Start from Scratch: Towards Automatically Identifying, Modeling, and Naming Visual Attributes

Hanwang Zhang[†], Yang Yang[‡], Huanbo Luan[†§], Shuicheng Yan[†], Tat-Seng Chua[†]
[†]National University of Singapore
[‡]University of Electronic Science and Technology of China
[§]Tsinghua University
{hanwangzhang,dlyyang,luanhuanbo}@gmail.com;{eleyans,dcscts}@nus.edu.sg

ABSTRACT

Higher-level semantics such as visual attributes are crucial for fundamental multimedia applications. We present a novel attribute discovery approach that can automatically identify, model and name attributes from an arbitrary set of image and text pairs that can be easily gathered on the Web. Different from conventional attribute discovery methods, our approach does not rely on any pre-defined vocabularies and human labeling. Therefore, we are able to build a large visual knowledge base without any human efforts. The discovery is based on a novel deep architecture, named Independent Component Multimodal Autoencoder (ICMAE), that can continually learn shared higher-level representations across the visual and textual modalities. With the help of the resultant representations encoding strong visual and semantic evidences, we propose to (a) identify attributes and their corresponding high-quality training images, (b) iteratively model them with maximum compactness and comprehensiveness, and (c) name the attribute models with human understandable words. To date, the proposed system has discovered 1,898 attributes over 1.3 million pairs of image and text. Extensive experiments on various real-world multimedia datasets demonstrate the quality and effectiveness of the discovered attributes, facilitating multimedia applications such as image annotation and retrieval as compared to the state-of-the-art approaches.

Categories and Subject Descriptors

H.3.3 [**Content Analysis and Indexing**]: Abstracting methods

Keywords

attribute discovery; deep learning; multimodal analysis

1. INTRODUCTION

With the evolution of machine understanding of visual attributes (*e.g.*, concepts, objects, and visual patterns) over the years[1], we have witnessed the advances in content-based multimedia applications such as annotation and retrieval, which have been reforming the way we interact with the world [26, 40]. It is believed that the success is mainly derived from the intelligent machinery built on large and diverse databases with high-quality attribute annotation. But constructing such databases is a challenging and expensive task. Recent studies, using crowdsourcing and bootstrapping strategies to collect attributes from the Web data at a vast scale with minimum human intervention, offer a promising route towards this task [8, 24, 28].

However, as we migrate these collected attributes to real-world applications, we may find that they adapt poorly to the ever-evolving visual world [35]. One fundamental issue is that until recently researchers have no idea on how many attributes are enough for multimedia applications. They generally resort to a pre-defined vocabulary constructed by domain experts. Since it is manually built with limited scale, the resultant attributes can hardly keep up with the ever-evolving interests of the general users. For instance, the overlap between the items of ImageNet [8]/LSCOM [24] and user-provided tags crawled from Flickr is only 1.4%/7.6%,

Vocabulary	ImageNet	LSCOM
#item	117,023	483,034
#item∩tag (% of #tag)	79,215 (1.42%)	427,440 (7.66%)

(a) Overlap of Pre-defined Vocabularies with User-Generated Tags

(b) Images Collected by Search Engine and User Tags

Figure 1: Problems with the conventional Web-scale attribute learning paradigm. (a) The pre-defined vocabularies (*i.e.*, items) hardly cover the user-generated tags with 623,464 non-repeated terms from Flickr. (b) Images collected from Google search engine hardly represent user-generated images from Flickr. Images averaged over the most relevant results are shown on the left.

[1]Besides the conventional attribute definition such as object parts, visual properties [12], we refer the term "attribute" to more general semantic meanings such as concepts, objects.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654915.

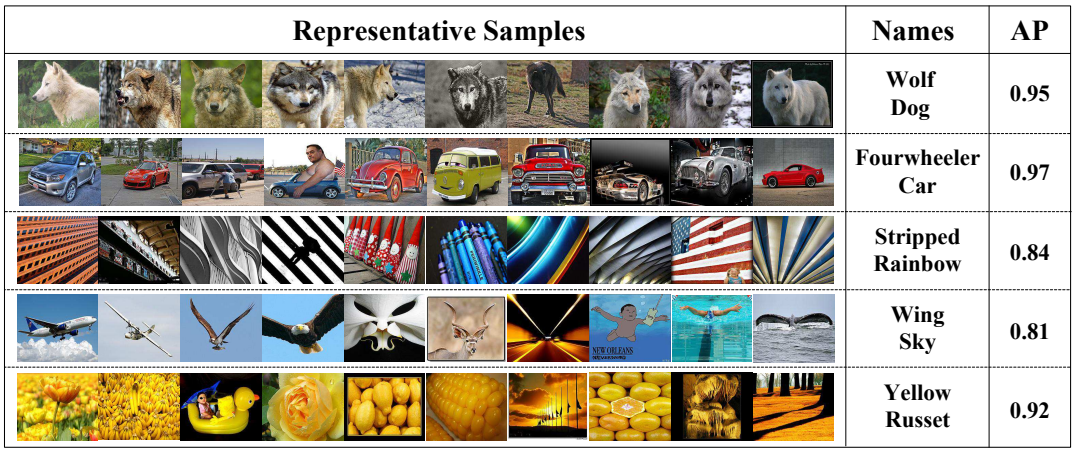


Figure 2: Illustrative examples of 5 discovered attributes. Our approach can automatically identify, model and name attributes without any human efforts. We show 10 representative images with the most model confidence, top 2 names and the performances of attribute models in terms of Average Precision (AP).

as shown in Figure 1(a). On the other hand, due to the noisy tags of the Web images, people exploit search engines to collect much more relevant images given the attributes as queries and invite human labelers to refine the results. However, despite the expensive human labor, the search results may not be informative and comprehensive because of duplicate or near-duplicate visual contents of images, resulting in attribute models heavily biased towards them [16]. For example, as shown in Figure 1(b), given certain attributes as queries, images from a photo sharing community (*e.g.*, Flickr) and a search engine are very different (*e.g.*, Google), causing the models built on the latter can hardly adapt to applications in the former, which nevertheless has begun to dominate the Web content [7].

Intrigued, we raise a question: is it possible to start from scratch to discover and build up the visual attributes directly from the data? In other words, we want to develop an intelligent system that can automatically acquire user-interested attributes solely from images and tags, which can be easily crawled on the Web without any constraints. Here, we refer the "unconstrained crawling" to dataset collection from any sources, rather than need to leverage on search engines to gather data related to a pre-defined set of attributes. In this paper, we describe a novel approach that automatically harvests visual attributes *without any human supervision*. Figure 2 shows some examples of attributes discovered by our system. We want to highlight three key components of our approach that make it effective and distinguishable from state-of-the-art methods [6, 20].

Identification. We do not need any pre-defined vocabulary of semantic attributes or search engines to collect attribute-related images. Instead, our system can automatically identify a set of attributes from a large number of image-text pairs. Motivated by the psychological and cognitive findings [23] that attributes are shared and constructive higher-level abstractions of multi-source sensory perceptions, we develop an unsupervised deep architecture that correlates image and textual data through the shared hidden layer, where the variables are encouraged to be de-correlated from each other. In this way, we are able to fully exploit the rich but noisy visual and semantic information of Web-scale multimodal data to discover useful patterns. Then, the hidden variables are expected to be informative semantic representations for both modalities and are thus ideal surrogates for identifying attributes. For example, a hidden variable may jointly represent the occurrence of the words "furry dog" and the visual properties of a furry dog image.

