Running head: LARGE IMAGE DATABASES

Exploring human cognition using large image databases

Thomas L. Griffiths and Joshua T. Abbott

Department of Psychology

University of California, Berkeley

Anne S. Hsu

School of Electronic Engineering and Computer Science

Queen Mary, University of London

Keywords: natural images, computer vision, categorization, representativeness, word learning, big data, randomness
Word count: 6004 (including captions)
Address for correspondence:
Tom Griffiths
University of California, Berkeley
Department of Psychology
3210 Tolman Hall # 1650
Berkeley CA 94720-1650
E-mail: tom_griffiths@berkeley.edu
Phone: (510) 642 7134
Fax: (510) 642 5293

Abstract

Most cognitive psychology experiments evaluate models of human cognition using a relatively small, well-controlled set of stimuli. This approach stands in contrast to current work in neuroscience, perception, and computer vision, which have begun to focus on using large databases of natural images. We argue that natural images provide a powerful tool for characterizing the statistical environment in which people operate, for better evaluating psychological theories, and for bringing the insights of cognitive science closer to real applications. We discuss how some of the challenges of using natural images as stimuli in experiments can be addressed through increased sample sizes, using representations from computer vision, and developing new experimental methods. Finally, we illustrate these points by summarizing recent work using large image databases to explore questions about human cognition in four different domains: modeling subjective randomness, defining a quantitative measure of representativeness, identifying prior knowledge used in word learning, and determining the structure of natural categories.

Exploring human cognition using large image databases

Over the last century, cognitive psychology has moved towards using more and more abstract stimuli in order to maximize experimental control. An example is the literature on category learning. Hull (1920), in his classic work exploring how people learn novel concepts, used Chinese characters as stimuli – a relatively naturalistic choice. In the 1950s, when the rigor of cognitive psychology was in question, Bruner, Goodnow, and Austin (1956) began studying concept learning using stimuli that had a very clear set of discrete features – the number, fill pattern, shape, and borders of a set of objects on a card. These abstract stimuli have become standard in category learning experiments (e.g., Shepard, Hovland, & Jenkins, 1961; Medin & Schaffer, 1978), with contemporary work using methods like multidimensional scaling to confirm the dimensions that people use to represent these stimuli (e.g., Nosofsky, 1987).

Abstract stimuli support precision. Research on category learning, for example, has reached the point where it is possible to test the fine-grained predictions of a variety of detailed mathematical models of human behavior (for an overview, see Pothos & Wills, 2011). However, this precision comes at the potential cost of ecological validity: by using ever more abstract stimuli to improve the precision of our measurements, there's the chance that the cognitive processes that we are measuring no longer correspond to the phenomena that we were originally interested in understanding. Do the same processes support learning abstract categories of geometric figures and the development of a child's ability to discriminate dogs from cats?

One approach to improving the ecological validity of cognitive psychology has been to use stimuli that allow people to make use of prior knowledge, making it possible to study the effects of this knowledge on category learning (Murphy & Medin, 1985). In this paper we highlight another axis along which methodological practices might change: the use of natural images, of the kind that can be found in large online image databases. This approach follows an emerging trend towards the use of natural images in research on computer vision (e.g., Deng et al., 2009; Torralba,

Fergus, & Freeman, 2008), neuroscience (e.g., Simoncelli & Olshausen, 2001; Naselaris, Prenger, Kay, Oliver, & Gallant, 2009), perception (e.g., Geisler, Perry, Super, & Gallogly, 2001; Geisler, 2008), and visual cognition (e.g., Brady, Konkle, Alvarez, & Oliva, 2008). Large image databases have facilitated significant advances in these fields, bringing theories and empirical results into closer alignment with the problems people face in everyday life.

Natural images can be used in cognitive science research in several different ways. In this paper, we explore these different uses of natural images, consider how some of the challenges of working with large image databases might be addressed, and use a series of case studies based on our own work to illustrate how these issues are negotiated in practice.

When are natural images valuable?

Natural images have been used in two ways in neuroscience and perception research: as a source of information about human environments, and as stimuli in experiments. These uses are sufficiently different that researchers can advocate one while arguing against the other (e.g., Rust & Movshon, 2005). We think that both uses are potentially important for the study of human cognition. However, in arguing for more widespread use of natural images in cognitive science research, we are not arguing against the utility of simple abstract stimuli. These are complementary methods, useful in exploring different kinds of questions about human cognition.

In the remainder of this section we highlight three contexts where natural images are particularly valuable: estimating distributions that characterize the environment in which human beings operate, evaluating psychological theories, and taking those theories outside the laboratory and turning them into real applications. These three contexts differ in the way in which natural images are used: the first treats images as data, while the second and third use images either as data or stimuli. In discussing these contexts, we thus also consider how the use of natural images interacts with more traditional experimental methods.

Estimating distributions

A natural question to ask about any aspect of cognition or perception is how much of people's behavior can be explained by the statistics of their environment. But estimating those statistics can be a challenge. Research on visual perception has made extensive use of images of natural scenes as a source of information about the probability distributions that arise in our natural environment. For example, Geisler et al. (2001) measured the frequencies with which edges co-occurred in images of natural scenes, and showed that the resulting distribution could be use to explain people's perception of contours.

Images are an obvious source of distributional information relevant to vision, but they can also potentially give us clues about other, more cognitive capacities. To return to our running example, consider what information might be extracted relevant to categorization. Current work in computer vision aims to develop automated systems for labeling the contents of images, and large databases of images annotated by humans already exist (e.g., Russell, Torralba, Murphy, & Freeman, 2008). Annotated images carry information about the baserates with which people encounter different categories and the correlations between categories (a source of contextual cues to category membership). When used in combination with tools for extracting high-level visual features (e.g., Donahue et al., 2013), these images also provide a source of hypotheses about the kinds of features people might find highly diagnostic of category membership. Exactly this approach has been used to explore human categorization of scenes (Greene, 2013), and can potentially be pursued for other aspects of categorization.

