ON THE COMPLEXITY OF BAYESIAN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider concept generalization at a large scale in a diverse and natural visual spectrum. Established computational modes (*i.e.*, rule-based or similarity-based) are primarily studied isolated and focus on confined and abstract problem spaces. In this work, we study the two modes when the problem space scales up and the *complexity* of concepts becomes diverse. Specifically, at the **representational level**, we seek to answer how the complexity varies when a visual concept is mapped to the representation space. Prior psychology literature has shown that two types of complexities (*i.e.*, subjective complexity and visual complexity) (Griffiths and Tenenbaum, 2003) build an inverted-U relation (Donderi, 2006; Sun and Firestone, 2021). Leveraging Representativeness of Attribute (RoA), we computationally confirm the following observation: Models use attributes with high RoA to describe visual concepts, and the description length falls in an inverted-U relation with the increment in visual complexity. At the **computational level**, we aim to answer how the complexity of representation affects the shift between the rule- and similarity-based generalization. We hypothesize that category-conditioned visual modeling estimates the co-occurrence frequency between visual and categorical attributes, thus potentially serving as the prior for the natural visual world. Experimental results show that representations with relatively high subjective complexity outperform those with relatively low subjective complexity in the rule-based generalization, while the trend is the opposite in the similarity-based generalization.

1 INTRODUCTION

Concept generalization can be categorized into two distinct schools of thought, rule- and similarity-based approaches (Sloman and Rips, 1998; Shepard, 1987). The first attempt to unify these two modes was presented by Tenenbaum (1999) using Bayesian inference (Tenenbaum, 1998; Tenenbaum and Griffiths, 2001a; Xu and Tenenbaum, 2007). Based on perception (Kersten et al., 2004), this paradigm reconstructs human's hypothesis space consisting of abstract features and incubates modern concept learning algorithms (Tenenbaum et al., 2011; Lake et al., 2015; Ellis et al., 2020). However, as most concept learners have only demonstrated in confined and abstract problem space, a challenging problem remains: When the problem space *scales up* (*e.g.*, using data collected from the natural world), is there a unified concept representation that combines the two established modes (*i.e.*, rule- and similarity-based)? If it does, how does the generalization shift between the two modes w.r.t. the *complexity* of concepts?

To give some concrete examples, let us consider the visual categorization problem in the natural world (Quine, 1960). (i) We can easily generalize *banana to cucumber* to *tennis to watermelon* considering the surface color (*i.e.*, *yellow to green*), but not easy to judge that *tennis* is more similar to *cucumber* or *watermelon*. (ii) We say that *tiger* is more similar to *cat* rather than *dog*, but hard to generalize the rule of *tiger to dog* to *cat to something*. (iii) We see digits *1* looks like *7* than *8*, but cannot figure out *7 to something* given *1 to 8*. How do these distinctions emerge from visual observations?

One hypothesis is that we tend to describe those visual concepts (*i.e.*, **visual complexity**, the complexity coded by pixels) by simple visual patterns (*i.e.*, **subjective complexity**, the coding length humans used to describe certain concepts) (Donderi, 2006; Wolfram et al., 2002). For a simple concept, we may only need one attribute, *e.g.*, the color yellow. As the concepts become more complex, we ought to adopt more attributes, such as *watermelons are green balls with waved stripes*. When the concept becomes even more complicated, we would choose not to describe it with many attributes but instead back to simple ones, such as *dogs has dog-like face*.

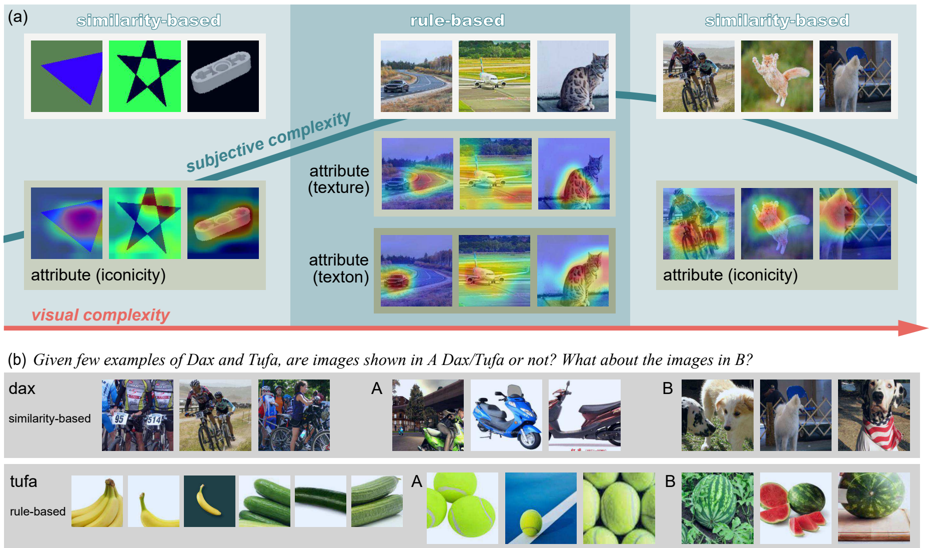


Figure 1: **The landscape of the computation-mode-shift *vs*. the concept complexity.** (a) *Representation level:* original visual concepts of diverse complexity and visualization of their representative attributes (around the peaks of heatmaps). (b) *Computation level:* an illustration of similarity- and rule-based generalization. The former is similar to word learning (Xu and Tenenbaum, 2007): Given very few examples of known concept *dax*, tell which is most likely to be *dax* in unseen examples. The latter is akin to concept learning (Salakhutdinov et al., 2012): Given a rule *tufa* over two known concepts, tell how *tufa* generates the examples of unknown concepts. As concept visual complexity increases, concept subjective complexity first increases, then decreases—and the computation mode shifts from similarity to rules as subjective complexity increases.

Together, we observe a shift in the continuous space spanned by the rule- and similarity-based approaches w.r.t. the increase of concept complexity. Intuitively, both very simple and complex concepts have a lower description length, which tends to generalize by similarity. In comparison, concepts neither too simple nor too complex have a higher description length, which tends to generalize by rules. This observation echos modern literature in both information theory and psychology, which demonstrate that subjective and visual complexity (Griffiths and Tenenbaum, 2003) come in an inverted-U relation (Donderi, 2006; Sun and Firestone, 2021).

In essence, we seek to quantify the relation between the prior-studied but mostly isolated modes (*i.e.*, rule- and similarity-based): What are the *relations* between the computation-mode-shift and the concept complexity, as we hypothesized above and illustrated in Fig. 1? Specifically, we disassemble the above question into two on the basis of Marr's (Marr, 1982) *representational level* and *computational level*, respectively: (i) How does the complexity change when a visual concept is mapped to the representation space? (ii) How does the complexity of representation affect the shift between rule- and similarity-based generalization? By answering these two questions, we hope to provide a new perspective and the very first pieces of evidence on unifying the two computational modes by mapping out the landscape of the concept complexity *vs*. the computation mode.

Representation *vs*. complexity Representing the natural visual world merely with human prior is insufficient (Griffiths et al., 2016) and oftentimes brittle to generalize. Despite that hierarchy empowers large-scale Bayesian word learning (Miller, 1998; Abbott et al., 2012), extending it to visual domains is yet challenging and may need costly elaboration. In comparison, modern discriminative models trained for visual categorization by leveraging large-scale datasets can capture the rich concept of attributes (Xie et al., 2020). These observations and progresses naturally lead to the problem of concept representation complexity: If we distinguish visual concepts using attributes, at least how many attributes should we use (Chaitin, 1977; Li et al., 2008)?

To tackle this problem, here we offer a new perspective by bridging the subjective complexity with the visual complexity via Representativeness of Attribute (RoA), which consists of (i) the probability of recalling an attribute z when referring to a concept c, and (ii) the probability of recalling other concepts $\hat{c}$ when referring to an attribute z. This design echoes the principles in rational analysis (Tenenbaum and Griffiths, 2001b) yet can be obtained by frequentist statistics for large problem spaces (*e.g.*, natural visual world (Abbott et al., 2011)).

Computation *vs*. complexity Modern statistical learning methods have demonstrated strong expressiveness in concept representation by implicitly calculating the co-occurrence frequency between visual attributes and categories (Wu et al., 2019; Xie et al., 2016), even when scaling up to the complex and large-scale visual domain—the learned representation fits the prior distribution of visual concepts conditioned on categorical description (Xie et al., 2020). It can also bridge sensory-derived and language-derived knowledge (Bi, 2021). Hence, this learning paradigm should somehow inherent semantic properties in addition to visual properties, such as iconicity (Fay et al., 2010; 2013) and disentanglement (Allen and Hospedales, 2019; Gittens et al., 2017; Mikolov et al., 2013).

