed ungulate*. To learn the embeddings for the supercategories, we map each data instance to be closer to its correct supercategory embedding than to its siblings: $\|\mathbf{W}\mathbf{x}_i - \mathbf{u}_s\|_2 + 1 \leq \|\mathbf{W}\mathbf{x}_i - \mathbf{u}_c\|_2 + \xi_{sc}$, $\forall s \in \mathcal{P}_{y_i}$ and $c \in \mathcal{S}_s$ where \mathcal{P}_{y_i} denotes the set of superclasses at all levels for class s , and \mathcal{S}_s is the set of its siblings. The constraints can be translated into the following loss:

$$\mathcal{L}_S(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i) = \sum_{s \in \mathcal{P}_{y_i}} \sum_{c \in \mathcal{S}_s} [1 + \|\mathbf{W}\mathbf{x}_i - \mathbf{u}_s\|_2^2 - \|\mathbf{W}\mathbf{x}_i - \mathbf{u}_c\|_2^2]_+.$$

Attributes. Attributes can be considered as a normalized basis vectors for the semantic space, whose combination represents a category. Basically, we want to maximize the correlation between the projected instance that possess the attribute, and its correct attribute embedding, as follows:

$$\mathcal{L}_A(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i) = \sum_{a \in \mathcal{A}_{y_i}} [\sigma - (\mathbf{W}\mathbf{x}_i)^\top y_i^a \mathbf{u}_a]_+,$$

$$\|\mathbf{u}_a\|_2^2 \leq 1, y_i^a \in \{0, 1\},$$

where \mathcal{A}_c is the set of all attributes for class c , σ is the margin (we simply use a fixed value of $\sigma = 1$), y_i^a is the label indicating presence/absence of each attribute a for the i_{th} training instance, and \mathbf{u}_a is the embedding vector for attribute a .

Semantic regularization. The previous multi-task formulation enables to implicitly associate the semantic entities, with the shared data embedding \mathbf{W} . However, we want to further explicitly impose structural regularization on the semantic embeddings \mathbf{U} , based on the intuition that an object class can be represented as its parent level class + a sparse combination of attribute as follows:

$$\mathcal{R}(\mathbf{U}, \mathbf{B}) = \sum_c^C \|\mathbf{u}_c - \mathbf{u}_p - \mathbf{U}^A \beta_c\|_2^2 + \gamma_2 \|\beta_c + \beta_o\|_2^2,$$

$$c \in \mathcal{C}_p, o \in \mathcal{P}_c \cup \mathcal{S}_c, 0 \preceq \beta_c \preceq \gamma_1, \forall c, p \in \{1, \dots, C + S\},$$

where \mathbf{U}^A is the aggregation of all attribute embeddings $\{\mathbf{u}_a\}$, \mathcal{C}_p is the set of children classes for class p , γ_1 is the sparsity parameter, and C is the number of categories. \mathbf{B} is the matrix whose j_{th} column vector β_j is the reconstruction weight for class j , \mathcal{S}_c is the set of all sibling classes for class c , and γ_2 is the parameters to enforce exclusivity. We require β to be non-negative, since it makes more sense to describe an object with attributes that it *has*, rather than attributes it does *not* have.

The exclusive regularization term is used to prevent the semantic reconstruction β_c for class c from fitting to the same attributes fitted by its parents and siblings. Such regularization will enforce the categories to be ‘semantically’ discriminated as well. With the sparsity regularization enforced by γ_1 , the simple sum of the two weights will prevent the two (super)categories from having high weight for a single attribute, which will let each category embedding to fit to exclusive attributes.

Unified semantic embeddings with semantic regularization. After augmenting the categorization objective in Eq. with the superclass and attributes loss and the sparse-coding based regularization presented in the previous paragraph, we obtain the following multitask learning formulation:

$$\min_{\mathbf{W}, \mathbf{U}, \mathbf{B}} \sum_{i=1}^N \mathcal{L}_C(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i) +$$

$$\mu_1 (\mathcal{L}_S(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i) + \mathcal{L}_A(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i)) + \mu_2 \mathcal{R}(\mathbf{U}, \mathbf{B})$$

$$\|\mathbf{w}_j\|_2^2 \leq \lambda, \|\mathbf{u}_k\|_2^2 \leq \lambda, 0 \preceq \beta_c \preceq \gamma_1,$$

$$\forall j \in \{1, \dots, d\}, \forall k \in \{1, \dots, m\}, \forall c, p \in \{1, \dots, C + S\},$$

where S is the number of supercategories, \mathbf{w}_j is \mathbf{W} ’s j_{th} column, and μ_1 and μ_2 are parameters to balance between the main and auxiliary tasks, and discriminative and generative objective.

The above equation can also be used for knowledge transfer when learning a model for a novel set of categories, by replacing \mathbf{U}^A in $\mathcal{R}(\mathbf{U}, \mathbf{B})$ with \mathbf{U}^S , learned on class set \mathcal{S} to transfer the knowledge from.