Evaluating theories

Using natural images to estimate distributions doesn't require a commitment to using natural images as stimuli in experiments. In visual neuroscience, this approach has been productively combined with traditional experimental methods using abstract stimuli (for a discussion, see Rust & Movshon, 2005). In the same way, distributions estimated from image databases concerning the

features of objects and how they relate to category membership can be valuable in calibrating models of human category learning, even if the experiments on which those models are tested use traditional abstract stimuli.

In neuroscience, images of natural scenes have come to play a key role in evaluating theories of neural representation. One example is the idea of "sparse coding" – the notion that the brain seeks to find a representation of its environment that results in relatively few neurons firing in response to any one stimulus. Olshausen and Field (1996) showed that this assumption, when paired with input corresponding to images of natural scenes, resulted in artificial neurons acquiring receptive fields very similar to those seen in primary visual cortex. Subsequently, this approach has been used to explain neural representations for several different aspects of perception, explaining results collected using traditional experimental methods (Simoncelli & Olshausen, 2001).

Similar potential exists for evaluating theories in cognitive psychology. In the context of categorization, annotated image databases might be used to examine whether the way that people extend labels to images conforms to standard models of category learning such as prototype (e.g., Reed, 1972) and exemplar (e.g., Medin & Schaffer, 1978; Nosofsky, 1986) models. By simply examining the way in which people have applied labels to images in existing datasets, we might be able to discover how natural categories are structured.

Taking the further step of using natural images as experimental stimuli offers a different way to probe the structure of human mental representations. Rather than having to rely on, say, categories learned in the laboratory, we can explore people's beliefs about the structure of categories that they have learned naturalistically – a learning process that unfolds over years rather than hours. The scale of this learning process – both in terms of the amount of time and the amount of data that it is based on – might support more complex category structures than those that can be studied in the laboratory. As a consequence, experiments that use natural images can provide a different kind of test of the assumptions behind theories of human cognition.

Developing applications

Using real images makes it easier to develop real applications for theories developed in cognitive science. Again, an example from the study of visual perception is instructive. A current growth area in neuroscience is the development of techniques for "brain reading" – identifying what people are seeing or thinking based on measurements of neural signals. Early work in this area showed that people's neural activity when viewing simple geometric figures could be used to reconstruct those figures (Thirion et al., 2006; Miyawaki et al., 2008). But the first system that might support real applications – such as recording the contents of mental imagery or dreams – was based on natural images (Naselaris et al., 2009). This system built a model of the responses of voxels in a functional magnetic resonance imaging (fMRI) scan to information contained in real images, and then used Bayesian inference to work back from the activity of a set of voxels to a likely image.

Conducting experiments with more realistic stimuli – and evaluating models on more realistic tasks – likewise creates the opportunity to put the findings of cognitive psychology in more direct contact with applications. Applied problems in computer vision such as image labeling and image search naturally involve elements that parallel cognitive tasks – categorization and generalization, respectively. Reducing the gap between the approaches taken by computer vision and cognitive psychology provides a natural way to increase the flow of ideas between these disciplines, allowing insights from cognitive science to be used more widely in machine learning and robotics.

Overcoming the challenges of working with natural images

Working with natural images poses challenges, particularly when those images are used as experimental stimuli. Even as natural images have become widely used as tools for estimating distributions and evaluating theories in neuroscience and vision research, the use of natural images as experimental stimuli has remained controversial (e.g., Rust & Movshon, 2005). The properties

of artificial stimuli are well understood, making it possible to conduct experiments with precision and make clear assumptions about how people are representing the stimuli. Working with natural images requires developing strategies to address the lower precision and greater uncertainty about representations that results.

Technological developments – in particular, the World Wide Web – make it far easier to get access to large image databases that can be used in research. This increased availability accounts in part for the increased use of natural images in other fields such as computer science. But the other part is that these fields have developed new methods for analyzing and representing natural images that make it possible to work with them in a rigorous fashion. Similar innovations are required for cognitive psychology to make the most effective use of large image databases. Some of the relevant technological advances are already in place, but others – such as defining new experimental methods that make the most of these complex stimuli – are a source of important open research questions.

Increasing precision with large samples

Careful control of experimental stimuli is important for reducing one source of noise from the already noisy signal provided by human behavior. Natural images are going to vary along many dimensions other than those the experimenter aims to manipulate – there might be multiple objects in a scene, factors that provide undesired cues to context, and so on. This introduces additional variability into experiments. But the alternative – using artificial stimuli – potentially introduces bias, if the goal is to develop theories that are applicable to the real world. Arguably, variance is a lesser evil than bias, as it can be reduced by increasing sample sizes.

In particular, the technological innovations that have made it possible to easily gain access to large image databases also offer a tool for addressing this problem: increasing the number of participants who take part in experiments, via crowdsourcing websites such as Amazon's Mechanical Turk (http://mturk.com; for further details of this population see Mason & Suri, 2012).

Crowdsourcing massively increases the bandwidth of psychological experimentation, providing the opportunity to collect quantities of behavioral data that rival the data about the brain produced by fMRI (and at similar cost – an hour of MRI time will pay for several hundred participants in an experiment on Mechanical Turk). Rather than using this increased bandwidth to run the same kind of experiments that would be run in the laboratory, we can make use of the greater precision it offers to give up some precision in control of stimuli and run experiments with greater ecological validity.