Leveraging this observation, we hypothesize that *rule-based* and *similarity-based* generalization reflects the *analogy* and *relatedness* properties in psycholinguistics (Gentner and Markman, 1997), where the former pairs are two ends of a continuum of concept representation and the latter pairs are two ends of a continuum of literal meaning. Visual categorization brings these two pairs together because linguistic analogy and relatedness come from generalizing the corresponding appearance instead of pure literal meaning—concepts with more easy-to-disentangle attributes (*e.g*., shape and color) are more likely to generalize by rules, while concepts represented with more iconicity (Fay et al., 2014) (*i.e*., those more likely to be viewed holistically) tend to generalize by similarity.

Computationally, the above hypothesis is consistent with the findings by Wu et al. (2008). Specifically in their visual space of natural scenes, textons (low-entropy) (Zhu et al., 2005) can be composed by very simple concepts (Wu et al., 2010), akin to rule-based generalization. In comparison, textures (high-entropy) (Julesz, 1962) cannot be represented by rules (Zhu et al., 1997); instead, they are evaluated and generalized in terms of similarity by "pursuit" (Zhu et al., 1998). As such, we hypothesize that generalization shifts from similarity to rules as subjective complexity increases.

In the remainder of the paper, we first present the new metrics, Representativeness of Attribute (RoA), to measure the subjective complexity and analyze the computation-mode-shift in Sec. 2. Next, through a series of experiments, we provide strong evidence to support our hypotheses in Sec. 3; we draw the following conclusions in response to the two problems raised at the beginning:

1. *Representation:* the subjective complexity significantly falls in an inverted-U relation with the increment of visual complexity.
2. *Computation:* rule-based generalization is significantly positively correlated with the subjective complexity of the representation, while the trend is opposite for similarity-based generalization.

2 Bayesian Generalization and Complexities

In this section, we formulate Bayesian generalization for visual concept learning (Sec. 2.1), followed by the definitions of subjective complexity and visual complexity (Sec. 2.2).

2.1 Bayesian Generalization for Large-scale Visual Concept Learning

Concept-conditional modeling Let us consider $f : \mathbb{R}^D \mapsto \mathbb{R}^d$, which maps the input $\mathbf{x} \in \mathbb{R}^D$ to a representation vector $\mathbf{z} \in \mathbb{R}^d$. Here, f might be part of a discriminative model trained for visual categorization tasks, such as a prefix for a convolutional neural network without the last fully-connected layer for mapping $\mathbf{z}$ to the category vector $\mathbf{c} \in \mathbb{R}^c$. Training a discriminator for image categorization is to estimate the likelihood of concept c given a set of samples X: $P(c|X) = \prod_{x\in X} P(c|x;\theta)$, where θ is the parameter of f. Here, we assume that f provides a good estimation of $P(z|X;\theta)$; Tishby and Zaslavsky (2015) provides empirical evidence that a discriminative model may first learn how to extract proper attributes to model images X conditioned on c, then learn to discriminate their categories based on the attribute distribution. Some dimensions of $\mathbf{z}$ (usually $5\% \sim 10\%$ of the total dimensions) capture concrete semantic attributes of visual concepts when the activation score $f_z(X)$ is relatively high (Bau et al., 2020). Combining this concept-conditional measurement with attribute modeling, we rewrite the category prediction considering the attribute as a latent variable, and the observable joint distribution (X, c) is marginalized over z:

$$P(c|X;\theta) = \sum_{z\in\mathcal{Z}} P(c, z|X;\theta) = \sum_{z\in\mathcal{Z}} P(c|z)P(z|X;\theta), \tag{1}$$

where $\mathcal{Z}$ is the space of all attributes. This expression is essentially a Bayesian prediction view of visual categorization, which can be derived to Bayesian generalization in the natural visual world.

Representativeness of Attribute (RoA) as an informative prior Statistically, we treat the concept-conditional attribute activation score as an estimation of the probability $P(z|c)$ that recalls an attribute z when referring to a concept c; this is similar to answering *"Describe how a dog looks like."* In the context of the natural visual world, we also have all activation scores generated by an attribute as an estimation of the probability $P(\hat{c}|z)$ that recalls all concepts $\hat{c} \neq c$ when referring to the attribute z; this is akin to answering *"What do you recall seeing a blue thing in a ball shape?"*

Given the above observations and inspiration from Tenenbaum and Griffiths (2001b)), we formally define the RoA of a specific attribute z_i for concept c as:

$$\mathrm{RoA}(z_i, c) = \log \frac{P(z_i|c)}{\sum_{\hat{c}\neq c} P(\hat{c})P(z_i|\hat{c})}, \tag{2}$$

where $P(\hat{c})$ is the prior of concepts in the context. We hypothesize that humans estimate $P(\hat{c})$ through both language derivation and visual experience, essentially calculating the co-occurrence frequency between visual attributes and categorical attributes over the joint distribution $P(z_i, \hat{c})$. Hence, modeling RoA with large-scale image datasets and language corpus should yield human-level prior modeling. On this basis, we use f to statistically estimate $P(z_i, \hat{c})$ (Xie et al., 2020):

$$\mathrm{RoA}(z_i, c) = \log \frac{P(z_i|c)}{\sum_{\hat{c}\neq c} P(\hat{c})P(z_i|\hat{c})} \propto \log \frac{P(z_i|c;\theta)}{\sum_{\hat{c}\neq c} P(\hat{c}|z_i;\theta)}, \tag{3}$$

where $P(z|c;\theta)$ and $P(\hat{c}|z;\theta)$ are estimations of $P(z_i|c)$ and $P(z_i, \hat{c})$, respectively.

Generalize to the unseen Given an appropriate modeling of $P(z|X;\theta)$, the goal is to generalize an unknown concept c' to a small set of unseen examples $\hat{X} = \{x_1, \cdots, x_n\}$, where n tends to be a small integer. The generalization function $P(c'|\hat{X})$ is given by:

$$\begin{aligned} P(c'|\hat{X}) &= \sum_{z\in\mathcal{Z}} P(c'|z)P(z|\hat{X};\theta) = \sum_{z\in\mathcal{Z}} \frac{P(c')P(z|c')}{\sum_{c\in\mathcal{C}} P(c)P(z|c)} P(z|\hat{X};\theta) \\ &\propto \underbrace{P(c')}_{\substack{\text{uninformative}\\\text{prior}}} \sum_{z\in\mathcal{Z}} \underbrace{\exp\left(\mathrm{RoA}(z, c')\right)}_{\substack{\text{informative}\\\text{prior}}} P(z|\hat{X};\theta), \end{aligned} \tag{4}$$

where the uninformative prior $P(c')$ encodes the computation-mode-shift. Specifically, the similarity-based generalization $\langle c :: c' \rangle$ between a pair of concepts is defined as $\exists c \in \mathcal{C}, \sigma_0(c, c') < \delta$, where δ is a relative small neighbour. Similarly, the rule-based generalization $\langle c_1 : c_2 :: c_3 : c' \rangle$ over a quadruple of concepts is defined as $\exists c_1, c_2, c_3 \in \mathcal{C}, \sigma_1(c_1 - c_2, c_3 - c') < \delta$, where $\sigma_N(\cdot,\cdot)$ is an arbitrary metric measurement with an N-order input. Inspired by Tenenbaum (1999), we define $P(c') \propto \sigma_0\sigma_1/(\sigma_0 + \sigma_1)$, resulting in the simplest hypotheses of concepts: The harmonic property keeps guide to similarity-based generalization if σ_0 is dominating, and vice versa.

2.2 Complexities

Visual complexity Visual concepts come with diverse complexity, from very simple geometry concepts such as *squares* and *triangles*, to very complex natural concepts such as *dogs* and *cats*. Inspired by Wu et al. (2008), we indicate concept-wise visual complexity by Shannon's information entropy (Shannon, 1948). Formally, for a set of images $X = \{x_1, x_2, \cdots\}$ belonging to a concept c, the concept-wise entropy is $H(X|c) = \mathbb{E}_{X\sim P(\cdot|c)}[\log P(X|c)]$. As shown in Fig. 2, we compute the visual complexity and order some commonly known image datasets: 2D geometries (El Korchi and Ghanou, 2020), single concepts (Gill et al., 2022), compositional-attribute objects (Zhang et al., 2021; Johnson et al., 2017), human-made objects (Deng et al., 2009), scenes (Zhou et al., 2017), and animals (Xian et al., 2018; Deng et al., 2009).