Modeling. As compared to recent automatic attribute modeling methods that rely on search engines to refine attribute training images [6, 20], our modeling approach incorporates the inherent correlations of visual and textual data encoded in the shared hidden variables, resulting in training images of larger diversity and leading to more generalizable attribute models that are evident in both modalities. Moreover, we propose a model update mechanism for the new data. Hence, besides discovering new attributes, our system can also update old attribute models by merging new but redundant ones. Therefore, the attributes can evolve into a more compact and comprehensive knowledge base from the inexhaustible amount of data.

Naming. Yet, the attributes discovered are not applicable to end-user applications since they have no meaningful utterable names. In order to make the machine-recognized attributes understandable by humans, we propose to name the attributes by using the associated tags of images. For each attribute, we rank the tags as tentative attribute names not only based on their noisy frequencies but also on their relatedness to the corresponding visual properties, namely visualness. This results in more accurate and informative names. It is worth noting that our naming strategy does not rely on any pre-defined vocabularies but directly mines from the user-generated data. The key advantage of this approach is that the attributes are no longer limited in semantic scale and are able to cover general user interest.

The overview of the proposed fully automatic approach is illustrated in Figure 3. We start with nothing but an arbitrary set of image-text pairs as the input data. In order to mine the correlations of the multimodal data, we propose a novel unsupervised deep architecture, called Independent Component Multimodal Autoencoder (ICMAE), which underpins the overall automatic system. ICMAE has two pathways for image and text modalities. The two pathways are then merged into a shared hidden layer. Given the higher-level representations of both pathways (*i.e.*, "Layer 2" as shown in Figure 3), we can learn higher-order correlations across modalities through the shared layer.

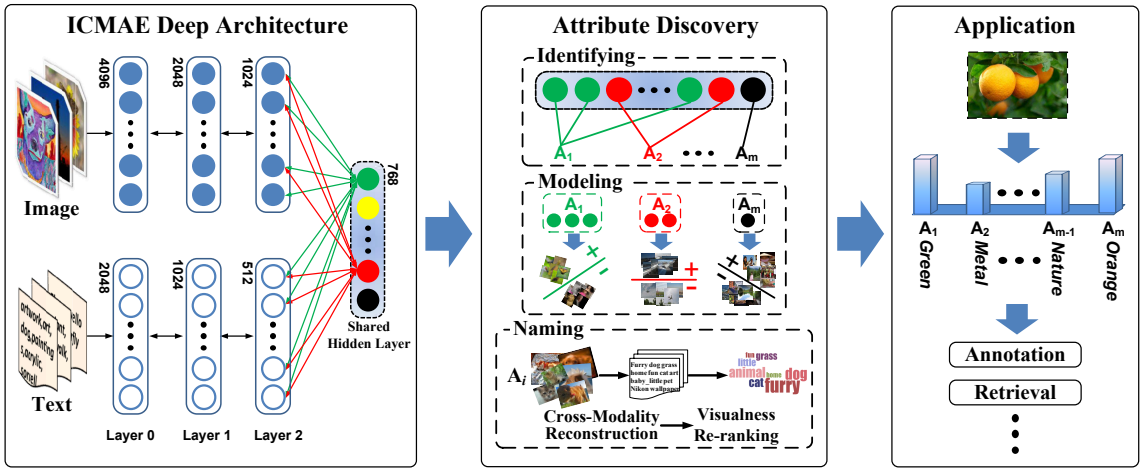


Figure 3: The overview of the proposed system towards automatically identifying, modelling and naming visual attributes. The proposed unsupervised deep architecture, Independent Component Multimodal Autoencoder (ICMAE), underpins the overall system. The number of neurons in each layer is marked.

Based on ICMAE, the attribute discovery works as follows. First, in order to constrain the hidden variables of the shared layer as potential attributes to be **identified**, we propose to impose Independent Component Analysis (ICA) constraints in the architecture to de-correlate the relationships among the variables (colored connections in Figure 3). Then, given a set of images, the response values of the hidden variables of an attribute indicate how strong the attribute is present, guiding us to divide the images into two clusters, namely, positive and negative samples, which are used to **model** the attribute. Finally, for **naming** an attribute, we provide the most confident images of the attribute model as input to the image pathway of ICMAE to reconstruct the textual counterpart. The reconstructed textual data of the images are considered as most visually related to the attribute. Together with their term frequencies, we finally score the text as potential attribute names. So far, the system has discovered 1,898 attributes from 1.3M images and tags pairs. Experimental results on real-world datasets demonstrate that the discovered attributes are effective in supporting fundamental multimedia applications such as image annotation and retrieval.

In the era of big data, we believe that our work has a great potential in relieving human labor for learning visual semantics, due to the following contributions.

- We propose a novel attribute discovery system that can constantly and automatically identify, model and name visual attributes from the exhaustible amount of Web data. To the best of our knowledge, this is the first work to explore the possibility of a fully machine-built visual knowledge base.
- We develop a novel deep architecture named ICMAE that effectively mines the higher-level correlations of noisy multimodal data from the images and the corresponding tags.
- Through fundamental applications such as image annotation and retrieval, we demonstrate that the attributes discovered by our automated system considerably outperform others with human efforts.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the machine learning framework of the proposed ICMAE. Section 4 illustrates the proposed automatic attribute discovery approach based on ICMAE, including identifying, modeling, and naming attributes. Experimental results and analysis are reported in Section 5, followed by conclusions in Section 6.

2. RELATED WORK

2.1 Attribute Discovery

The research trend of learning semantic attributes has changed from using small benchmark datasets to large-scale knowledge base, towards real-world applications [35, 38]. However, constructing a desired large dataset is challenging. With the development of crowdsourcing techniques such as Mechanical Turk, large-scale high-quality datasets in terms of size and diversity such as ImageNet [8], LSCOM [24] and Visipedia [28] are constructed at a reasonable cost. Such valuable datasets have successfully spawned many widely-used attribute models such as Classemes [36], Picodes [4], and DeCAF [10]. On the other hand, several studies focus on automatically mining attributes from the Web. For example, Li *et al.* [20] used a multiple instance learning framework to learn sub-image-level features from images returned by querying attribute names from the search engine. Moreover, some bootstrapping strategies are applied to improve the quality of the collected data and their induced attribute models. Li *et al.* [19] developed a system that incrementally updates attribute models, which in turn refines the collected data. Chen *et al.* [6] developed a never-ending image learner that cycles between discovering attribute relationships and then retrain attribute models constrained by the relationships.