Finding appropriate representations for images

Using simple stimuli makes it easier to make uncontroversial claims about how those stimuli are represented. Geometric shapes with clear binary features (e.g., Bruner et al., 1956; Medin & Schaffer, 1978; Shepard et al., 1961) can be assumed to be represented using those features. Stimuli that vary along easily identified dimensions can likewise be put into correspondence with multidimensional scaling solutions to confirm that people represent those dimensions (e.g., Nosofsky, 1986). But natural images – with many complex features, and no easily identified underlying dimensions – pose quite a different challenge. One strategy is to use information that accompanies the images, rather than the images themselves, as a source of representational assumptions. For example, ImageNet (Deng et al., 2009) provides 14 million images that are identified with the nodes of the directed graph comprising WordNet (Miller & Fellbaum, 1998). Consequently, WordNet provides a source of representational assumptions for those images, and a way of measuring the similarity between them without analyzing the images themselves. Other image databases, such as the Corel database, are labelled with tags that carry semantic information.

A second strategy is to try to identify representations directly from the images. Research in computer vision has resulted in a variety of schemes for identifying the features of images, such as SIFT features (Lowe, 1999), the GIST descriptor (Oliva & Torralba, 2001), and spatial-temporal "words" extracted from video (Niebles, Wang, & Fei-Fei, 2008). These features have already been

used in psychological models (e.g., Torralba, Oliva, Castelhano, & Henderson, 2006; Greene & Oliva, 2009; Buchsbaum, Canini, & Griffiths, 2011). More recently, computer vision researchers have started to use features generated by "deep" neural networks, which have resulted in significantly higher performance on a range of computer vision tasks (Krizhevsky, Sutskever, & Hinton, 2012; Donahue et al., 2013) and are motivated in part by parallels with the hierarchical processing of images in human visual cortex (e.g., Riesenhuber & Poggio, 1999).

Image features developed by computer scientists are useful for developing computational models of human inferences from images, but may be a poor proxy for psychological representations. For example, recent work has highlighted ways in which the features discovered by deep networks differ from those identified by the human visual system (Nguyen, Yosinski, & Clune, 2014). However, these features provide a starting point for designing experiments that might gather more precise representational information. For example, graph-based methods for identifying representations such as Isomap (Tenenbaum, De Silva, & Langford, 2000) only require a measure of similarity that is accurate for the most similar stimuli – something that the kinds of features used in computer vision might provide.

Defining new methods

One of the main arguments that Rust and Movshon (2005) made against the use of natural images as stimuli is that existing methods for identifying people's representations from those stimuli – such as reverse correlation (Ahumada & Lovell, 1971) – work poorly with natural images. Rather than an insurmountable obstacle, we view this as a research challenge: can we develop new behavioral methods that can make the best use of these stimuli?

In addressing this challenge, it may also be instructive to look to computer science. Statisticians and computer scientists have developed a variety of sophisticated methods for estimating complex unknown quantities from high-dimensional data, and these methods can be adapted to behavioral research. We present an example of one such method – Markov chain Monte

Carlo with people – below.

Examples of using natural images to explore human cognition

Having discussed the general value and use of natural images for research in cognitive science, we now turn to specifics. In this section we present four examples from our own research of ways in which natural images can be used to address questions about human cognition. These four examples illustrate the uses for large image databases discussed above, and show how some of the challenges posed by using real images in cognitive psychology research might be overcome.

Estimating distributions: Modeling subjective randomness

People have strong intuitions about whether a sequence of heads and tails or a pattern of dots seems random (Falk & Konold, 1997). A natural question is where these intuitions come from. A Bayesian analysis of the problem of detecting randomness suggests that subjective randomness might be viewed as reflecting the relative likelihood that a stimuli was produced from a random generating process, rather than from one with a more regular structure (Feldman, 1997; Griffiths & Tenenbaum, 2003). For example, Griffiths and Tenenbaum (2003) defined the following measure of randomness:

$$\text{random}(X) = \frac{p(X|\text{random})}{p(X|\text{regular})} \tag{1}$$

where $p(X|\text{random})$ and $p(X|\text{regular})$ are the probability of a stimulus X being generated by a random and regular process respectively. But this raises an important problem: how do we know what constitutes a regular process, or what distribution over stimuli it implies?

Hsu, Griffiths, and Schreiber (2010) addressed this problem using the strategy introduced above: looking to natural images as a source of information about a probability distribution. They hypothesized that the perceived randomness of a binary array (in this case, a 4×4 grid of black and white squares) could be determined in part by estimating the probability of that array occurring

in an image of a natural scene. For these stimuli, $p(X|\text{random})$ was readily calculated assuming that each cell in the array takes on a value of 1 or 0 with equal probability. $p(X|\text{regular})$ was estimated by an exhaustive tabulation of the frequency with which all 4×4 patterns appeared in a set of natural images.

The images used to estimate $p(X|\text{regular})$ consisted of 62 photographs of natural scenes containing trees, flowers, and shrubs (Doi, Inui, Lee, Wachtler, & Sejnowski, 2003). There were no images of humans, animals, or cityscapes. Image patches of varying sizes were extracted from each natural image to measure statistics at a range of scales. A total of 700,000 patches were sampled at random from among the 62 images using $n \times n$ patches, for $n = 4$, 8, 16, 32, 64, 128, and 256 pixels. All the patches were then reduced through averaging and intensity thresholding down to 4×4 binary arrays. The resulting 4,900,000 binary arrays were then divided into the 216 possible patterns, and the frequency of each pattern was recorded. Normalizing these frequencies provided an estimate for the probability distribution $p(X|\text{regular})$ in the equation for $\text{random}(X)$.