Numerical optimization. Eq. is not jointly convex, and has both discriminative and generative terms. The problem is similar to the problem in (Mairal et al. 2008), and can be optimized using a similar alternating optimization, while alternating between the following two convex sub-problems: 1) Optimization of the data embedding \mathbf{W} and parameters \mathbf{B} , and 2) Optimization of the category embedding \mathbf{U} .

	Method	Flat hit @ k (%)			Hierarchical precision @ k (%)	
		1	2	5	2	5
No semantics	Ridge Regression	38.39 ± 1.48	48.61 ± 1.29	62.12 ± 1.20	38.51 ± 0.61	41.73 ± 0.54
	LME	44.76 ± 1.77	58.08 ± 2.05	75.11 ± 1.48	44.84 ± 0.98	49.87 ± 0.39
Implicit semantics	ALE (Akata et al. 2013)	36.40 ± 1.03	50.43 ± 1.92	70.25 ± 1.97	42.52 ± 1.17	52.46 ± 0.37
	HLE (Akata et al. 2013)	33.56 ± 1.64	45.93 ± 2.56	64.66 ± 1.77	46.11 ± 2.65	56.79 ± 2.05
	AHLE (Akata et al. 2013)	38.01 ± 1.69	52.07 ± 1.19	71.53 ± 1.41	44.43 ± 0.66	54.39 ± 0.55
Explicit semantics	LME-MTL-S	45.03 ± 1.32	57.73 ± 1.75	74.43 ± 1.26	46.05 ± 0.89	51.08 ± 0.36
	LME-MTL-A	45.55 ± 1.71	58.60 ± 1.76	74.67 ± 0.93	44.23 ± 0.95	48.52 ± 0.29
USE	USE-No Reg.	45.93 ± 1.76	59.37 ± 1.32	74.97 ± 1.15	47.13 ± 0.62	51.04 ± 0.46
	USE-Reg.	46.42 ± 1.33	59.54 ± 0.73	76.62 ± 1.45	47.39 ± 0.82	53.35 ± 0.30

Table 1: Multiclass classification performance on the AWA dataset (4096-D DeCAF features).

Results

We validate our method for multiclass categorization performance and knowledge transfer on the Animals with Attributes dataset (Lampert, Nickisch, and Harmeling 2009), which consists of 30, 475 images on 50 animal classes, with 85 class-level attributes¹. We use the Wordnet hierarchy to generate supercategories. Since there is no fixed training/test split, we use {30,30,30} random split for training/validation/test. For the features, we use the provided 4096-D DeCAF features obtained from a deep convolutional neural network.

We compare USE against multiple existing embedding-based categorization approaches, that either do not use any semantic information, or use semantic information but do not explicitly embed semantic entities. For non-semantic baselines, we use **Ridge Regression**, a linear regression with ℓ_2 norm, and **LME**, a base large-margin embedding (Eq.) solved using alternating optimization. For implicit semantic baselines, we consider **ALE**, **HLE**, and **AHLE**, which are our implementation of Akata et al. (Akata et al. 2013). The method inputs the semantic information by representing each class with structured labels that indicate the class’ association with superclasses and attributes. We implement variants that use attributes (ALE), leaf level + superclass labels (HLE), and both (AHLE) labels.

We implement multiple variants of our model to analyze the impact of each semantic entity and the proposed regularization. **1) LME-MTL-S:** The multitask semantic embedding model learned with supercategories. **2) LME-MTL-A:** The multitask embedding model learned with attributes. **3) USE-No Reg.:** The unified semantic embedding model learned using both attributes and supercategories, without semantic regularization. **4) USE-Reg:** USE with the sparse coding regularization. We find the optimal parameters for the USE model by cross-validation on the validation set.

Multiclass categorization. We first evaluate the USE framework for categorization performance. We report the average classification performance and standard error over 5 random training/test splits in Table 1, using both flat hit@k,

¹Attributes are defined on color (black, orange), texture (stripes, spots), parts (longneck, hooves), and other high-level behavioral properties (slow, hibernate, domestic) of the animals.

which is the accuracy at the top-k prediction made, and hierarchical precision@k from (Frome et al. 2013), which is a precision the given label is correct at k , at all levels.

The implicit semantic baselines, ALE-variants, underperformed even the ridge regression baseline with regard to the top-1 classification accuracy², while they improve upon the top-2 and hierarchical precision. This shows that hard-encoding structures in the label space do not necessarily improve the discrimination performance, while it helps to learn a more semantic space.

Explicit embedding of semantic entities using our method improved both the top-1 accuracy and the hierarchical precision, with USE variants achieving the best performance in both. USE-Reg. made substantial improvements on flat hit and hierarchical precision @ 5, which shows the proposed regularization’s effectiveness in learning a semantic space that also discriminates well.