However, all the above methods start from a pre-defined vocabulary with poor coverage on general user interests and rely on the less informative images retrieved by search engines. Alternatively, we are able to harvest attributes directly from image-text pairs without the help of any vocabularies and search engines. In this way, our work relates to several recent studies, such as Bergamo *et al.* [4] and Rastegari *et al.* [30] that considered attributes as random split [12] of the visual space enhancing the classification performance. But these discovery methods require class-level image annotations and the discovered attributes have no human understandable names. Instead, our approach does not only dis-

cover them but also name them explicitly. Parikh and Grauman [27] invited humans in the discovering process to name attributes by active learning strategy. Similarly, Kankuekul *et al.* [17] incrementally labeled attributes via online interaction with users. As compared to these semi-automatic methods, our naming is fully automatic. There are also automatic attribute discovery and naming methods that fall into specific domains like butterfly [37] and fashion productsbengio2009learning recogntion [3]. However, we tackle the problem in general domain which is more fundamental and challenging.

2.2 Multimodal Deep Learning

Recent advances in machine learning community have examined that deep architecture can be trained to produce useful representations for visual [14], acoustic [9] and textual [21] modalities. However, there are only few studies focusing on multimodal deep learning. Ngiam *et al.* [25] proposed a deep autoencoder that jointly learns shared representations from visual and acoustic modalities. This model correlates the multimodalities by reconstructing one modality given that others missing. Our deep architecture follows this cross-modality reconstruction strategy since it is very similar to the reconstructive and reusable properties of attributes. Andrew *et al.* [1] proposed to use Canonical Correlation Analysis criterion in the overall fine-tuning of independent deep networks of multimodalities. But this model is not reconstructive and hence fails to learn representations optimized for attributes. Srivastava and Salakhutdinov [32] developed a multimodal Deep Boltzmann Machines that jointly model the image and text data. This model learns similar representations as that of Ngiam *et al.* [25] but it may suffer from the early-stopping issue since there is no explicit objective when training Deep Boltzmann Machines. It is worth noting that the learnt representations by the above models are not optimized for identifying informative attributes since the shared representations are highly correlated, resulting in very redundant attributes. Therefore, we propose to incorporate the Independent Correlation Analysis criterion [18] in a deep Multimodal Autoencoder to overcome this issue.

3. INDEPENDENT COMPONENT MULTIMODAL AUTOENCODER

In this section, we introduce the architecture and details of the proposed Independent Component Multimodal Autoencoder (ICMAE), which underpins the subsequent attribute discovery. Technically, learning ICMAE includes layer-wise pre-training and overall fine-tuning. Note that this learning procedure can be applied once again to update the old ICMAE if a new batch of crawled data comes in. Here, we use the parameters of the old model as initializations for the new round of pre-training using the new data. Thus, we can always obtain an updated ICMAE that adapts to the ever-evolving visual world. Without loss of generality, in this section, we only introduce how we train ICMAE based on one batch of training data.

3.1 Architecture Overview

The basic architecture of the network we use is similar to the one used in [25]. As shown in Figure 3, it has two parallel pathways for two input modalities and one shared hidden layer. Each input datum is a feature pair of image $\mathbf{x}^0 \in \mathbb{R}^{4096}$ and its associated text $\mathbf{t}^0 \in \mathbb{R}^{2048}$ (cf. Section 5.1.2 for details of feature extraction from noisy data). Then, $(\mathbf{x}^0, \mathbf{t}^0)$ is *encoded* (*i.e.*, fed forward) to the shared layer along both pathways, namely, $(\mathbf{x}^0, \mathbf{t}^0) \rightarrow (\mathbf{x}^1, \mathbf{t}^1) \rightarrow (\mathbf{x}^2, \mathbf{t}^2) \rightarrow \mathbf{z}$, where the dimensions of the variables $\mathbf{x}^1$ (or $\mathbf{t}^1$) and $\mathbf{x}^2$ (or $\mathbf{t}^2$) are respectively 2048 (or 1024) and 1024 (or 512), and the dimensions of the hidden variable $\mathbf{z}$ is 768. Here, we gradually reduce the dimensionality through each pathway in order to learn higher-level, nonlinear abstractions [14], and merge them into the shared hidden layer to model the relationships between the two modalities. Finally the stimuli of the shared layer is *decoded* (*i.e.*, fed backward) through the two pathways to reconstruct the input data of both modalities, namely, $\mathbf{z} \rightarrow (\widehat{\mathbf{x}}^2, \widehat{\mathbf{t}}^2) \rightarrow (\widehat{\mathbf{x}}^1, \widehat{\mathbf{t}}^1) \rightarrow (\widehat{\mathbf{x}}^0, \widehat{\mathbf{t}}^0)$. Such multimodal deep autoencoder aims to jointly model the distributions of the image and text data, resulting in shared hidden variables that have strong connections to variables from both modalities. Interestingly, this property meets the meanings of semantic attributes [23] and therefore the variables of the shared hidden layer are ideal as attribute candidates. Next, we detail the encoding and decoding functions of ICMAE.

Without loss of generality, we only formulate the image pathway. The encoding functions are computed as

$$\begin{aligned} &\mathbf{x}^1 = \sigma(\mathbf{W}_v^0 \mathbf{x}^0 + \mathbf{c}^1),\ \mathbf{x}^2 = \sigma(\mathbf{W}_v^1 \mathbf{x}^1 + \mathbf{c}^2), \\ &\mathbf{z} = \sigma\left([\mathbf{W}_v^2\ \mathbf{W}_t^2] \begin{bmatrix} \mathbf{x}^2 \\ \mathbf{t}^2 \end{bmatrix} + \mathbf{c}^3\right), \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ is the element-wise sigmoid function that has been shown to be useful for autoencoders [14]. $\mathbf{W}_v^i$ and $\mathbf{c}^i$, $i = 0, 1$ are trainable weights and encoding biases of the image pathway, respectively. $\mathbf{W}_v^2$ and $\mathbf{W}_t^2$ are the weights of the image and text pathways, respectively. $\mathbf{c}^3$ is the bias of the shared layer. The decoding functions are computed as

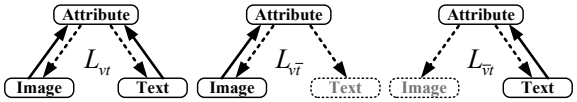
$$\widehat{\mathbf{x}}^2 = \sigma({\mathbf{W}_v^2}^T \mathbf{z} + \mathbf{b}^2),\ \widehat{\mathbf{x}}^1 = \sigma({\mathbf{W}_v^1}^T \widehat{\mathbf{x}}^2 + \mathbf{b}^1),\ \widehat{\mathbf{x}}^0 = {\mathbf{W}_v^0}^T \widehat{\mathbf{x}}^0 + \mathbf{b}^0, \tag{2}$$

where $\mathbf{b}^i$ is the decoding bias. Note that the decoding function from Layer 1 to the input Layer 0 is not sigmoidal. This is because we need to retain the intrinsic Gaussian distribution for the features of the input data.