To evaluate this approach, a set of stimuli were selected from 50 evenly spaced quantiles on either side of the neutral stimulus with $\text{random}(X) = 0$, for a total of 100 images. The full set of stimuli, ordered by $\text{random}(X)$, is shown in Figure 1. A group of 77 participants were asked to label each test image as either random or not-random. A significant linear correlation was found between $\text{random}(X)$ and the probability that the stimulus would be classified as random ($r(98) = .75$, $p < .001$), and the rank-order correlation (taking into account only the relative ordering of these different measures) was $r(98) = .75$, $p < .01$. This strong correlation suggests that natural images provide a good source for estimates of $p(X|\text{regular})$, and that the environment may influence the regularities people naturally identify when they are assessing randomness.

Extending this work to more complex stimuli requires estimating the probabilities of more complex perceptual objects. Small binary arrays were used in order to make it possible to simply enumerate the full distribution, but working with larger arrays – or stimuli that ultimately look more like images – will require specifying a probabilistic model that describes the joint

probabilities of pixel values. Defining such a model will require making some choices about the kinds of regularities expressed in the distribution. For example, while seeing a plaid elephant is extremely unlikely, this complex perceptual stimulus is composed of many simple parts, all of which might arise fairly commonly in natural images. Identifying what the components are that people use to represent the content of images is a major theoretical challenge, but perhaps one that we can begin to explore through a more systematic exploration of human subjective randomness.

Evaluating theories: Measuring the representativeness of images

The images in Figure 2 (a) have all been labeled as having some aspect of "coast" in them. Clearly, some of these images are bad examples of coasts, but a few are good examples. How do people determine what makes something a good example of a concept? A common proposal in cognitive psychology is that people use representativeness, a similarity-based heuristic, to make these decisions (e.g., Kahneman & Tversky, 1972). However, exactly what our intuitive sense of representativeness corresponds to remains elusive.

The notion of "representativeness" appeared in cognitive psychology as a proposal for a heuristic that people might use in the place of performing a probabilistic computation (Gigerenzer, 1996; Kahneman & Tversky, 1972). For example, we might explain why people believe that the sequence of heads and tails HHTHT is more likely than HHHHH to be produced by a fair coin by saying that the former is more representative of the output of a fair coin than the latter. This proposal seems intuitive, but raises a new problem: How is representativeness itself defined? Various proposals have been made, connecting representativeness to existing quantities such as similarity (Kahneman & Tversky, 1972), or likelihood (Gigerenzer, 1996).

Tenenbaum and Griffiths (2001) took a different approach to this question, providing a rational analysis of representativeness by trying to identify the problem that such a quantity solves. They proposed that one sense of representativeness is being a good example of a concept, and then showed how this could be quantified via Bayesian inference. Given some observed data d and a set

of of hypothetical sources, $\mathcal{H}$, Tenenbaum and Griffiths (2001) defined the representativeness of d for h to be the evidence that d provides in favor of a specific h relative to its alternatives,

$$R(d,h) = \log \frac{p(d|h)}{\sum_{h' \neq h} p(d|h')p(h')} \tag{2}$$

where $p(h')$ in the denominator is the prior distribution on hypotheses, re-normalized over $h' \neq h$. Essentially, being a good example means providing strong evidence for the target concept relative to possible alternatives. The resulting model outperformed alternative accounts of representativeness based just on similarity in predicting human representativeness judgments for two kinds of simple stimuli. However, this formal model was not evaluated beyond these specific, hand-constructed domains.

The Bayesian measure of representativeness introduced by Tenenbaum and Griffiths (2001) indicated the representativeness of data d for a hypothesis h. However, in many cases we might not know what statistical hypothesis best describes the concept that we want to illustrate through an example. For instance, in an image retrieval problem, we might just have a set of images that are all assigned to the same category, without a clear idea of the distribution that characterizes that category. To address these concerns, Abbott, Heller, Ghahramani, and Griffiths (2011) extended this Bayesian measure of representativeness to apply to sets of objects and showed that the resulting model was closely mathematically related to an existing machine learning method of clustering-on-demand known as Bayesian Sets (Ghahramani & Heller, 2005). More formally, given a data collection $\mathcal{D}$, and a subset of items $\mathcal{D}_s = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{D}$ representing a concept, the Bayesian Sets algorithm ranks an item $\mathbf{x}^* \in \{\mathcal{D} \setminus \mathcal{D}_s\}$ by the following scoring criterion

$$\text{Bscore}(\mathbf{x}^*) = \frac{p(\mathbf{x}^*, \mathcal{D}_s)}{p(\mathbf{x}^*)p(\mathcal{D}_s)}. \tag{3}$$

This ratio intuitively compares the probability that $\mathbf{x}^*$ and $\mathcal{D}_s$ were generated by some statistical model with the same, though unknown, model parameters θ, versus the probability that $\mathbf{x}^*$ and $\mathcal{D}_s$

were generated by some statistical model with different model parameters θ_1 and θ_2. Extending the rational model of representativeness from Tenenbaum and Griffiths (2001) to connect with the Bayesian Sets algorithm of Ghahramani and Heller (2005) provides a link between the cognitive science literature on representativeness and the machine learning literature on information retrieval, allowing for the evaluation of psychological theories on large-scale datasets.