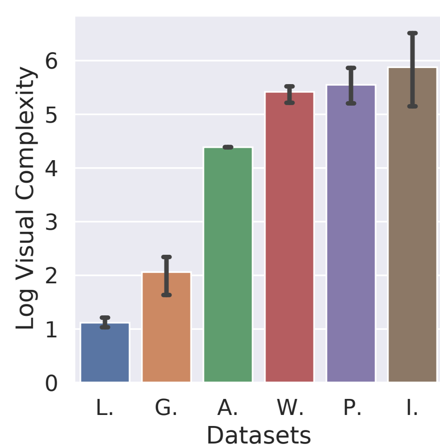


Figure 2: **Visual complexity of various datasets.**

Subjective complexity We quantify the subjective complexity over the prior model by Kolmogorov Complexity (Li et al., 2008). We calculate the minimum description length, *i.e.*, the minimum number of attributes we use to discriminate a concept. Specifically, for each concept c, we rank all attributes $z \in \mathcal{Z}$ by $\mathrm{RoA}(z, c)$ decreasingly, such that $\forall i, j \in [1, d], i < j, \mathrm{RoA}(z_i, c) \geq \mathrm{RoA}(z_j, c)$. Starting from $K = 1$, for each iteration, we select the top-K attributes and check whether these attributes can distinguish the concept c from the others. This process continues if the current iteration cannot distinguish it from the others. Formally, we define subjective complexity of visual concept $L(\hat{c})$ as:

$$L(\hat{c}) = \min_{K} \mathbb{1}\big(P(\hat{c} \neq c) < \epsilon \mid c = \arg\max_{c} P(c|z_1, \cdots, z_K; \phi)\big), \tag{5}$$

where ϵ is the error rate threshold, and ϕ is the parameter of the suffix of f in the same discriminative model for visual categorization (*e.g.*, the fully-connected layer). We calculate $P(c|z_1, \cdots, z_K; \phi)$ by removing the neurons' effects corresponding to $z_{K+1}, \cdots, z_d$ (Bau et al., 2020). Instead of maintaining all error rate thresholds, we leverage the *accuracy gain* between every two iterations to search for the minimum K. This process yields the concept-wise subjective complexity in RoA.

3 Empirical Analysis

In this section, we provide evidences and analyses to validate the above hypotheses. We (i) conduct empirical analyses at both *representation* (Sec. 3.1) and *computational* (Sec. 3.2) level; (ii) provide quantitative analyses the computation-mode-shift w.r.t. the concept complexity in Sec. 3.2); and (iii) provide qualitative analyses to interpret from the aspect of natural image statistics in Sec. 3.3.

3.1 Representation *vs.* Complexity

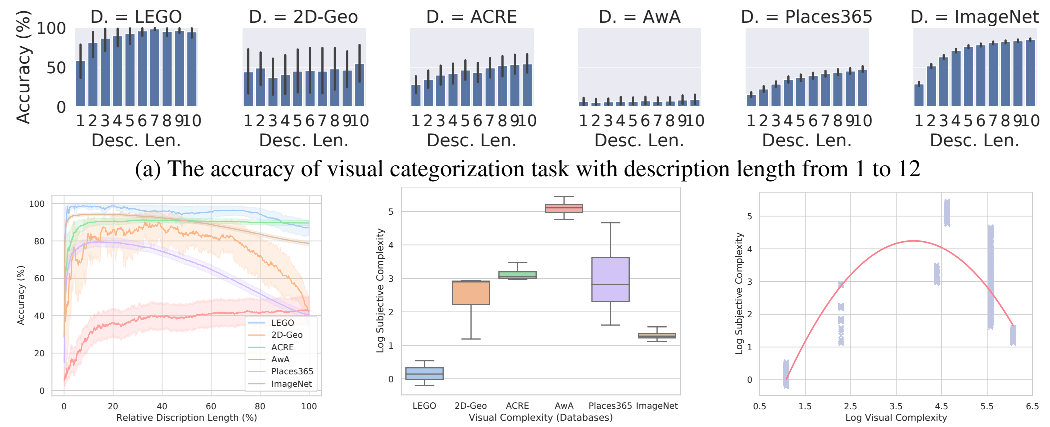
This experiment investigates the subjective complexity of visual concepts through visual categorization. Our predictions were that models use attributes with high RoA to describe visual concepts, and the description length falls in an inverted-U relation with the increment of visual complexity.

Method Six discriminative models are trained from scratch on six datasets with the supervision of concept labels. The six datasets are LEGO (Tatman, 2017), 2D-Geo (El Korchi and Ghanou, 2020), ACRE (Zhang et al., 2021), AwA (Xian et al., 2018), Places (Zhou et al., 2017), and ImageNet (Deng et al., 2009), ordered as the increment of concept-wise visual complexity. The six models are all optimized to convergence on the training set and are tuned to the best hyper-parameters on the validation set. The readers are referred to Appendix B for details.

During the evaluation, RoA is calculated for each attribute in the context of all concepts for each dataset. Following the protocol described in Sec. 2.2, the models conduct visual categorization tasks from exploiting only one attribute with the highest RoA to exploiting the entire attribute space.

Results The main quantitative results are illustrated in Fig. 3. Subjective complexity shows significant diversity between the datasets. The logarithm values are as follows; see Fig. 3c. LEGO: $.10\,(CI = [-.10, .52],\, p < .05)$, 2D-Geo: $2.91\,(CI = [1.21, 2.95],\, p < .05)$, ACRE: $3.08\,(CI = [2.99, 3.46],\, p < .05)$, AwA: $5.08\,(CI = [4.82, 5.36],\, p < .05)$, Places: $2.74\,(CI = [1.63, 4.72],\, p < .05)$, ImageNet: $1.28\,(CI = [1.16, 1.51],\, p < .05)$. All models rely on only a few (less than 20% of all) attributes to reach the prediction accuracy comparable with prediction accuracy exploiting all attributes; see Fig. 3b; most models (5 out of 6) exploit very few (less than 5% of all) attributes to reach a higher accuracy than that of all attributes; see Fig. 3b. The models for the simplest dataset (*i.e.*, LEGO) and for the most complex dataset (*i.e.*, ImageNet) obtain a large accuracy gain (over 10%) with the description length from 0 to 3 and obtain smaller accuracy subsequently. In comparison, the models for ACRE and Places obtain relatively small accuracy gain (about 5%) with description length from 0 to 8; see Fig. 3a. Fig. 3d shows the estimated inverted-U relation between subjective complexity and visual complexity. Following the "two-lines" test proposed by Simonsohn (2018), the relation is relatively robust across the datasets, decomposing the non-monotonic relation via a "breakpoint"; the positive linear relation ($b = 1.10, z = 253.76, p < 1e - 4$) and the negative linear relation ($b = -2.57, z = -659.26, p < 1e - 4$) are both significant.

Discussion The above results reveal that (i) the representation helps the models to describe concepts with very few attributes, (ii) representation trained from very simple or very complex datasets usually have a shorter concept description length than those trained on other datasets, and (iii) the subjective complexity significantly comes in an inverted-U relation with the visual complexity.



(a) The accuracy of visual categorization task with description length from 1 to 12

(b) Visual categorization accuracy with various description lengths

(c) Average subjective complexity on different datasets

(d) Estimated inverted-U relation between visual and subjective complexity

Figure 3: **Quantitative results of *Representation* vs. *Complexity*.** (vector graphics; zoom for details)

3.2 COMPUTATION *vs.* COMPLEXITY

This experiment evaluates the capability of rule- and similarity-based generalization by the representations in Sec. 3.1. Our predictions were that under the same evaluation protocol, representations with relatively high subjective complexity outperform those with low subjective complexity in rule-based generalization, while the trend is opposite in similarity-based generalization.

Method The evaluation of generalization is designed with two phases: in-domain and out-of-domain generalizations. The former consists of unseen samples from the test set of ACRE and ImageNet, whereas the latter contains unseen samples of unknown concepts collected from the internet. Each phase has a dataset with pairs for similarity-based generalization evaluation and a dataset with quadruples for rule-based generalization evaluation.

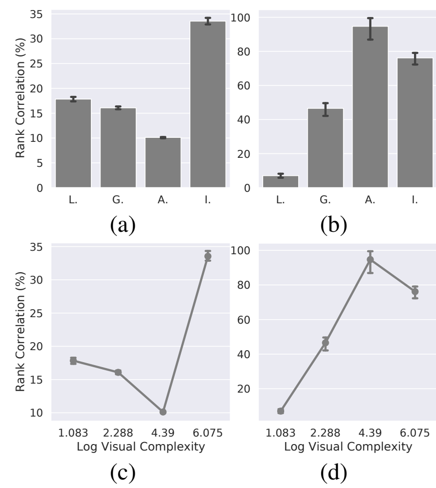


Figure 4: **Quantitative results of *Computation* vs. *Complexity*.** (a)(b) The rank correlation of similarity- and rule-based generalization with the four representations trained from four datasets. (c)(d) The rank correlation of similarity- and rule-based generalization according to the visual complexity. (L: LEGO, G:2D-Geo, A: ACRE, I: ImageNet) These plots reflect the landscape in Fig. 1.

The evaluation protocol for similarity-based generalization extends its definition in Sec. 2.1. Formally, given unknown concept c' and known concepts $c \in \mathcal{C}$, there is a ranking of the pairwise metric measurement $S = \{\sigma_0(c_i, c') \geq \sigma_0(c_j, c') | c_i, c_j \in \mathcal{C}\}$. The representation ranking S_r is obtained by the cosine similarity between two representation vectors $cos(z_i, z')$. The ground-truth ranking S_h is obtained by human judgment. Hence, the generalization capability of the representation can be quantified through the rank correlation coefficient introduced in Spearman (1961) as an accuracy measurement. Similarly, the evaluation protocol for rule-based generalization is defined as follows. Given incomplete rule $r'(c_3, c')$ and known rules $r_i(c_1, c_2) \in \mathcal{R}$, the ranking score R_r of representation is reduced to a cosine similarity calculation $cos(z_2 - z_1 + z_3, c')$, and the ground-truth ranking R_h is obtained by human judgment (Mikolov et al., 2013). We obtain the ground-truth concepts by literal meanings through the language representation model GloVe (Pennington et al., 2014). The image examples are retrieved from datasets (in-domain) or the internet (out-of-domain) with label embedding matching (Vendrov et al., 2015).