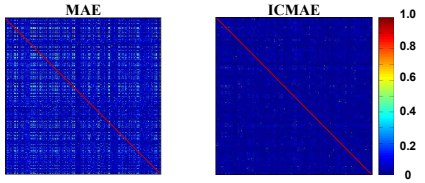
3.2 Layer-wise Pre-training

It is well-known that deep architecture only works well if the trainable parameters are properly initialized to a good solution. In this section, we introduce how to use the Restricted Boltzmann Machine (RBM) [2] to pre-train the proposed ICMAE. The RBM is an undirected graphical model that connects two layers of random variables. Without loss of generality, we denote $\mathbf{v}$ as visible variables and $\mathbf{h}$ as hidden variables corresponding to any two connected layers in ICMAE. In particular, for the combined image and text variables at Layer 2, we have $\mathbf{v} \leftarrow [{\mathbf{x}^2}^T, {\mathbf{t}^2}^T]^T$, and for the shared hidden layer, we have $\mathbf{h} \leftarrow \mathbf{z}$. The connections in RBM are parameterized by $\mathbf{W}$ between $\mathbf{v}$ and $\mathbf{h}$. The optimization is to minimize the negative logarithm of the likelihood $p(\mathbf{v}) = \sum_{\mathbf{h}} p(\mathbf{v}, \mathbf{h}) = \sum_{\mathbf{h}} \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}$, where Z is a partition function.

One form of RBM is to assume that $\mathbf{v}$ and $\mathbf{h}$ are $\{0, 1\}$-valued binary variables. This RBM type is consistent with the sigmoidal decoding/encoding functions in Eq. (1) and (2).



(a) Cross-modality Reconstruction



(b) Correlation of Hidden Variables

Figure 4: Illustration of the two objectives of ICMAE. (a) Three types of cross-modality reconstruction error as shown in Eq. (3). The solid and dashed arrows denote encoding and decoding, respectively. The lighter font color denotes missing modalities. (b) Visualizations of the correlation matrices of the resultant hidden variables of MAE (Multimodal AE [25]) and the proposed ICMAE. We can see that most of the hidden variables of ICMAE are de-correlated with each other by the use of independent component constraints as shown in Eq. (4).

Another form of RBM is to assume $\mathbf{h}$ as binary but $\mathbf{v}$ as real-valued Gaussian variables with unit variance and is consistent with the linear encoding/decoding functions in Eq. (2). Therefore, we use the Gaussian-binary RBM to initialize the parameters that connect between Layer 0 and Layer 1 of ICMAE and use the binary-binary RBM to initialize parameters of other layers. In total, we train 5 RBMs (4 along the two pathways, and 1 for the joint layer and the shared hidden layer) for ICMAE pre-training. As it is intractable to compute the gradient of the log-likelihood, we learn the parameters of the RMBs using contrastive divergence as in [34]. Moreover, in order to learn sparse models in an efficient way, we set the initial biases of the RMBs as sufficiently small (*e.g.*, -2).

3.3 Overall Fine-tuning

After pre-training ICMAE by using RBMs as above, we are able to fine-tune the entire ICMAE using stochastic gradient descent to minimize the objective function of ICMAE. If we only define the objective function as the reconstruction error between the mutimodal input $(\mathbf{x}^0, \mathbf{t}^0)$ and output $(\widehat{\mathbf{x}}^0, \widehat{\mathbf{t}}^0)$, the deep autoencoder may easily overfit to each of the modality and fail to learn useful cross-modality correlations. For example, it is possible for shared hidden layer to find representations such that some of the variables are tuned only for images while others are tuned only for tags. So, the shared hidden variables may result in attributes that are only responsive to one single modality, thus losing important information for the subsequent attribute discovery. Inspired by [25], we propose to encourage cross-modality reconstruction. As shown in Figure 4(a), given one modality present and the other one absent, we hope to reconstruct both modalities. According to the encoding/decoding functions in Eq. (1) and (2), we can see that the reconstructed of any modality is a function of both input modality, *i.e.*, $\widehat{\mathbf{x}}^0 = \widehat{\mathbf{x}}^0\left(\mathbf{x}^0, \mathbf{t}^0\right)$. For example, we can define the reconstruction error when the image modality is present while the text modality is absent as

$$L_{v\bar{t}} = \|\mathbf{x}^0 - \widehat{\mathbf{x}}^0\left(\mathbf{x}^0, \mathbf{0}\right)\|_2^2 + \|\mathbf{t}^0 - \widehat{\mathbf{t}}^0\left(\mathbf{x}^0, \mathbf{0}\right)\|_2^2, \quad (3)$$

where the subscript $v\bar{t}$ denote the presence of the image and the absence of text modalities. L_{vt} and $L_{\bar{v}t}$ can be defined in a similar way.

On the other hand, we want to regularize the hidden variables to be statistically independent with each other. Intuitively, attributes in nature are independent semantic meanings that constitute the visual or semantical world [23]. For example, an image of "beach sunset" may be composed by independent attributes such as "round", "water", "sun", "horizontal line". Therefore, we propose to impose the Independent Component Analysis (ICA) criterion on the overall objective function of ICMAE. Different from traditional ICA [15] that optimizes a ℓ_1-norm penalty with "hard" orthonormality constraints, we use a "soft" version with the reconstruction cost [18], which meets our formulation in Eq. (3). In particular, the penalty term is the sum of the activation value of the hidden variables $\mathbf{z}$. For example, when the input modality is image while the text modality is absent, the ICA regularization term is defined as

$$I_{v\bar{t}} = \left\|\mathbf{z}(\mathbf{x}^0, \mathbf{0})\right\|_1, \quad (4)$$

and I_{vt}, $I_{\bar{v}t}$ can be defined in a similar way.

Hence, the overall objective function of ICMAE that we are going to minimize is

$$F\left(\mathbf{x}, \mathbf{t}; \mathcal{W}\right) = L_{vt} + L_{v\bar{t}} + L_{\bar{v}t} + \lambda(I_{vt} + I_{v\bar{t}} + I_{\bar{v}t}) + \gamma R(\mathcal{W}), \quad (5)$$

where $\mathcal{W} = \{\mathcal{W}^i\} = \{(\mathbf{W}_v^i, \mathbf{W}_t^i, \mathbf{b}^i, \mathbf{c}^i)\}$, $i = 0, 1, 2$, is the set of trainable parameters, $R(\mathcal{W})$ is an ℓ_2-norm regularizer, λ and γ are trade-off parameters. We update $\mathcal{W}$ using the stochastic gradient descent method with a dynamic momentum tric [29].

4. ATTRIBUTE DISCOVERY

The deep architecture ICMAE have mined the strong correlations of multimodal data and learnt useful higher-level representations, in which the attributes are encoded. In this section, we will detail how we identify, model and name attributes based on ICMAE.

4.1 Identification

Recall that we identify attributes without any pre-defined vocabularies. Instead, we identify them from the learnt shared hidden variables of ICMAE. Recall that such variables are (a) higher-level features which can reconstruct both modalities; and (b) as independent as possible with each other. On one hand, the hidden variables are high-level abstractions of visual and semantic evidences from a large amount of image-text data and are in turn encouraged to reconstruct these evidences [23]. For example, attribute "furry" is abstracted from visual and semantic cues such as furry animals. Meanwhile, "furry" is indispensable to compose those evidences. On the other hand, we humans in nature hope that attributes are compact and independent semantic meanings which compose the world. Therefore, the hidden variables are expected to be an ideal feature pool to extract attributes.