Abbott et al. (2011) exploited this link to provide a detailed evaluation of a set of representativeness models using a large database of natural images. Here, the problem is formulated as one of determining how representative an image is of a labeled set of images. The dataset that was used was first presented by Heller and Ghahramani (2006), being a subset of images taken from the Corel database commonly used in content-based image retrieval systems. The images in the dataset are partitioned into 50 labeled sets depicting unique categories, with varying numbers of images in each set (the mean is 264). The dataset is of particular interest for testing models of representativeness as each image from the Corel database comes with multiple labels given by human judges. The labels have been criticized for not always being of high quality (Müller, Marchand-Maillet, & Pun, 2002), which provides an additional (realistic) challenge for the models of representativeness that were evaluated. The images in this dataset are represented as 240-dimensional feature vectors, composed of 48 Gabor texture features, 27 Tamura texture features, and 165 color histogram features. The images were additionally preprocessed through a binarization stage based on the skewness of feature distributions, transforming the entire dataset into a sparse binary matrix that represents the features which most distinguish each image from the rest of the dataset. Details of the construction of this feature representation are presented in Heller and Ghahramani (2006).

Abbott et al. (2011) compared their Bayesian model against a likelihood model and two similarity models, building upon a simple leave-one-out framework to allow a fair comparison of the different representativeness models. Given a set of images with a particular category label from the set of 50 above, the leave-one-out algorithm iterates through each image in the set and compute

a score for how well this image represents the rest of the set. Abbott et al. (2011) then took the top 10 and bottom 10 ranked images from the output of each model and used each as stimuli in a large experiment run on Amazon Mechanical Turk. Each of 500 participants was shown a series of images and was asked to rate how good an example each image was of the assigned category. Figure 2 presents an example of this ranked output for the category "coast". Overall, the Bayesian model of representativeness provided the best account of peoples judgments of which images were good and bad examples of the different categories.

Using a large database of natural images allowed Abbott et al. (2011) to extend and evaluate a model of representativeness under a rational analysis, with results that provide strong evidence for this characterization of representativeness. In addition, utilizing a standard image database from the computer vision community opens up the opportunity to test theories of representativeness in more applied settings and over other image databases and ontologies (Sun, Wang, Yao, & Zhang, 2013).

Developing applications: Large-scale word learning

How do people learn to appropriately apply new labels to concepts from only a few example observations? Xu and Tenenbaum (2007) examined how people generalize a novel word to new objects based on the diversity and number of objects shown as examples of the word. For example, a participant saw the word "FEP" applied to three images of Dalmatians and was asked to select what other objects they would label "FEP" from a set of other Dalmatians, other dogs, other animals, vehicles, and vegetables (see Figure 3 for a set of example trials). Xu and Tenenbaum (2007) found that participants would use "FEP" to label only the other Dalmatians in this case, but if they saw "FEP" applied to three different types of dogs, participants would extend the label to represent all dogs, not just Dalmatians, and not to other (non-dog) animals as well. To account for these results, Xu and Tenenbaum (2007) developed a Bayesian word learning model and found a high correspondence in the degree of generalization shown by the model and by people.

While Xu and Tenenbaum (2007) showed that their Bayesian word learning model did a good job of capturing human judgments, both the model and the experiment used to evaluate it were based on a very small set of stimuli – a total of 45 objects divided into animals, vegetables, and vehicles. Making and evaluating a model of word learning that can be used in real applications – such as teaching new words to a robot designed to interact with humans – requires scaling this up significantly. The Bayesian word learning model assumes that hypotheses about the meanings of words correspond to a taxonomic hierarchy, being subsets of one another. For example, the same object might be represented at three different levels of abstraction: subordinate (e.g., a Dalmatian), basic (e.g., a dog), and super-ordinate (e.g., an animal). To be able to use the model, a taxonomy was constructed from similarity ratings that each person provided for each pair of objects in the stimulus set. With just 45 images, this corresponded to roughly 400 judgments per participant. This approach cannot scale, as it requires on the order of n^2 judgments for n objects.

Abbott, Austerweil, and Griffiths (2012) presented an alternative approach to hypothesis space construction, making it possible to use the Bayesian word learning framework with a natural set of concepts on a large scale without eliciting any judgments from participants. They developed a hypothesis space automatically derived from the structure of a large online word ontology, WordNet (Miller & Fellbaum, 1998). WordNet is a lexical database of English represented as a network of words linked by directed edges denoting semantic relatedness. As WordNet is hierarchically structured like the hypothesis space used by Xu and Tenenbaum (2007), it is an ideal candidate for constructing the hypothesis space. Furthermore, for each node in WordNet, there are at least 500 high-quality images in the ImageNet database (Deng et al., 2009). These naturalistic images can be used to generate better features as input to models, and as the source of stimuli for large-scale behavioral experiments.

From the 82,115 noun-node subtree of Wordnet, Abbott et al. (2012) created a hypothesis space that is a binary matrix, $\mathcal{H}$, whose rows are the objects (64,958 leaf nodes from the subtree) and columns are the hypotheses (82,115 nodes, 17,157 of which are inner nodes and 64,958 are

leaf nodes). Each entry (i, j) of the matrix $\mathcal{H}$ denotes whether or not hypothesis node j is an ancestor of leaf node i in the WordNet graph (with a 1 indicating it is). The leaf nodes are included as hypotheses so that the model distinguishes between subordinate objects. While this space has no clear taxonomic labelling (e.g., there is no well-defined "basic-level"), this allows the testing of generalization for categories at different levels of abstraction. The prior probability of different hypotheses was defined to be Erlang distributed in the size of the hypothesis (the number of leaf nodes under the hypothesis node h), a standard prior over sizes in Bayesian models (Shepard, 1987; Tenenbaum, 2000).

Using this hypothesis space and the experimental paradigm from Xu and Tenenbaum (2007), Abbott et al. (2012) ran a large online experiment via Amazon Mechanical Turk. Stimuli were images sampled from ImageNet for the three object taxonomies of animals, vehicles, and vegetables used in Xu and Tenenbaum (2007). Figure 3 displays example results for the domain of animals. Overall, the Mechanical Turk participants displayed the characteristic patterns of generalization similar to adults in Xu and Tenenbaum (2007). The Bayesian word learning model with a hypothesis space derived from WordNet also captured these trends, providing validation of the original model in a naturalistic setting.