Results The quantitative results for in-domain generalization evaluation are illustrated in Fig. 4. In similarity-based generalization, the representation trained from ImageNet outperforms others (over 15%), and LEGO outperforms its more complex counterparts 2D-Geo and ACRE (over 10%). In rule-based generalization, the representation trained from ACRE outperforms its more complex counterpart ImageNet (over 20%). Though the models trained on ImageNet and ACRE reach the highest accuracy on similarity- and rule-based generalization, this is not likely due to over-fitting in training: The objective of visual categorization is different from that of generalization, thus the over-fitting on one visual categorization would not result in an over-fitting on other objectives. Intuitively,

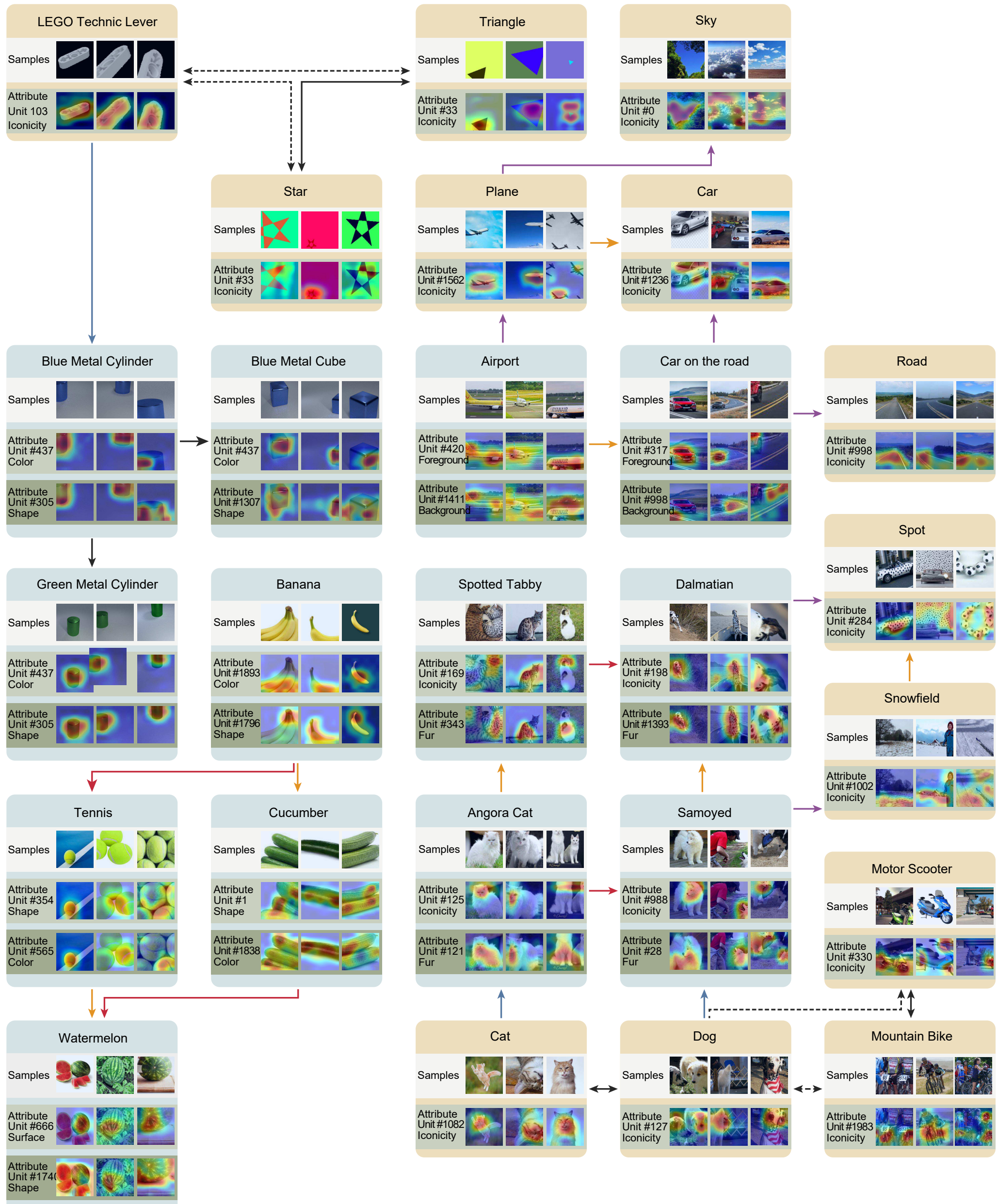


Figure 5: **A landscape of similarity- and rule-based generalization over concepts with relatively high and low subjective complexity, considering both concept complexities and concept hierarchy. Bidirectional arrows** denote the similarity judgment between concepts, wherein concepts linked by solid lines are more similar than those linked by dashed lines. Arrows denote rules over concepts. Rule-based generalization in basic-level generalizes **given rules** to **unknown rules**. Similarity shifts to rules when the sample hierarchy goes **from superordinate-level to subordinate-level** (*e.g.*, from *block* to *blue cylinder*, from *cat* to *angora cat*). Rules shift to similarity as the sample hierarchy goes **from subordinate-level to superordinate-level** (*e.g.*, from *car on the road* to *car*, from *dalmatian* to *spot*). The **confusing similarity judgment** between blue cylinder, blue cube, and green cylinder with distinct and shared attributes is also an interesting observation.

representations trained on more complex dataset span more complex attribute spaces. However, the result implies that the shift between similarity- and rule-based generalization is non-monotonic as the dataset complexity increases; it is more correlated to the subjective complexity based on Sec. 3.1. Hence, there is a significant negative relationship between the similarity-based generalization and the subjective complexity ($r = -.48$, $p < .05$), and a significant positive relationship between the rule-based generalization and the subjective complexity ($r = .68$, $p < .01$).

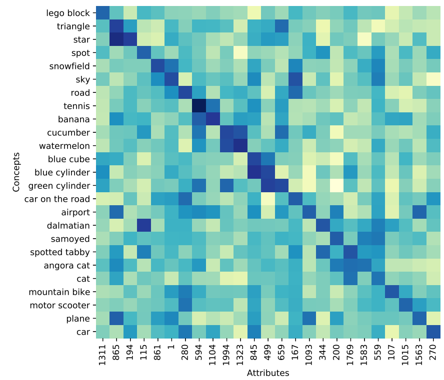


Figure 6: **Visualization of the RoA matrix.** Most (21 out of 25) of the concepts are unknown; high saturation indicates high RoA value. The diagonal elements are the most representative attributes for all concepts.

Fig. 5 illustrates the qualitative results for out-of-domain generalization. As shown in Fig. 6, though never tuned on the unseen examples, the representation model also captures representative attributes for unknown concepts, which supports our argument in Sec. 2.1 that RoA has the potential to serve as a prior for Bayesian generalization. Further, we visualize the most representative attributes of each concept by upsampling the activated feature vector to the size of the original image (Bau et al., 2020); the attributes are located around the peaks. Most attributes with high RoA are explainable, such as the shape attribute shared by *blue cylinder* and *green cylinder*, shape and color captured by two distinct attributes in *banana* and *watermelon*, and foreground object (plane, car) and background (road, field) attributes in *airport* and *car on the road*. Those concepts with more than one meaningful attributes are sensitive to rule-based generalization. By contrast, those concepts with only one meaningful attribute, such as *dog-like face* for *dog*, *car-like shape* for *car*, are sensitive to similarity-based generalization.

Discussion The above experiment reveals that (i) both similarity- and rule-based generalizations are not significantly related to the visual complexity of datasets, (ii) the capability of similarity-based generalization has a significant negative relationship with the subjective complexity of representation, and (iii) the capability of rule-based generalization has a positive relationship with the subjective complexity of representation. We empirically articulate that the computation-mode-shift significantly exists, and similarity shifts to rules as the subjective complexity increases (see Appendix C.3).