However, there are still correlated hidden variables relating to the same attribute. Hence, it is possible to identify several clusters, each of which holds the variables with large correlation values. We adopt the Affinity Propagation (AP) [13] that can cluster data given a pair-wise similarity matrix without fixing the number of clusters. In our case, the similarity matrix is the correlation matrix of the

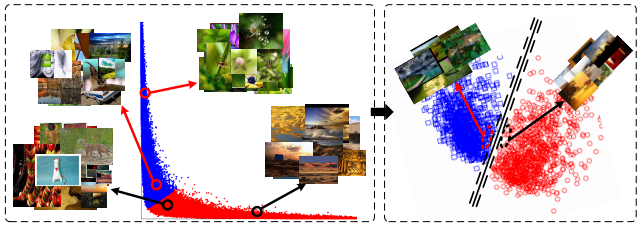


Figure 5: Illustrations of an attribute identified in the hidden space and modeled in the visual space. The attribute corresponds to two hidden variables, which are strongly correlated (left). Images that are sufficiently far from the MMC boundary are chosen. Then, the attribute is modeled by the chosen samples in the visual space using SVM (right, visualized by using PCA to 2-D). The support vectors of images (marked near the hyperplane) are considered as the attribute model representations, which are used to check model redundancy.

hidden variables based on training data. Then, each resultant cluster is considered as an attribute, denoted as a while its hidden variables is denoted as $\mathbf{a}$, where $\mathbf{a} \subseteq \mathbf{z}$. For an identified attribute a, we should further identify its positive and negative samples for the subsequent attribute modeling. To achieve this, we use Maximum Margin Clustering (MMC) [41] to automatically divide samples in the hidden space into two clusters with maximum marginal separation. Denoting $\{\mathbf{a}_i\}$ as the hidden representations of samples with respect to attribute a, MMC iteratively seeks for two clusters by Support Vector Regression (SVR) [31]. It is worth noting that, as shown in Figure 5, images associated with the hidden variables near the decision boundary are ambiguous and not as visually separable as those associated with the hidden variables far from the boundary. Therefore, for better attribute modeling, we identify the high-quality positive and negative images with respect to the attribute as $Pos(a) = \{\mathbf{x}_i | \mathbf{w}^T \mathbf{a}_\mathbf{i} > T\}$ and $Neg(a) = \{\mathbf{x}_i | \mathbf{w}^T \mathbf{a}_\mathbf{i} < -T\}$, where $\mathbf{w}$ is the model parameter of SVR and $T > 0$ is a positive threshold.

4.2 Modeling

Provided with the training samples $Pos(a)$ and $Neg(a)$ of attribute a, we use a linear Support Vector Machine (SVM) to model the attribute in the visual space. Due to the imperfection of AP clustering of the hidden variables, some attributes may still correlate with each other, resulting in redundant attribute models. In order to obtain a compact set of attribute models, we propose to merge redundant models.

Our merging strategy is based on checking the Mutual Information (MI) shared between model $\mathcal{M}_a$ of attribute a and model $\mathcal{M}_b$ of attribute b. MI measures how much information the knowledge of either $\mathcal{M}_a$ or $\mathcal{M}_b$ provides for the other. For instance, if the MI between $\mathcal{M}_a$ and $\mathcal{M}_b$ is small, it indicates that $\mathcal{M}_a$ provides minimal information for determining whether $\mathcal{M}_b$ is redundant as compared to $\mathcal{M}_a$. Specifically, we represent any model $\mathcal{M}$ by a set of images which are support vectors $\mathcal{S}$ of the model since the support vectors sufficiently characterize the decision boundary of the attribute model (see the right of Figure 5). Therefore, the MI between attribute model $\mathcal{M}_a$ and $\mathcal{M}_b$ is defined as

$$I(\mathcal{M}_a; \mathcal{M}_b) = \sum_{\mathbf{x}' \in \mathcal{S}_b} \sum_{\mathbf{x} \in \mathcal{S}_a} p(\mathbf{x}, \mathbf{x}') \log \left(\frac{p(\mathbf{x}, \mathbf{x}')}{p(\mathbf{x}) p(\mathbf{x}')} \right), \quad (6)$$

where we use kernel density estimation [22] to estimate the distribution $p(\mathbf{x})$, $p(\mathbf{x}')$ and $p(\mathbf{x}, \mathbf{x}')$. Note the estimation is efficient since the number of support vectors is relatively small as compared to the number of training samples. We accept to merge two attribute models if the MI between them is larger than a pre-defined threshold. For merging the two models, we first align the direction of their hyperplanes and then combine the samples on either side with response concensus on both models. Finally, we train a merged attribute model with the combined samples.

We can also use this merging strategy to update old attribute models by merging new but redundant ones which are discovered from a new batch of data. Thus, we will acquire more compact and comprehensive attribute models from the inexhaustible amount of Web data.

4.3 Naming

Yet, the discovered attributes are not assigned with any human nameable annotations. In this section, we show how to name attributes directly from the tags associated with representative images (*i.e.*, images with the most confidence of the attribute models). Since the attribute models are binary classifiers, representative images on both sides are considered. That is to say, we will assign *two* names for *one* identified attribute. As compared to standard attribute naming that only express one property (*e.g.*, "green" or "not green"), our naming is more descriptive. Note that this is reasonable since images on the opposite side of "flower" could be endowed various semantics and thus should not be coarsely named as "non-flower" [30].

A straightforward approach to name an attribute is to mine highly frequent tags associated with the representative images of the attribute. However, the user-generated tags are very noisy and thus are usually irrelevant to the visual properties of the attribute. For example, as illustrated in Figure 6, highly frequent words in user tags can be very general concept (*e.g.*, "nature", "sky"), device (*e.g.*, "cannon", "nikon"), or sentiment (*e.g.*, "a-big-fave", "diamond-class photographer"). Moreover, some obvious visual properties are unlikely to ever be mentioned by users. For example, the frequencies of "green", "violet" are low as shown in Figure 6. Therefore, naming attributes only from tag frequency is not a proper approach.

We propose to incorporate the visualness of words to name attributes. Here, visualness refers to how strong the textual words relate to visual properties. Conventional visualness measurement methods [39, 16, 3] still fundamentally rely on word statistics and hence cannot recover missing words with high visualness. Fortunately, thanks to ICMAE that correlates the cross-modality data of images and textual tags, we can directly choose tags which are responsive to the target visual attributes. First, as shown in Figure 6, we feed the representative images of the attribute into ICMAE and reconstruct the text counterpart. In particular, the textual reconstruction is formulated as $\widetilde{\mathbf{t}}^0(\mathbf{x}^0, \mathbf{0})$, which is the cross-modality reconstruction given images but missing text (cf. Section 3.3). Then, since each dimension of the textual data refers to several words (cf. Section 5.1.2), the value of the dimension is considered as the visualness of these words with respect to the attribute. As analogous to word frequencies, we can also visualize the visualness of words in Figure 6. We can see that words like flower names (*e.g.*, "fuchsia", "rosebuds") are strongly related to the images. Moreover,

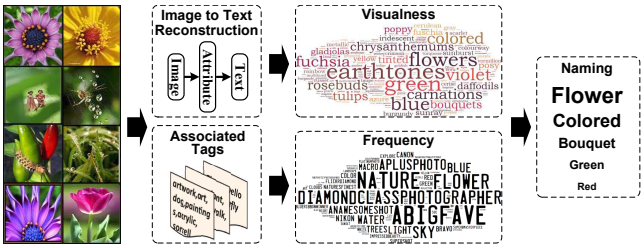


Figure 6: Illustration of attribute naming by jointly considering word visualness and frequency, which are visualized as word clouds. Images shown on the left are representative samples on one side of the attribute model.

some obvious properties like "colored", "green" and "red" have emerged. Finally, we respectively normalize visualness and frequencies, and add the two scores as the final scores for potential attribute names. Figure 2 illustrates the top 2 ranked names for several discovered attributes.