With ImageNet as a source of natural stimuli and WordNet as the source of their representations, it was possible to extend the Bayesian word learning framework from 45 objects to over 14 million images. One of the key advantages of adopting this framework is that it is relatively simple to conduct new experiments with different sets of stimuli and hypothesis spaces using exactly the same word learning model. Abbott et al. (2012) demonstrated this in an additional experiment with the same design as above, but with three new domains, and obtained qualitatively similar gradients of generalization between model and people. This model has most recently been extended to incorporate perceptual uncertainty, gaining leverage from a low-level visual classifier itself trained using ImageNet (Jia, Abbott, Austerweil, Griffiths, & Darrell, 2013). Here, rather than using Imagenet as just a source for experimental stimuli, Jia et al. (2013) used

ImageNet to train a convolutional neural network and provide a more robust featural representation for the Bayesian word learning model. The resulting hybrid model outperforms state-of-the-art computer vision systems for the problem of appropriately generalizing from a set of input images, and approaches human performance on this task. Given the scale of ImageNet and WordNet, these results bring us closer to using the Bayesian word learning framework in real-world applications.

Defining new methods: Exploring natural categories

The growing number of large image databases presents new opportunities for psychological research. However, it is challenging to collect relevant human judgments using these databases. Because of the sheer number of images they contain, only a small proportion are likely to be relevant to a particular research question. Thus, new experimental methods are needed in order to make the most of the opportunities they offer.

One example of such a method focuses on the question of how to measure people's beliefs about the structure of categories using large sets of discrete stimuli, such as images. This method is called discrete Markov chain Monte Carlo with people (d-MCMCP) (A. S. Hsu, Martin, Sanborn, & Griffiths, 2012), and is based on ideas from the well-known class of Markov chain Monte Carlo (MCMC) algorithms from computer science. Previous work has shown how Markov chain Monte Carlo algorithms can be adapted to be used as the basis for psychological experiments, provided the stimuli can be represented in terms of underlying features or dimensions (Sanborn, Griffiths, & Shiffrin, 2010). d-MCMCP extends this approach to work with discrete, realistic stimuli such as the contents of large online databases.

A brief overview of the d-MCMCP method is as follows: First, it is assumed that people's representation of a category C can be expressed as a probability distribution over a set of items X, $p(X|C)$. Given this assumption, d-MCMCP works by combining the basic methods of a standard MCMC algorithm with human judgments. The human judgments are obtained using a two alternative forced choice task where people are asked to make a choice between a pair of items

e.g., “Which image looks more like a dog?”. The d-MCMCP procedure presents pairs of items in a manner such that the sequence of chosen items forms a “chain”, with each pair consisting of the previously chosen item and a closely-related variant. With some general assumptions about people’s pair-wise choice behavior (see Sanborn et al., 2010), this chain can be treated as providing samples from $p(X|C)$, providing insight into the structure of the corresponding category.

Hsu et al. (2012) used d-MCMCP to explore peoples representations of seasonal images defined over images from a large online database. Specifically, they explored the categories associated with the seasons Spring, Summer, Autumn, and Winter, using images obtained from online image databases. A set of 4000 colored season-related images was assembled by searching for public domain web images using the phrases “spring season”, “summer season”, “autumn season”, and “winter season” in Google Image Search and on Flickr.com. The top 500 results for searches on Google and Flickr for each season were downloaded. All images were resized so that the maximum dimension was 250 pixels, while preserving the original ratio of image height to width. Each participant made pairwise choices between images by answering questions such as “Which image is more representative of Spring?”. The study was run through Amazon Mechanical Turk, making it possible to recruit a large number of participants.

The top ten images that were chosen most often over all three chains for each season are shown in Figure 4. Clearly, the images are very indicative of each season. Convergence within each category was quantified by calculating the distance between 11-bin color histograms for cumulative images, both between chains for the same season and between chains corresponding to different seasons. Within-chain distance decreased over numbers of trials, and was typically lower than the similarity between chains, supporting the idea that chains are converging towards different parts of the space of images. The right hand column of Figure 4 shows a simple example of the kind of statistical analyses that can be done on the resulting samples. The color histograms for the different seasons are quite different from one another, and correspond to palettes that intuitively match the seasons. These results illustrate that the d-MCMCP method can be used to extract

psychologically meaningful information about the structure of categories from large image databases.

Conclusions

Large image databases provide unique opportunities for cognitive psychology, making it possible to explore the statistical structure of people's environment, test theories in a more realistic way, and work towards more direct applications of those theories. The challenges posed by using natural images as stimuli are significant – a decrease in precision, and difficulty in identifying appropriate representations. However, those challenges can be overcome by developing new experimental methods that make use of the additional resources provided by running experiments online, and exploring new ways to combine ideas from cognitive psychology and computer science.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2012). Constructing a hypothesis space from the Web for large-scale Bayesian word learning. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Abbott, J. T., Heller, K. A., Ghahramani, Z., & Griffiths, T. L. (2011). Testing a Bayesian measure of representativeness using a large image database. In *Advances in Neural Information Processing Systems 24.*

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *49*, 1751-1756.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325-14329.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Buchsbaum, D., Canini, K. R., & Griffiths, T. L. (2011). Segmenting and recognizing human action using low-level video features. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Doi, E., Inui, T., Lee, T., Wachtler, T., & Sejnowski, T. J. (2003). Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Computation*, *15*, 397-417.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2013). DeCAF: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a bias for judgment. *Psychological Review*, *104*, 301-318.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, *41*, 145-170.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.*, *59*, 167–192.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*, 711-724.