3.3 A Statistical Interpretation

Subjective complexity in natural image statistics According to algorithmic information theory (Chaitin, 1977), a concept's subjective complexity is proportional to the probability to perceive this concept. This is consistent with the subjective complexity of visual concepts defined in our work. An attribute z is representative for concept c when $\mathrm{RoA}(z, c)$ is relatively high; we have a high probability to observe the attribute by the concept (*e.g.*, $P(z|c) = 1$) or only by the concept (*i.e.*, $\sum_{\hat{c}\neq c} P(\hat{c}|z)$ is small). Specifically, complex concepts (*e.g.*, *dog*, *cat*), though consist of many attributes (*e.g.*, *fur*, *ear*), tend to have a unique attribute of *view as a whole* to distinguish these concepts from others because we can hardly observe them in other concepts. Conversely, simple concepts (*e.g.*, *circle*, *cylinder*) can be observed by many other concepts (*e.g.*, *wheel*, *chimney*) and also have other attributes (*e.g.*, *number of angles*, *smoothness*). Nevertheless, the attribute *shape* is one of the simple attributes to describe these concepts; representation of these concepts emerges iconicity (Guo et al., 2003; Fay et al., 2010; 2013).

Meanwhile, for those concepts that are either too simple or too complex (*e.g.*, *watermelon*, *airport*), no unique or a simple attribute can distinguish them from others; *i.e.*, $\mathrm{RoA}(z, c)$ is not high. In these cases, we have to describe them with more attributes. Of note, this interpretation is also in line with the principle of rational reference (Frank and Goodman, 2012; Goodman and Frank, 2016).

From similarity to rules Since similarity gradient can be viewed as a partial order defined on a single set (Tenenbaum, 1999), sorting hypotheses requires numerical comparison in the same domain. Hence, similarity judgment in a single attribute space z_i is simply calculating the similarity between concepts c_j and c_k by $d(z_i^{(j)}, z_i^{(k)})$, where $d(\cdot, \cdot)$ can be an arbitrary similarity or distance metric (Ontañón, 2020). As the number of independent attribute spaces increases (*i.e.*, subjective complexity increases), the similarity becomes subtle as we have to consider multiple independent attributes. Of note, the attribute spaces are those obtained after dimension reduction (Xie et al., 2016). According to high-dimension geometry, those concept representations are almost distributed uniformly (Blum et al., 2020), unless we assign weights to different attribute spaces by only considering very few attributes. For example, *watermelon* is similar to *tennis* in the attribute space of *shape*, but it becomes *cucumber* in the attribute space of *color*; *airport* is similar to *plane* in the attribute space of *foreground object* and is similar to *land and sky* in the attribute space of *background context*. In this work, we reduce similarity judgment over multiple attribute spaces to rules defining relations over two concepts: At least one shared attribute space serves as the bridge to connect the two concepts.

From rules to similarities As the number of independent attribute spaces (*i.e.*, subjective complexity) decreases, rules are moved back to similarity. For example, we have the rule relating *dalmatian* to *spotted tabby* by *fur texture*, and can generalize it to *samoyed to angora cat*. However, when the concepts are more complex (*e.g.*, *dalmatian* and *samoyed* fall in *dog*, or *spotted tabby* and *angora cat* belong to *cat*), rules are difficult on these concepts; instead, we directly apply similarity judgment.

Concept complexities and hierarchy When visual complexity moves from low to high, we have visual concepts move from *simple and universal* to *complex and unique*. Inspired by Xu and Tenenbaum (2007), we argue that these two ends consist of *superordinate* concepts, usually on higher hierarchies. Objects such as *watermelon*, attribute-specified animals such as *samoyed*, are subordinate concepts of *ball* and *dog*, respectively; scenes such as *airport* are compositions of subordinate concepts like *plane* and *land and sky*. In a top-down view, we have concepts with increasing subjective complexity and more shared attribute spaces to generalize by rules. In a bottom-up view, the attribute spaces are reduced to the *simple* or *unique* ones, easy for similarity judgment.

4 Discussion and Conclusion

We have analyzed the complexity of concept generalization in natural visual world, in Marr's *representational* and *computational level* respectively, with the rational analysis of Representativeness of Attribute (RoA). At the representational level, the subjective complexities significantly fall in an inverted-U relation with the increment of visual complexity. At the computational level, the rule-based generalization is significantly positively correlated with the subjective complexity of the representation, while the trend is the opposite in similarity-based generalization. RoA bridges the two levels by unifying the frequentist properties of natural images (sensory-based) and the Bayesian properties of concepts (knowledge-derived) (Bi, 2021). It is easy to obtain based on observations, is flexible to an extent, and captures contextual rationality, thus may serve as humans' *visual common sense* (Zhu et al., 2020). The readers are referred to Appendix A for additional remarks.

The appropriateness of the computational modeling Thanks to Marr's paradigm (Marr, 1982), we could separate the computational-level problem and the representational-level problem, where we study computation problems regardless of their algorithmic representation or physical implementation in either humans or machines (Lake et al., 2015). Hence, under the same computation problem, whether the algorithm is neural networks or brain circuits is not the problem in the scope.

Since the two parts of our computation problem—Bayesian generalization (Tenenbaum and Griffiths, 2001a) and subjective complexity (Lake et al., 2017)—have established solid backgrounds in human cognition, we have a sufficient prerequisite for studying the complexities in the natural visual world. Though there may be infinite interpretations of human cognitive models (Lieder and Griffiths, 2020), constrained by previous theories and the principle of resource-rational analysis (Lieder and Griffiths, 2020; Gershman et al., 2015), we can make assumptions about the Bayesian derivations.

Limitations and future work The limitations of this work lead to several future directions: (i) We only demonstrated the inverted-U relation and the correlation empirically. Can we provide them theoretically, from the aspect of information theory and statistics? (ii) On the basis of this work, can we extend the generalization evaluation to a larger scale, that helps to probe the continuum space between similarity and rules quantitatively? (iii) Are our findings consistent with those in other environments, where the concepts are represented in different modalities (*e.g.*, language, audio, and tactile)? (iv) If using only a few attributes with high RoA improves the accuracy of the visual categorization task, as Sec. 3.1 suggests, can we build an algorithm that samples from RoA adaptively to task and data distribution, to obtain stronger generalization ability? (v) If RoA reflects humans' visual commonsense, can we model the communications between individuals toward commonsense knowledge as a pursuit of the common grounds on representative attributes for the concepts to be communicated (Tomasello, 2010)? With many questions unanswered, we hope to shed light on future research on Bayesian generalization.

REFERENCES

Abbott, J., Austerweil, J., and Griffiths, T. (2012). Constructing a hypothesis space from the web for large-scale bayesian word learning. In *Annual Meeting of the Cognitive Science Society (CogSci)*. 2

Abbott, J. T., Heller, K. A., Ghahramani, Z., and Griffiths, T. (2011). Testing a bayesian measure of representativeness using a large image database. In *Advances in Neural Information Processing Systems (NIPS)*. 2

Allen, C. and Hospedales, T. (2019). Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning (ICML)*. 3

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences (PNAS)*, 117(48):30071–30078. 3, 5, 8, 14

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, 25(10):883–895. 3, 9, 14

Blum, A., Hopcroft, J., and Kannan, R. (2020). *Foundations of data science*. Cambridge University Press. 8

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47. 16

Chaitin, G. J. (1977). Algorithmic information theory. *IBM Journal of Research and Development*, 21(4):350–359. 2, 8

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4, 5, 16

Donderi, D. C. (2006). Visual complexity: a review. *Psychological Bulletin*, 132(1):73. 1, 2

El Korchi, A. and Ghanou, Y. (2020). 2d geometric shapes dataset–for machine learning and pattern recognition. *Data in Brief*, 32:106090. 4, 5

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., Hewitt, L., Solar-Lezama, A., and Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*. 1

Fay, N., Arbib, M., and Garrod, S. (2013). How to bootstrap a human communication system. *Cognitive Science*, 37(7):1356–1367. 3, 8, 14

Fay, N., Ellison, M., and Garrod, S. (2014). Iconicity: From sign to system in human communication and language. *Pragmatics & Cognition*, 22(2):244–263. 3, 14

Fay, N., Garrod, S., Roberts, L., and Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, 34(3):351–386. 3, 8, 14

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998. 8

Gentner, D. and Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45. 3

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278. 9

Gill, H. S., Murugesan, G., Khehra, B. S., Sajja, G. S., Gupta, G., and Bhatt, A. (2022). Fruit recognition from images using deep learning applications. *Multimedia Tools and Applications*, pages 1–22. 4

Gittens, A., Achlioptas, D., and Mahoney, M. W. (2017). Skip-gram- zipf+ uniform= vector additivity. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 3

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829. 8

Griffiths, T. and Tenenbaum, J. (2003). From algorithmic to subjective randomness. In *Advances in Neural Information Processing Systems (NIPS)*. 1, 2

Griffiths, T. L., Abbott, J. T., and Hsu, A. S. (2016). Exploring human cognition using large image databases. *Topics in Cognitive Science*, 8(3):569–588. 2

Guo, C.-e., Zhu, S.-C., and Wu, Y. N. (2003). Towards a mathematical theory of primal sketch and sketchability. In *International Conference on Computer Vision (ICCV)*. 8

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 14

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4, 16

Julesz, B. (1962). Visual pattern discrimination. *IRE Transactions on Information Theory*, 8(2):84–92. 3

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, 55:271–304. 1, 14

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., and Murphy, K. (2017). Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*. 16

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338. 1, 9

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40. 9

Li, M., Vitányi, P., et al. (2008). *An introduction to Kolmogorov complexity and its applications*. Springer. 2, 5