5. EXPERIMENTS

We systematically evaluate the effectiveness of the proposed ICMAE deep architecture and the quality of attribute identification, modeling and naming methods.

5.1 Data and Methodology

5.1.1 Experimental Data Sets

We crawled the image-text multimodal data from a photo-sharing website Flickr. The crawling protocol is as follows. First, we started from 500 seed users who contribute most interesting photos using Flickr API. Next, for each of the seed user, we select at most 100 users from his/her contact. This results in around 50,000 users. Finally, we crawled all their uploaded photos which have tags. In this paper, we used 1,321,496 pairs which are crawled in 10 continuous days of Jan, 2014. In order to examine how our approach can evolve to new data, we split the data into four batches of equal size from earlier to latter periods. We trained ICMAE and discovered attributes from one batch to another. The trained parameters of ICMAE using the former batch were used to initialize the training using the next one.

In order to quantitatively evaluate the quality of the discovered attributes, we conducted two fundamental multimedia tasks, *i.e.*, image annotation and content-based image retrieval, on two benchmark data sets: **MIR-Flickr** that contains 25,000 images across 24 concepts, and **NUS-WIDE** that contains 269,648 images across 81 concepts. For annotation experiments, we followed the "train/test" split protocol in the official documentary, *i.e.*, "15,000/10,000" for MIR-Flickr and "161,789/107,859" for NUS-WIDE. For retrieval experiments on both data sets, we randomly selected 10% images from each concept as query examples.

5.1.2 Visual and Textual Features

For image features, we adopted the 4096-D DeCAF generic visual feature [10], which is the activations of the 6-th layer of a deep CNN trained in a fully supervised fashion on ImageNet [8]. This feature has been demonstrated to be effective on modern benchmark data sets. For each batch of the image data, we normalized the visual features into a zero-mean unit-variance Gaussian distribution. For textual feature, we developed a novel feature towards extreme sparse and noisy tags. The extraction pipeline was conducted as follows.

First, we transformed each word in a tag list into a 300-D vector using Word2Vec [21], which is a model trained on Google News data of 100 billion words. Note that there are about 100,000 words from our data contained in this model. Second, we learnt a 2048-D dictionary from the overlapped 100,000 words using K-means. Third, for each word in a tag list, we obtained a 2048-D sparse codes using any off-the-shelf sparse coding toolbox. Finally, each tag list can be transformed into a 2048-D textual feature by max pooling all the words in it. As compared to traditional word-frequency feature which is extremely sparse (*e.g.*, 1-5 non-zero dimensions using most frequent 2,048 words), our textual features have 30-50 non-zero dimensions. The features have richer semantic meanings, resulting in better representation of textual data. We normalized the textual features into a zero-mean unit-variance Gaussian distribution.

5.1.3 Implementation Details

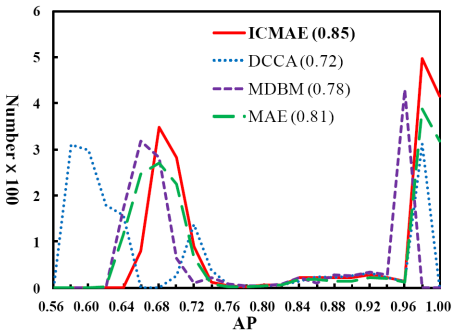
For training the RBMs (cf. Section 3.2), we initialized the weights uniformly as in the range of $[-0.05, +0.05]$ and the biases as -2. For stochastic gradient of the likelihood, we used Persistent Contrastive Divergence (PCD) [34] with 5-step Gibbs sampling of 100 Markov chains, and the coefficient of the ℓ_2-norm weight decay was set as 0.0001. There were 500 training data in a mini-batch. The initial learning rate was set as 0.1 and we decayed it by dividing 1.000015 at each epoch. We stopped the training of each RBM after 1000 epochs. It took about 5 days to train RBMs on each batch of the multimodal data along each modality pathway.

For training the ICMAE (cf. Section 3.3), we initialized the parameters using the pre-trained RBMs. The hyper-parameters λ and γ in Eq. (5) are set as 1 and 0.0001, respectively. The size of the mini-batch of the stochastic gradient descent is 500 and the initial learning rate was set to 0.001, which will be dynamically adjusted by monitoring the objective function. We stopped the training of ICMAE after 600 epochs. It took about 6 days to train on each batch of the multimodal data. The above deep architectures were built using PyLearn2 and trained on an NVIDIA TITAN 600 GPU.

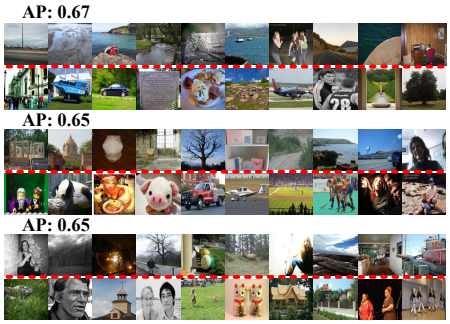
For attribute identification (cf. Section 4.1), the AP [13] and MMC [41] clustering was implemented by using the codes provided by the authors. Besides using the default parameters, we set the preference in AP algorithm as 0.5 to obtain clusters of hidden variables with sufficient correlations. The threshold T was set as such that the selected samples contribute at least 90% of the sum of the confidence scores. For attribute modeling (cf. Section 4.2) and image annotation experiments (cf. Section 5.2.2), we used the linear SVMs and tuned the trade-off parameter within the set $\log C \in \{-3, -2, ..., 3\}$ on the 10% training data, which were used as the validation set.

5.1.4 Compared Methods

To evaluate the effectiveness of the proposed **ICMAE** in attribute identification, we compared it against the following three state-of-the-art multimodal deep architectures: a) **DCCA**, Deep Canonical Correlation Analysis [1], which exploits Canonical Correlation Analysis to connects the hidden layer of the two pathways of multimodal data in order to fine-tune the overall deep networks, b) **MDBM**, Multimodal Deep Boltzman Machine [32], which connects two Deep Boltzmann Machine of multimodalities by a shared



(a) AP Distribution of Attributes



(b) Attributes with Low AP

Figure 7: Evaluations of the attribute identification. (a) Distributions of the number of the discovered attributes by four multimodal deep architectures over a range of AP, with mAP shown in bracket. (b) 10 representative images on both side of the model hypothesis (red dashed line) are shown.

hidden layer, and c) **MAE**, Multimodal Autoencoder, [25] which connects two deep autoencoders of multimodalities by a shared hidden layer. For fair comparisons, all the compared deep architectures have the same number of layers and the same number of neurons in each layer, and we used the same pre-training methods (cf. Section 3.2). We used Average Precision (AP) of the discovered attribute models to evaluate the quality, *i.e.*, the strength of the visual separability of the automatically identified attributes.