Qualitative analysis. Besides learning a space that is both discriminative and generalizes well, our method’s main advantage is its ability to generate compact, semantic description of each category it has learned. This is a great caveat, since in most models, including the state-of-the-art deep convolutional networks, humans cannot understand what has been learned; by generating human-understandable explanation, our model can *communicate* with the human, allowing understanding of rationale behind the categorization decision, and to possibly provide feedback for correction.

To show the effectiveness of using supercategory+attributes in the description, we report the learned reconstruction for our model, compared against the description generated by ground-truth attributes in Table 2. The results show that our method generates compact description of each category, focusing on its *discriminative* attributes. For example, our method selects *flippers* for otter, and *stripes* for skunk, instead of common nondiscriminative attributes such as *tail*. Further, our method selects attributes for each supercategory, while there is no provided attribute label for supercategories.

One-shot/Few-shot learning. Our method is expected to be especially useful for few-shot learning, by generating a

²We did extensive parameter search for the ALE variants.





Category	Ground-truth attributes	Supercategory + learned attributes
 Otter	An animal that swims, fish, water, new world, small, flippers, furry, black, brown, tail, ...	A musteline mammal that is quadrapedal, flippers, furry, ocean
 Skunk	An animal that is smelly, black, stripes, white, tail, furry, ground, quadrapedal, new world, walks, ...	A musteline mammal that has stripes
 Deer	An animal that is brown, fast, horns, grazer, forest, quadrapedal, vegetation, timid, hooves, walks, ...	A deer that has spots, nestspot, longneck, yellow, hooves
 Moose	An animal that has horns, brown, big, quadrapedal, new world, vegetation, grazer, hooves, strong, ground, ...	A deer that is arctic, stripes , black
Equine	N/A	An odd-toed ungulate, that is lean and active
Primate	N/A	An animal, that has hands and bipedal

Table 2: Semantic description generated using ground truth attributes labels and learned semantic decomposition of each categories. For ground truth labels, we show top-10 ranked by their human-ranked relevance. For our method, we rank the attributes by their learned weights. Incorrect attributes are colored in red.

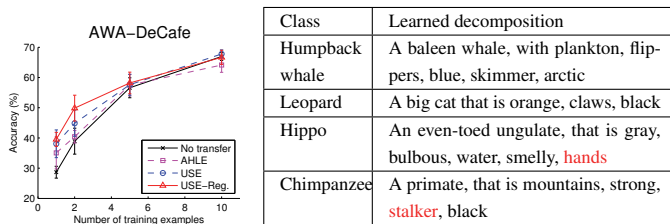


Figure 2: Few-shot experiment result on the AWA dataset, and generated semantic decompositions.

richer description than existing methods that estimates the new input category using only trained categories, or attributes. For this experiment, we divide the 50 categories into predefined 40/10 training/test split, and compare the knowledge transfer using AHLE. We assume that no attribute labels are provided for test classes. For AHLE, and USE, we regularize the learning of \mathbf{W} with \mathbf{W}^S learned on training class set \mathcal{S} by adding $\lambda_2 \|\mathbf{W} - \mathbf{W}^S\|_2^2$, to LME (Eq.). For USE-Reg, we use the reconstructive regularizer to regularize the model to generate semantic decomposition using U^S . While all methods made improvements over the no-transfer baseline, USE-Reg achieves the largest improvement, improving two-shot result on AWA-DeCafe from 38.93% to 49.87%. Most learned reconstruction look reasonable, and fit to discriminative traits that help to discriminate between the test classes, which in this case are colors (See Figure 2).

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-Embedding for Attribute-Based Classification. 819–826.

Bengio, S.; Weston, J.; and Grangier, D. 2010. Label Embedding Trees for Large Multi-Class Task. In *NIPS 2010, Twenty-Fourth Annual Conference on Neural Information Processing Systems*.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing Objects by their Attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Frome, A.; Corrado, G.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In *Proceedings of the Neural Information Processing Systems*.

Gao, T., and Koller, D. 2011. Discriminative learning of relaxed hierarchy for large-scale visual recognition. *Computer Vision, IEEE International Conference on* 0:2072–2079.

Griffin, G., and Perona, P. 2008. Learning and using taxonomies for fast visual categorization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8.

Hwang, S. J.; Sha, F.; and Grauman, K. 2011. Sharing features between objects and their attributes. In *CVPR*, 1761–1768.

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

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A. 2008. Supervised dictionary learning. In *NIPS*, 1033–1040.

Marszalek, M., and Schmid, C. 2008. Constructing category hierarchies for visual recognition. In *Proceedings of the European Conference on Computer Vision*.

Weinberger, K. Q., and Chapelle, O. 2009. Large margin taxonomy embedding for document categorization. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Proceedings of the Neural Information Processing Systems*, 1737–1744.