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43. 9

Lind, J. T. and Mehlum, H. (2010). With or without u? the appropriate test for a u-shaped relationship. *Oxford bulletin of economics and statistics*, 72(1):109–118. 15

Marr, D. (1982). *Vision*. W. H. Freeman and Company. 2, 9

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*. 3, 6

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT Press. 2

Ontañón, S. (2020). An overview of distance and similarity functions for structured data. *Artificial Intelligence Review*, 53(7):5309–5351. 8

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. 6

Quine, W. V. O. (1960). *Word and object*. MIT Press. 1

Salakhutdinov, R., Tenenbaum, J., and Torralba, A. (2012). One-shot learning with a hierarchical nonparametric bayesian model. In *ICML Workshop on Unsupervised and Transfer Learning*. 2

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. 4

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323. 1

Simonsohn, U. (2018). Two lines: A valid alternative to the invalid testing of u-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science*, 1(4):538–555. 5

Sloman, S. A. and Rips, L. J. (1998). *Similarity and symbols in human thinking*. MIT Press. 1

Spearman, C. (1961). *The proof and measurement of association between two things*. Appleton-Century-Crofts. 6

Sun, Z. and Firestone, C. (2021). Seeing and speaking: How verbal "description length" encodes visual complexity. *Journal of Experimental Psychology: General*. 1, 2

Tatman, R. (2017). The lego parts, sets, colors, and inventories of every official lego set. 5

Tenenbaum, J. B. (1998). Bayesian modeling of human concept learning. In *Advances in Neural Information Processing Systems (NIPS)*. 1

Tenenbaum, J. B. (1999). Rules and similarity in concept learning. In *Advances in Neural Information Processing Systems (NIPS)*. 1, 4, 8, 14

Tenenbaum, J. B. and Griffiths, T. L. (2001a). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640. 1, 9

Tenenbaum, J. B. and Griffiths, T. L. (2001b). The rational basis of representativeness. In *Annual Meeting of the Cognitive Science Society (CogSci)*. 2, 4

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285. 1

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*. 3

Tomasello, M. (2010). *Origins of human communication*. MIT press. 9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 16

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2015). Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*. 6

Wolfram, S. et al. (2002). *A new kind of science*. Wolfram Media Champaign. 1

Wu, Y. N., Gao, R., Han, T., and Zhu, S.-C. (2019). A tale of three probabilistic families: Discriminative, descriptive, and generative models. *Quarterly of Applied Mathematics*, 77(2):423–465. 3

Wu, Y. N., Guo, C.-E., and Zhu, S.-C. (2008). From information scaling of natural images to regimes of statistical models. *Quarterly of Applied Mathematics*, pages 81–122. 3, 4, 14

Wu, Y. N., Si, Z., Gong, H., and Zhu, S.-C. (2010). Learning active basis model for object detection and recognition. *International Journal of Computer Vision (IJCV)*, 90(2):198–235. 3

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(9):2251–2265. 4, 5

Xie, J., Gao, R., Nijkamp, E., Zhu, S.-C., and Wu, Y. N. (2020). Representation learning: A statistical perspective. *Annual Review of Statistics and Its Application*, 7:303–335. 2, 3, 4

Xie, J., Lu, Y., Zhu, S.-C., and Wu, Y. N. (2016). A theory of generative convnet. In *International Conference on Machine Learning (ICML)*. 3, 8

Xu, F. and Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, 114(2):245. 1, 2, 9, 17

Zhang, C., Jia, B., Edmonds, M., Zhu, S.-C., and Zhu, Y. (2021). Acre: Abstract causal reasoning beyond covariation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4, 5, 16

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2017). Places: A 10 million image database for scene recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6):1452–1464. 4, 5

Zhu, S.-C., Guo, C.-E., Wang, Y., and Xu, Z. (2005). What are textons? *International Journal of Computer Vision (IJCV)*, 62(1):121–143. 3

Zhu, S.-C., Wu, Y. N., and Mumford, D. (1997). Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8):1627–1660. 3

Zhu, S.-C., Wu, Y. N., and Mumford, D. (1998). Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision (IJCV)*, 27(2):107–126. 3

Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y. N., et al. (2020). Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345. 9

A Additional Remarks

A.1 The Uniqueness of the Natural Visual World

Why do we only use the modality of vision to investigate the complexity of Bayesian generalization? Vision is unique for the diverse complexities both in the natural visual world and in the semantic space (Kersten et al., 2004), which relates vision to the discussion of levels of abstraction (Wu et al., 2008). To some extent, vision serves as the bridge between abstract language-derived knowledge and perceptual sensory-derived knowledge (Bi, 2021). The two ends of the continuum of Bayesian generalization touch the functional essences of rule-based symbolic signals and similarity-based perceptual signals (Tenenbaum, 1999). In particular, the very final development of symbolic signals leads to the emergence of language, based on the compositionality of symbols and rules as the basic feature of language. The psychology literature supports the hypothesis that language is emerged from visual communications by abstracting visual concepts toward hieroglyphs through their iconicity (Fay et al., 2010; 2013; 2014). Both simple and universal visual concepts, such as geometric shapes, and complex and unique visual concepts, such as animals and artificial objects, are all related to corresponding abstract concepts by iconicity. By contrast, those concepts that are neither simple nor unique are unlikely to be abstracted by iconicity since they are described by multiple representative attributes—though each attribute can be generalized through iconicity respectively, putting different attribute spaces together is not making sense—by contrast, those concepts naturally satisfy the compositionality of language, thus are appropriate for rule-based generalization. In this sense, vision is not only a modality of data but is the hallmark of human intelligence, evolving perceptual sensory toward language for communications. Hence, vision is meaningful and sufficient for investigating the complexity of Bayesian generalization.

Consider other modalities, say audio, the second common resource of sensory input. Although we could define audio complexity and try to correlate it with subjective complexity, audio is only a perceptual sensory—abstraction of raw audio is not related to any semantic meaning, thus does not provide much insight on human intelligence; also the diversity of audio complexity is far less than its visual counterpart. Hence, generalizing the experiments to audio data may be a bonus but never provides us insights as deep as that provided by visual data.

B Implementation Details

B.1 Implementing basic discriminative models

The basic discriminative models are employed from ResNet (He et al., 2016), thus the feature space is spanned by a 512-d or 2048-d feature vector (dimensions are different by the different depths of ResNet architecture). All models are trained on eight NVIDIA A100 80GB GPUs.

B.2 Implementing RoA

In general, the RoA computes a score for each attribute z_i over each concept c. The output of RoA is a matrix where the column space is the context of all the concepts in the natural visual world, and the row space is all the attributes. Assume we have three samples $\{x^{(1)}, x^{(2)}, x^{(3)}\} \in X$ of concept c, then the output of f provides the attribute vectors $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \mathbf{z}^{(3)} \in \mathbb{R}^{H\times W\times D}$ respectively. We then adaptively pool each feature map $\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}, \mathbf{z}_i^{(3)} \in \mathbb{R}^{H\times W\times d}$ in each dimension of the attribute vector to a scalar $z_i^{(1)}, z_i^{(2)}, z_i^{(3)}$, thus $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \mathbf{z}^{(3)} \in \mathbb{R}^d$. $P(z_i|c)$ is calculated by normalizing over the dimensions of centroid vector of all $\mathbf{z}$ given the set of samples of concept c, *e.g.*, $\widetilde{\overline{\mathbf{z}}^{(k)}}, k = 1, 2, 3$.

B.3 Implementing Subjective Complexity Measurement

Since the calculation of the absolute value of $L(\hat{c})$ may encounter multiple solutions, we employ accuracy gain (Bau et al., 2020) to compute a relative $L(\hat{c})$ specifically for $\hat{c}$. The accuracy gain approach considers the categorization accuracy difference for a single concept before and after removing the effect of a specific neuron, defined as:

$$\begin{aligned}\Delta Acc_K(\hat{c}) = P\big(\hat{c} = c \big| c = \arg_c \max P(c|z_1, \ldots, z_K; \phi)\big) \\ -P\big(\hat{c} = c \big| c = \arg_c \max P(c|z_1, \ldots, z_{K-1}; \phi)\big),\end{aligned} \tag{6}$$

where $K \geq 2$ and $\Delta Acc_1(\hat{c}) = P\big(\hat{c} = c \big| c = \arg_c \max P(c|z_1; \phi)\big)$. Hence, the relative $L(\hat{c})$ is exactly computed by:

$$L_{relative}(\hat{c}) = \min_K \max \; Acc_K(c), \tag{7}$$

which serves as the heuristic to search for the minimum K to calculate absolute $L(\hat{c})$.