Ghahramani, Z., & Heller, K. A. (2005). Bayesian sets. In *Advances in Neural Information Processing Systems 18.*

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological review*, *103*(3), 592.

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, *4*.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137–176.

Griffiths, T. L., & Tenenbaum, J. B. (2003). Probability, algorithmic complexity, and subjective randomness. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*

Heller, K. A., & Ghahramani, Z. (2006). A simple Bayesian framework for content-based image retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, *2*, 2110–2117.

Hsu, A., Griffiths, T., & Schreiber, E. (2010). Subjective randomness and natural scene statistics. *Psychonomic Bulletin & Review*, *17*, 624-629.

Hsu, A. S., Martin, J. B., Sanborn, A. N., & Griffiths, T. L. (2012). Identifying representations of categories of discrete items using Markov chain Monte Carlo with people. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Hull, C. (1920). Quantitative aspects of the evolution of concepts. *Psychological Mongraphs*,

XXVIII(1).

Jia, Y., Abbott, J. T., Austerweil, J. L., Griffiths, T. L., & Darrell, T. (2013). Visual concept learning: combining machine vision and Bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems 26.*

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430-454.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision* (Vol. 2, pp. 1150–1157).

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, *44*, 1-23.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Miller, G. A., & Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., … Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*(5), 915-929.

Müller, H., Marchand-Maillet, S., & Pun, T. (2002). The truth about Corel - evaluation in image retrieval. *International Conference on Image and Video Retrieval*.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902-915.

Nguyen, A., Yosinski, J., & Clune, J. (2014). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv preprint arXiv:1412.1897*.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, *79*(3), 299–318.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87-108.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607-609.

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization.* Cambridge, UK: Cambridge University Press.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393-407.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*, 157-173.

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, *8*(12), 1647–1650.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*, 63-106.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science.

Science, *237*, 1317-1323.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*. (13, Whole No. 517)

Simoncelli, E. P., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193-1216.

Sun, X., Wang, X.-J., Yao, H., & Zhang, L. (2013). Exploring implicit image statistics for visual representativeness modeling. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 516–523).

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319-2323.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (p. 1036-1041).

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, *33*(4), 1104-1116.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(11), 1958-1970.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*(4), 766.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245-272.

Author Note

This work was supported by grants FA-9550-10-1-0232 and FA9550-13-1-0170 from the Air Force Office of Scientific Research and SMA-1228541 from the National Science Foundation.

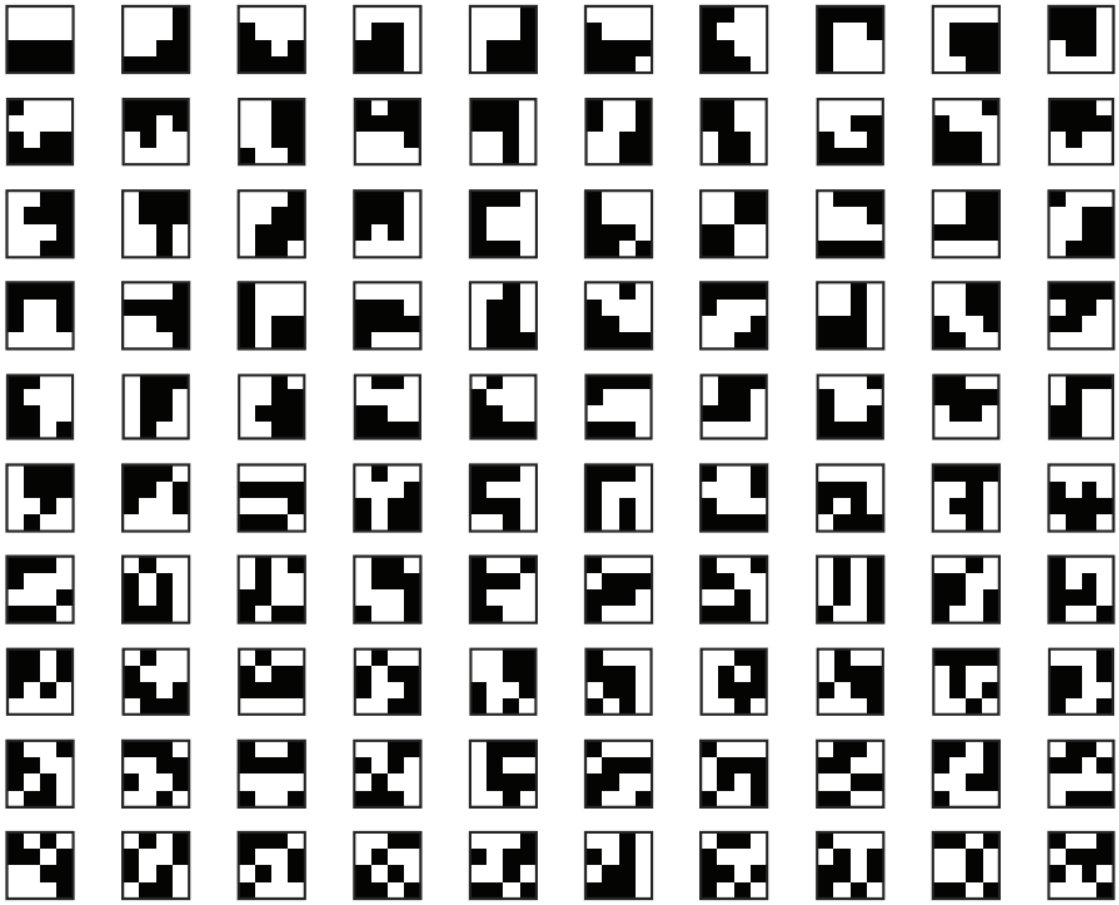
Figure Captions

Figure 1. One hundred 4×4 binary arrays, ordered by their predicted subjective randomness based on the frequency with which they appear in natural scenes.