To evaluate the effectiveness of attribute modeling, we represented images by the responses of the discovered attribute classifiers. Such representations of images can be considered as semantic features, which are expected to support high performance image annotation and retrieval tasks. We compared our semantic features, named **Auto**, with three state-of-the-art semantic features: a) **Classeme** [36], the 2659-D output of classifiers trained on images collected from Bing Image Search by querying 2,659 concepts from LSCOM [24], b) **DeCAF** [10], the 1000-D output of a deep CNN trained on 1,000 classes of ImageNet [8], and c) **PiCoDes** [4], the 2048-D binary codes obtained by optimizing the classification across 2,659 classes of ImageNet. All the above three compared features were extracted by the software provided by the authors. For annotation, we trained 1-vs-all concept classifiers using the semantic features and used the mean AP (mAP) of the classifiers as the evaluation metric. For retrieval, all the query images and the gallery images are represented using these features and then the retrieval was then performed by similarity search. In particular, we used ℓ_1-norm distance for Classeme, DeCAF and the proposed Auto since they are classifier outputs and used Hamming distance for PiCoDes since it is binary code. We used the mAP at top K, for $K \in \{1, .., 100\}$, and averaged over all the queries as the evaluation metric for retrieval.

To evaluate the quality of attribute naming, we conducted two experiments: quantitative evaluation and human evaluation. For quantitative evaluation, given an image in MIR-Flickr or NUS-WIDE, we calculated the average cosine similarities of the predicted attribute names and the ground-truth labels. Specifically, we applied all the discovered attribute classifiers on an image and collected the corresponding top 5 names of each attribute. Recall that each attribute name or the ground-truth label can be represented by a 300-D vector, we then calculated the cosine similarity between any pair of a name and the label. We averaged the similarities over all the pairs as the final similarity between the attribute names and the ground-truth label. For human evaluation, we invited 30 graduate students and showed them 20 representative images and top 5 names of each attribute. We asked them to judge whether the naming is "good", "fair" or "bad". We compared the proposed naming strategy using term frequency and visualness (cf. Section 4.3), named **Freq+Visual** with that using only term frequency (**Freq**) or visualness (**Visual**). Note the Freq naming method is widely adopted in state-of-the-art attribute naming approaches [37, 3].

5.2 Experimental Results

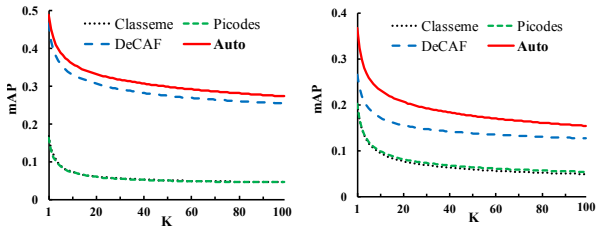
5.2.1 *Evaluations of Identification*

Figure 7(a) plots the distributions of the number of the discovered attributes by four multimodal deep architectures: the proposed ICMAE, DCCA [1], MDBM [32] and MAE [25]. We can see that all the multimodal deep learning methods can identify a lot of attributes with good AP (*e.g.*, above 0.9). This demonstrates that jointly deep learning representations from visual and textual modality is useful in identifying meaningful visual patterns. In particular, there are 1898, 1691, 1502 and 1778 attributes identified by ICMAE, DCCA, MDBM and MAE, respectively. We can also see that the discovered attributes by ICMAE have stronger visual evidences in terms of the number of good attribute models and the overall mAP. The superiority of ICMAE to other methods arises from the following aspects. a) ICMAE is designed to learn strong cross-modality representations that can reconstruct any of the modality while the other is missing. This guarantees that the identified attributes are good higher-level abstractions from both visual and textual sources, resulting in high-quality labeled images which are strongly related to attributes. b) ICMAE aims to learn hidden representations that are as independent as possible and thus the resultant attributes will have less redundant information. This helps to identify a compact and comprehensive set of attributes.

Figure 2 shows some successful examples of the discovered attributes. However, there are also some attributes with relative low AP (*e.g.*, below 0.7). Figure 7(b) illustrates some examples of attributes with AP around 0.65. We can observe that there is no consistent visual patterns on either side of the attribute model hypothesises. This indicates that the corresponding hidden variables fail to capture any stable semantic meanings. One possible reason is that these variables are isolated from clusters to which they are supposed to belong, due to the imperfection of AP clustering.

Table 1: Performance (mAP%) of image annotation on MIR-Flickr and NUS-WIDE. Note that our Auto with simple linear SVM considerably outperforms the best published results (43.67% on MIR-Flickr [33] and 27.1% on NUS-WIDE [11]), which were obtained by complicated features and classification models.

Dataset/Method	Classeme	PiCoDes	DeCAF	Auto
MIR-Flickr	15.90	15.98	52.67	**61.02**
NUS-WIDE	24.09	20.29	27.08	**32.71**



(a) MIR-Flickr (b) NUS-WIDE

Figure 8: Performance (mAP@K) of image retrieval over 10% random queries of MIR-Flickr and NUS-WIDE

5.2.2 *Evaluations of Modeling*

Table 1 and Figure 8 respectively shows the performance of the four semantic features in two fundamental multimedia tasks: image annotation and retrieval. In particular, Figure 10 presents the detailed annotation and retrieval results over all the concepts. Due to page limit, we here only report the 81 concepts of NUS-WIDE. It is worth noting that as compared to the three semantic features which require huge human efforts, *e.g.*, exploiting manually built knowledge base such as ImageNet [8], the proposed feature Auto was extracted from 1,898 attributes discovered in a fully automatic way. From the results, we can observe that Auto considerably outperforms the other three state-of-the-art semantic features. The reasons are two folds. First, our attributes were discovered automatically without the limitations of a pre-defined vocabulary. Hence, the derived feature Auto has a better representation of image semantics. Second, our discovery does not rely on the results of any search engine, which may only provide duplicate and simple images with respect to an attribute. In contrast, we directly mine an arbitrary set of Web images, resulting in attribute models that retain the intricate visual patterns of the visual world.

Table 2: Averaged cosine similarities between the predicted attribute names and the ground-truth labels for images of the combined NUS-WIDE and MIR-Flickr

Method	Freq	Visual	Freq+Visual
Avg. Cos. Similarity	0.10	0.21	**0.28**

5.2.3 *Evaluations of Naming*

Table 2 lists the averaged cosine similarities between the top 5 predicted attribute names and the ground-truth labels for images in NUS-WIDE and MIR-Flickr. From the results, we can see that Freq+Visual improves the naming quality significantly to 0.28. Note that this value suggests that the attribute names are generally relevant to the ground-truth labels. For reference, the cosine similarity between "car" and "fourwheeler" is 0.48 and that between "airplane" and "sky" is 0.23. Interestingly, using solely visualness can obtain 0.21 similarity. This demonstrates that ICMAE can recover missing texts meaningfully using only visual data. Figure 9 shows the average percentage of "Good", "Fair" and "Bad" attribute naming judgement from 30 regular users over the discovered 1,898 attributes. We can clearly see that the users are more satisfied with the Freq+Visual naming strategy. The better performance of jointly considering term frequency and visualness is because the visualness of words obtained by textual reconstruction can effectively eliminate the irrelevant but highly frequent words in the very noisy user-provided tag set.