C Method Appropriateness Checking

C.1 Checking the assumptions of the two-lines test

The "two-lines test" requires a weaker assumption than the mostly used quadratic regression test for testing U-shapes (Lind and Mehlum, 2010), hence the former is employed instead of the latter. Let $y = f(x)$ be the ground-truth function, the U-shape assumes only a sign flip effect in discrete data, where there exists x_c such that $f'(x), x \leq x_c$ and $f'(x), x \geq x_c$ has opposite signs (Lind and Mehlum, 2010). To note, since the data is originally discrete, there is no need to check the existence of $f'(x)$ because it is estimated based on the discrete data points. Hence, the basic hypothesis of the U-shape is that at least one such x_c exists, and the null hypothesis is that no such x_c exists. The null hypothesis is rejected by estimating many x_c values and run two separate linear regressions for $x \leq x_c$ and $x \geq x_c$ respectively. The fact that two regression lines are of opposite sign rejects the null hypothesis. By contrast, the quadratic regression test assumes that the first-order derivative function $f'(x)$ is continuous in the domain. Hence, there is no need to employ the quadratic regression test.

C.2 Checking the assumptions of the linear regression test

The assumptions of the test are (1) linearity of the data; (2) x values are statistically independent; (3) the errors are homoscedastic and normally distributed. We did test the applicability of the linear regression test: (1) the two relations between rank correlation and subjective complexity are intuitively in lines ($[(0.1, 7.8), (1.28, 79.1), (2.91, 46.7), (3.08, 99.5)]$ and $[(0.1, 17.1), (1.28, 33.2), (2.91, 15.8), (3.08, 10.2)]$); (2) all the evaluations are run separately with different random seeds, thus the predictors are statistically independent; (3) since the only independent variable is the dataset, which is not likely to be the source of constant variance of the errors, the errors are homoscedastic. Consider the null hypothesis of the linear regression test that the coefficient β_1 is zero, which leads to a trivial solution. However, the p-values of both the positive and negative relations are less than 0.05, rejecting the null hypothesis.

C.3 The correctness for combining representation and computation

As illustrated in Fig. 4, we integrated the results in representation *vs*. complexity into this plot to use these plots to demonstrate the computation-mode-shift—the two U-shapes come with opposite trends intuitively show the landscape for concept complexity *vs*. the computation mode, that similarity-based generalization tends to emerge in concepts with very low or very high visual complexity (*i.e.*, the concepts with low subjective complexity, on the left and right ends of the visual complexity axis), and rule-based generalization tends to emerge in concepts with neither very low nor very high visual complexity (*i.e.*, the concepts with high subjective complexity, in the middle of the visual complexity axis). This is the exact claim of the paper. The quantitative results on the significant positive relation between rule-based generalization rank correlation and subjective complexity, and the significant negative relation between similarity-based generalization rank correlation and subjective complexity, both support the claim.

D Dataset Construction

D.1 Definition of the Vocabulary

We leverage a fully-connected probabilistic graph model to obtain the representativeness of every attribute for every concept, where each node is a piece of natural language that serves as either a concept or an attribute describing other concepts. We exploit the RoA in language to generate the in-domain and out-of-domain visual datasets for Bayesian generalization. Technically, we use the

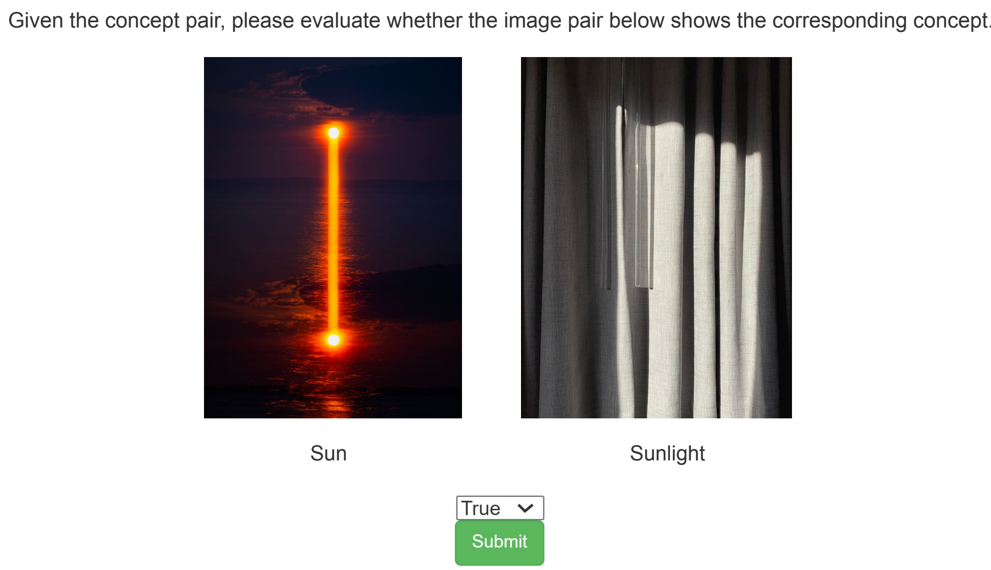


Figure 7: **The Amazon Mechanical Turk (AMT) interface used to collect human judgments.**

vocabulary from a WordPiece model (e.g., the base version of Bert Vaswani et al. (2017)), where a word is tokenized into word pieces (also known as subwords) so that each word piece is an element of the dictionary. Non-word-initial units are prefixed with the sign "##" as a continuation symbol. In this way, there is no Out-Of-Vocabulary. This brings the benefit of generalization over all words. Using all these words as attributes or features leads to sufficient coverage. Moreover, some symbols are reserved for unused placeholders, leaving room for features that the language cannot describe. The readers can refer to `vocab.txt` in the supplementary materials for more details about the attribute list.

D.2 Human-in-the-loop dataset validation

We constructed the *similarity-based generalization* and the *rule-based generalization* datasets using both manual approaches and automatic approaches. Details of all datasets are demonstrated in Tab. 1.

For *in-domain similarity-based generalization*, a concept pair with a human-annotated similarity score was first retrieved from MEN dataset (Bruni et al., 2014) and ImageNet dataset (Deng et al., 2009). Next, we used AMT to crowd-source the image aligned to the concept. In total, 305 pairs were selected from 500 candidates. One image was aligned to each concept.

For *in-domain rule-based generalization*, we generated dataset using objects of easy-to-disentangle attributes (*e.g.*, shape and color) (Zhang et al., 2021; Johnson et al., 2017). Based on these attributes, we constructed the quadruple relation (*e.g.*, `blue cube:red cube::blue cylinder:red cylinder`). In total, 4800 images and 24 quadruple relations were collected.

For *out-of-domain similarity-based generalization* and *out-of-domain rule-based generalization*, we collected images from an open internet image dataset (Krasin et al., 2017) based on a predefined set of similarity pairs and rule quadruples. Of note, all the selected pairs or rules were uniformly sampled from the dataset instead of manually picked. All the selected images were under human validation.