Figure 2. Assessing the representativeness of images. (a) A diverse set of images that have been given the label "coast" by human judges. (b) the top nine and (c) bottom nine ranked images according to the Bayesian representativeness model.

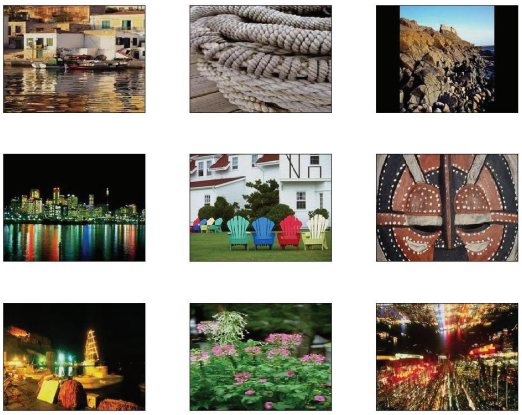
Figure 3. Bayesian word learning with natural images. The images in the left-most column are the stimuli presented as an example observations of a concept. Panel (a) displays three subordinate examples (Dalmatians), (b) three basic examples (dogs), and (c) three superordinate examples (animals). The images on the right are the test stimuli to select from, with the grey bars next to them indicating the percentage of participants selecting that image as an extension of the test word.

Figure 4. Distributions over images for different seasons estimated using discrete Markov chain Monte Carlo with people. The images on the left are the top ten images (out of 4000) for each season. The distribution over colors associated with each season, estimated by averaging over the images selected for that season, are shown on the right.

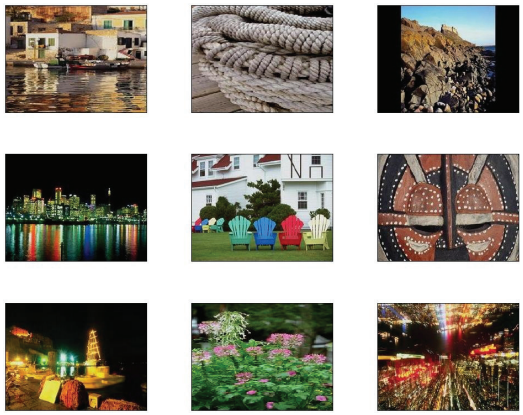


(a)

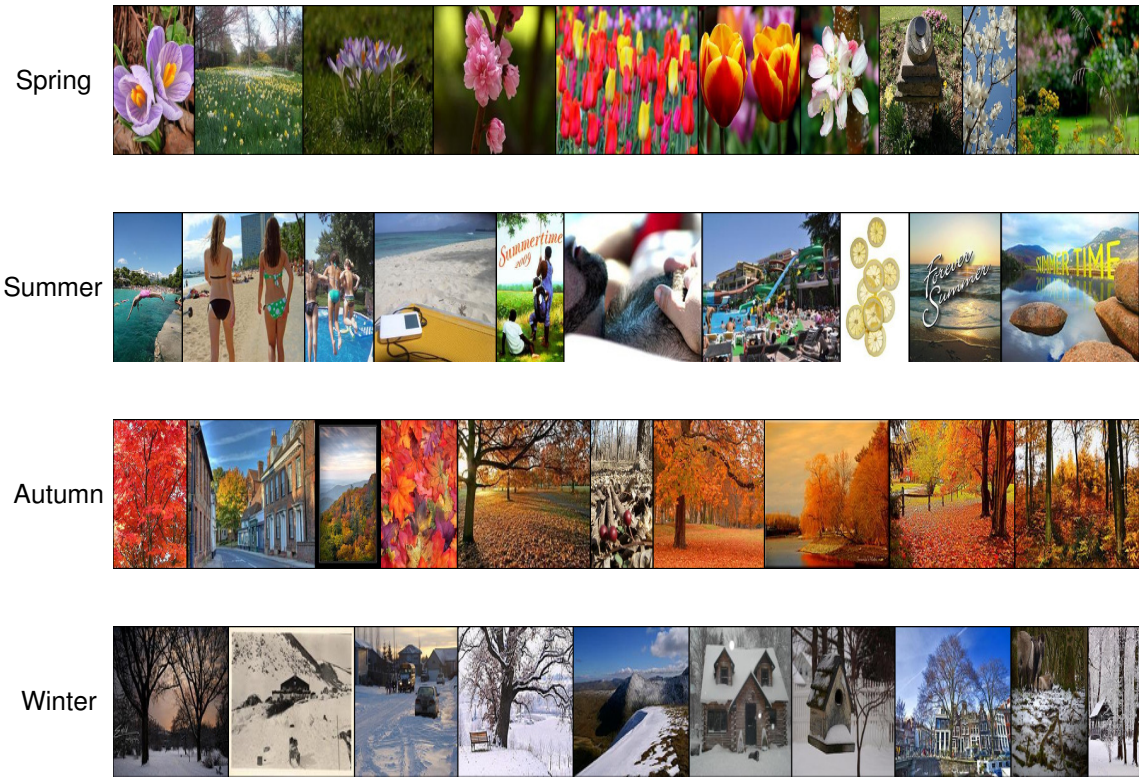
(b)



(c)




Observations
Generalizations
(a)
(b)
(c)
1
.5
0


Spring
Summer
Autumn
Winter