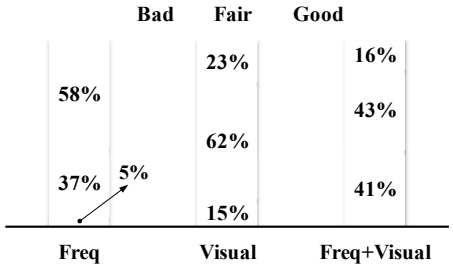


Figure 9: Average percentage of "Good", "Fair", "Bad" attribute naming judgement from the 30 regular users over the discovered 1,898 attributes

6. CONCLUSIONS

We presented a novel automatic attribute discovery approach that can automatically identify, model and name attributes from the inexhaustible image and tag pairs on the Web. We started from scratch without any pre-defined vocabulary or human labelling. Hence, our approach requires no human efforts. In particular, We proposed a novel deep architecture, called Independent Component Multimodal Autoencoder (ICMAE), to mine useful higher-level representations, based on which we developed strategies to identify, model and name attributes. So far, 1,898 attributes have been discovered from 1.3 million crawled data pairs. Extensive experiments on benchmark datasets have demonstrated the feasibility and effectiveness of the proposed approach, which has a great potential in relieving human labor for learning visual semantics from big Web data. In the near future, we will launch a live system based on this work. Peer researchers are welcome to download the attribute models to facilitate their research.

Although this work is an ambitious attempt to build a fully machine-built visual knowledge base, it has left the following two open issues not addressed. First, the system is not a real end-to-end system because we rely on "engineered" but not pixel- or word-level "raw" features. Second, there is still a large gap between human and machine attribute naming qualities (*e.g.*, our result is only 0.28 as compared to 0.50, which is the result we empirically find that human can achieve), especially the names corresponding to attribute models of relatively low reliability. One possible solution for the first issue is to connect CNNs to the inputs of the two pathways [5]. However, it is still unknown on how to design a proper learning strategy to synchronize the behavior of the multimodal heterogeneous network. For the second issue, it is mainly due to the limitation of multimodal correlation. Therefore, we will investigate more effective corre-

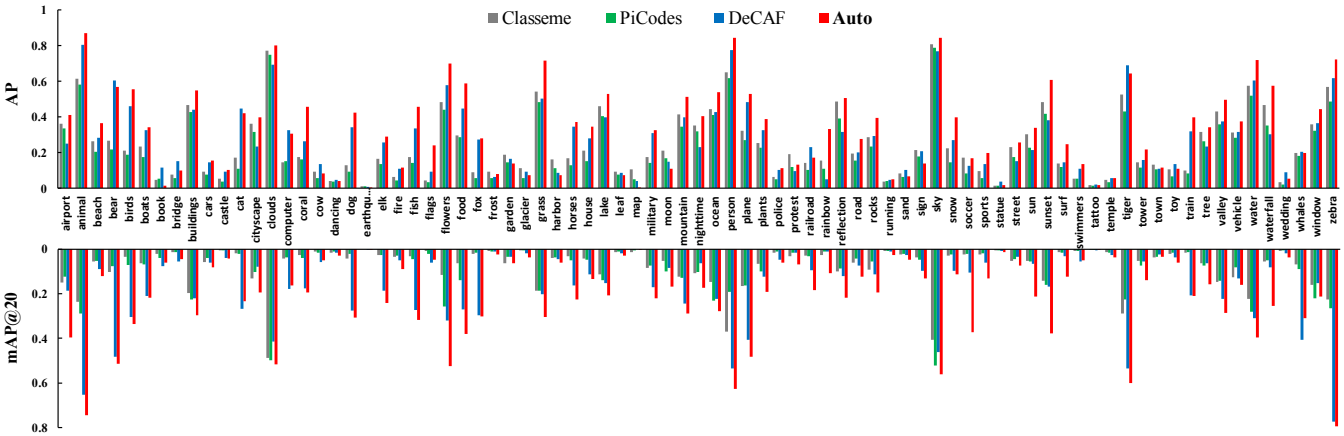


Figure 10: Detailed performance of annotation (AP) and retrieval (mAP@20) over the 81 concepts of NUS-WIDE

lation strategy such as adding new fine-tune objectives [1] and deepening the shared hidden layers.

7. ACKNOWLEDGMENTS

This work was supported by NUS-Tsinghua Extreme Search (NExT) project under the grant No.: R-252-300-001-490 and the National Natural Science Foundation of China, No. 61303075.

8. REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, 2013.

[2] Y. Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2009.

[3] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[4] A. Bergamo, L. Torresani, and A. W. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *NIPS*, 2011.

[5] P. Blunsom, E. Grefenstette, N. Kalchbrenner, et al. A convolutional neural network for modelling sentences. In *ACL*, 2014.

[6] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013.

[7] T.-S. Chua, H. Luan, M. Sun, and S. Yang. Next: Nus-tsinghua center for extreme search of user-generated content. *MultiMedia*, 2012.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, et al. Recent advances in deep learning for speech research at microsoft. In *ICASSP*, 2013.

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[11] J. Dong, B. Cheng, X. Chen, T.-S. Chua, S. Yan, and X. Zhou. Robust image annotation via simultaneous feature and sample outlier pursuit. *TOMCCAP*, 2013.

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

[13] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 2007.

[14] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006.

[15] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[16] J.-W. Jeong, X.-J. Wang, and D.-H. Lee. Towards measuring the visualness of a concept. In *CIKM*, 2012.

[17] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa. Online incremental attribute-based zero-shot learning. In *CVPR*, 2012.

[18] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *NIPS*, 2011.

[19] L.-J. Li and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *IJCV*, 2010.

[20] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[22] Y.-I. Moon, B. Rajagopalan, and U. Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 1995.

[23] G. L. Murphy. *The big book of concepts*. The MIT Press, 2004.

[24] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *MultiMedia*, 2006.

[25] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[26] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quéenot. Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2012.

[27] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.

[28] P. Perona. Vision of a visipedia. *Proceedings of the IEEE*, 2010.

[29] N. Qian. On the momentum term in gradient descent learning algorithms. *NN*, 1999.

[30] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*, 2012.

[31] B. Scholkopf and A. Smola. Learning with kernels, 2002.

[32] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012.

[33] B. Thomee and A. Popescu. Overview of the imageclef 2012 flickr photo annotation and retrieval task. In *CLEF*, 2012.

[34] T. Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *ICML*, 2008.

[35] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[36] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*. 2010.

[37] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009.

[38] M. Wang, R. Hong, X.-T. Yuan, S. Yan, and T.-S. Chua. Movie2comics: Towards a lively video content presentation. *TMM*, 2012.

[39] L. Xie and X. He. Picture tags and world knowledge: Learning tag relations from visual semantic sources. In *MM*, 2013.

[40] H. Zhang, Z.-J. Zha, S. Yan, J. Bian, and T.-S. Chua. Attribute feedback. In *MM*, 2012.

[41] K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. *TNN*, 2009.