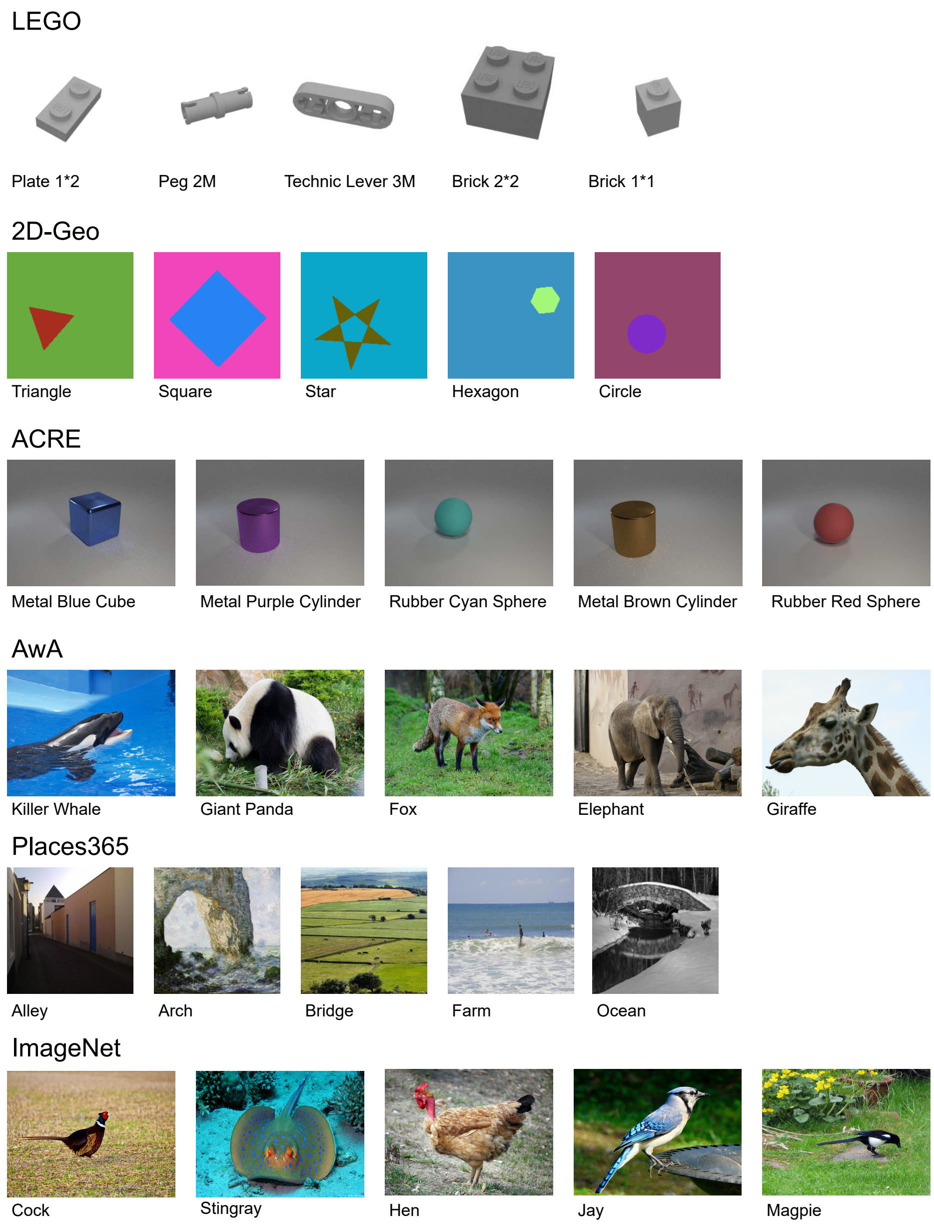


Figure 8: **Examples of datasets used in our work.**

In the study, AMT workers recruited have acceptance rates higher than or equal to 90% and approved hits more than 500. Each AMT worker was compensated at the rate of 0.01 dollar per selection. In total, we have tested 1000 judgments for 500 concept pairs; two judgments per pair. Fig. 7 shows an example of the AMT interface.

Table 1: **Details of the datasets for generalization evaluation.** Subordinate level indicates the concept being generalized to is a subordinate concept of the known ones, whereas superordinate level indicates the concept being generalized to is a superordinate concept of the known ones. Subordinate level, basic level, and superordinate level are terms introduced in Xu and Tenenbaum (2007).

Group	In-domain		Out-of-domain			
Generalization type	Similarity-based	Rule-based	Similarity-based	Rule-based		
Concept hierarchy	basic level	basic level	basic level	basic level	subordinate level	superordinate level
Test-set size	305	24	21	10	10	10

E ADDITIONAL RESULTS

E.1 THE CONVERGENCE OF REPRESENTATION *vs.* GENERALIZATION

Does the training setting of the representation model affect its generalization ability? Fig. 9 shows the rank correlation on in-domain generalization evaluation w.r.t. the number of training epochs for visual categorization. This result empirically shows that the generalization ability converges when the representation models are well trained after 6-10 epochs, and that ability is stable after convergence. The regression line is significantly vertical to the y-axis ($b = .03, a = 69.67, p < 1e-4$). Hence, we can assume that there are no significant distinctions of generalization ability between representation models being trained to convergence but with different training settings.

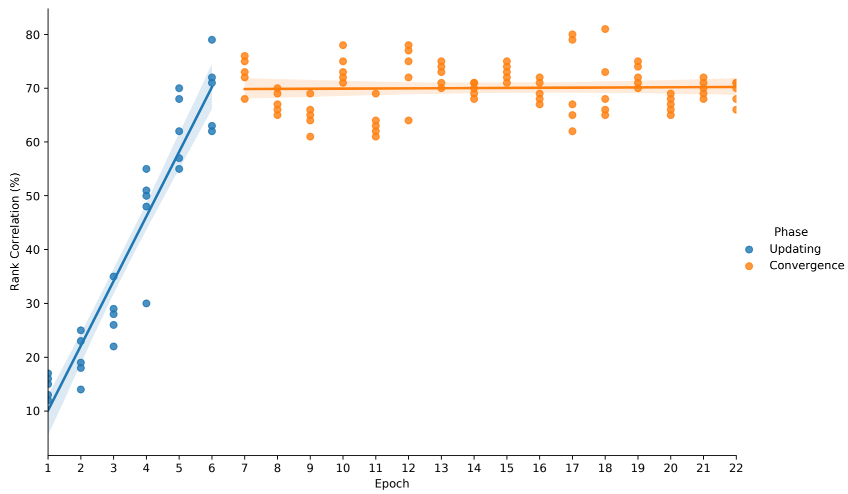


Figure 9: **Rank correlation of generalization w.r.t. the number of training epochs for visual categorization.**

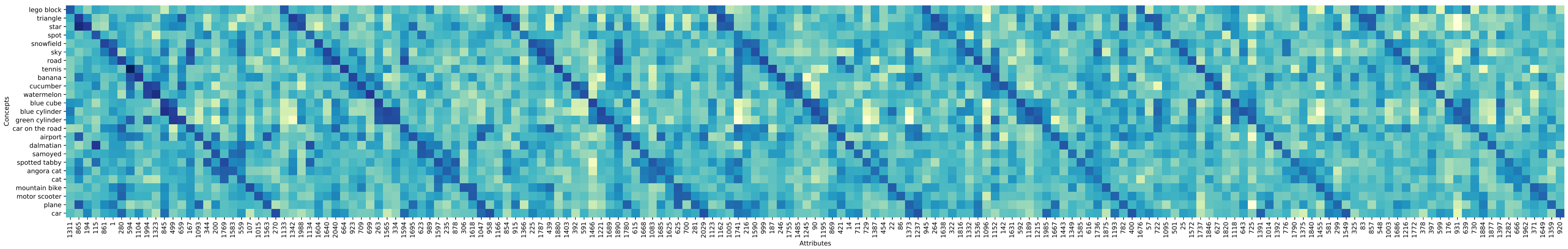
E.2 ADDITIONAL VISUALIZATION RESULTS OF ROA

Additional visualization results of RoA are illustrated in Fig. 10. Most (21 out of 25) concepts are unknown; high saturation indicates high RoA value.

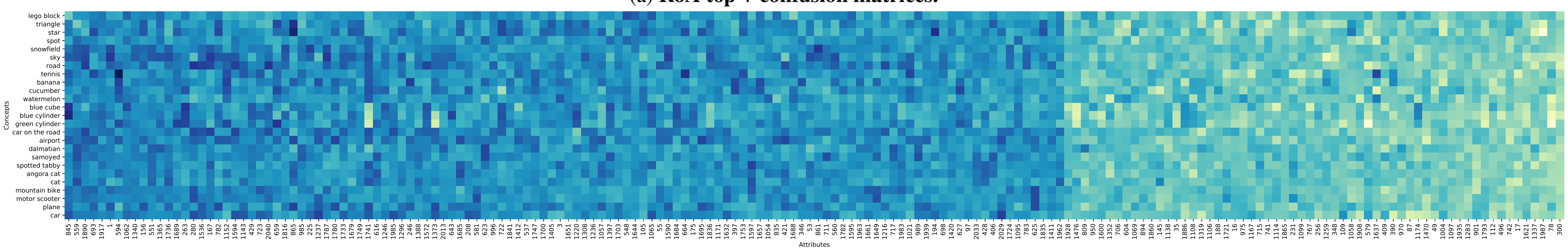
Fig. 10a shows the concatenation of the 7 confusion matrices where the n-th diagonal indicates the n-top RoA of the concepts.

Fig. 10b shows the concatenation of 120 highest (from the left) and 60 lowest (from the right) attributes with the mean of RoA in the context.

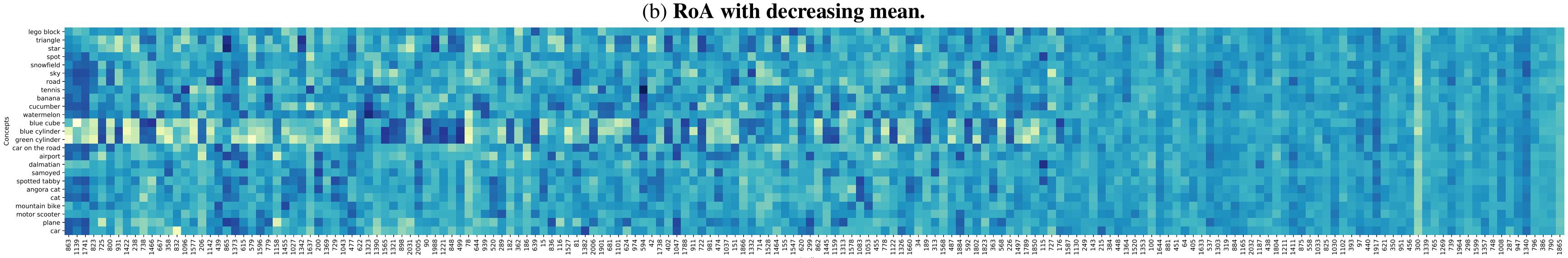
Fig. 10c shows the concatenation of 120 highest (from the left) and 60 lowest (from the right) attributes with the variance of RoA in the context.



(a) **RoA top-7 confusion matrices.**



(b) **RoA with decreasing mean.**



(c) **RoA with decreasing variance.**

Figure 10: **Additional visualizations of RoAs.**