# Not just a matter of semantics: the relationship between visual and semantic similarity

Clemens-Alexander Brust[1][0000−0001−5419−1998] and Joachim Denzler[1,2]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Michael Stifel Center Jena, Germany

**Abstract.** Knowledge transfer, zero-shot learning and semantic image retrieval are methods that aim at improving accuracy by utilizing semantic information, *e.g.,* from WordNet. It is assumed that this information can augment or replace missing visual data in the form of labeled training images because semantic similarity correlates with visual similarity.
This assumption may seem trivial, but is crucial for the application of such semantic methods. Any violation can cause mispredictions. Thus, it is important to examine the visual-semantic relationship for a certain target problem. In this paper, we use five different semantic and visual similarity measures each to thoroughly analyze the relationship without relying too much on any single definition.
We postulate and verify three highly consequential hypotheses on the relationship. Our results show that it indeed exists and that WordNet semantic similarity carries more information about visual similarity than just the knowledge of "different classes look different". They suggest that classification is not the ideal application for semantic methods and that wrong semantic information is much worse than none.

## 1 Introduction

There exist applications in which labeled training data cannot be acquired in sufficient amounts to reach the high accuracy associated with contemporary convolutional neural networks (CNNs) with millions of parameters. These include industrial [14, 18] and medical [15, 27, 31] applications as well as research in other fields like wildlife monitoring [4, 5, 7]. *Semantic methods* such as knowledge transfer and zero-shot learning often process information about the semantic relationship between classes from databases like WordNet [19] to allow high-accuracy classification even when training data is insufficient or missing entirely [24]. *They can only function when the unknown visual relationships between classes are predictable from the given semantic relationships.*

In this paper, we analyze and test this crucial assumption by evaluating the relationship between visual and semantic similarity in a detailed and systematic manner. To guide our analysis, we formulate three highly consequential, non-trivial hypotheses around the visual-semantic relationship. The exact nature of the links and the similarity terms is specified in section 4. Our first hypothesis concerns the relationship itself:

(a) A deer and a forest. By taxonomy only, their semantic similarity is weak. Visual similarity however is strong.

(b) An orchid and a sunflower. Their semantic similarity very strong due to them both being flowers. The visual similarity between them is weak.

**Fig. 1.** Examples of semantic-visual disagreement.

$\mathcal{H}_1$*: There is a link between visual and semantic similarity.* It seems trivial on the surface, but each individual component requires a proper, non-trivial definition to ultimately make the hypothesis verifiable (see section 4). The observed effectiveness of semantic methods suggests that knowledge about semantic relationships is somewhat applicable in the visual domain. However, counter-examples are easily found, *e.g.,* figs. 1 and 4. Furthermore, a crude approximation of semantic similarity is already given by the expectation that "different classes look different" (see section 2.1). A similarity measure based on actual semantic knowledge should be a stronger predictor of visual similarity than this simple baseline.

Semantic methods seek to improve accuracy and in turn reduce model confusion, but the relationship between confusion and visual similarity is non-trivial. Insights about the low-level visual similarity may not apply to the more abstract confusion. To cover not only largely model-free, but also also model-specific notions of visual similarity, we formulate our second and third hypotheses:

$\mathcal{H}_2$*: There is a link between visual similarity and model confusion.* Low inter-class distance in a feature space correlates with confusion, but it could also indicate strong visual similarity. This link depends on the selected features and classifier. It is also affected by violations of "different classes look different" in the dataset.

$\mathcal{H}_3$*: There is a link between semantic similarity and model confusion.* This link should be investigated because it directly relates to the goal of semantic methods, which is to reduce confusion by adding semantic information. It "skips" the low-level visual component and as such is interesting on its own. However, the expectation that "different classes look different" can already explain the complete confusion matrix of a perfect classifier. We also expect it to partly explain a real classifier's confusions. So, to consider $\mathcal{H}_3$ verified, we require semantic similarity to show an even stronger correlation to confusion than given by this expectation.

Our main contribution is an extensive and insightful evaluation of this relationship across five different semantic and visual similarity measures respectively. It is based on the three aforementioned hypotheses around the relationship. We show quantitative results measuring the agreement between individual measures and across visual and semantic similarities as rank correlation. Moreover, we analyze special cases of agreement and disagreement qualitatively. The results and their various implications are discussed in section 5.5. They suggest that, while the relationship exists even beyond the "different classes look different" baseline, tasks different from classification need further research. The semantically reductive nature of class labels suggests that semantic methods may perform better on more complex tasks.

### 1.1   Related Work

The relationship between visual and semantic similarity has been subject of previous investigation. In [6], Deselaers and Ferrari consider a semantic similarity measure by Jiang and Conrath (see section 2.4 and [11]) as well as category histograms, in conjunction with the ImageNet dataset. They propose a novel distance function based on semantic as well as visual similarity to use in a nearest neighbor setting that outperforms purely visual distance functions. The authors also show a positive correlation between visual and semantic similarity for their choice of similarity measures on the ImageNet dataset. Their selections of Jiang-Conrath distance and the GIST feature descriptor are also evaluated in our work, where we add several different methods to compare.

Bilal *et al.* observe the confusion matrix of a convolutional network trained on the ImageNet-1k dataset [26] in [2]. They use visual analytics to show that characteristics of the class hierarchy can be found in the confusion matrix, a result related to our hypothesis $\mathcal{H}_3$.

## 2   Semantic Similarity

The term *semantic similarity* describes the degree to which two concepts interact semantically. A common definition requires taking into account only the taxonomical (hierarchical) relationship between the concepts [9, p. 10]. A more general notion is *semantic relatedness*, where any type of semantic link may be considered [9, p. 10]. Both are *semantic measures*, which also include distances and dissimilarities [9, p. 9]. We adhere to these definitions in this work, specifically the hierarchical restriction of semantic similarity.

### 2.1   Prerequisites

In certain cases, it is easier to formulate a semantic measure based on hierarchical relationships as a distance first. Such a distance $d$ between two concepts $x, y$ can be converted to a similarity by $1/(1+d(x,y))$ [9, p. 60]. This results in a measure bounded by $(0, 1]$, where 1 stands for maximal similarity, *i.e.,* the distance is zero.

We will apply this rule to convert all distances to similarities in our experiments. We also apply it to dissimilarities, which are comparable to distances, but do not fulfill the triangle inequality.

*Semantic Baseline* When training a classifier without using semantic embeddings [1] or hierarchical classification techniques [29], there is still prior information about semantic similarity given by the classification problem itself. Specifically, it is postulated that "classes that are different look different" (see section 4). Machine learning can not work if this assumption is violated, *i.e.,* different classes look identical. We encode this "knowledge" in semantic similarity measure, defined as 1 for two identical concepts and zero otherwise. It will serve as a baseline for comparison with all other similarities.

## 2.2   Graph-based Similarities

We can describe a directed acyclic graph $G(C, \textbf{is-a})$ using the taxonomic relation **is-a** and the set of all concepts $C$. The notions of semantic similarity described in this section can be expressed using properties of $G$. The graph distance $d_G(x, y)$ between two nodes $x, y$, which is defined as the length of the shortest path $xPy$, is an important example. If required, we reduce the graph $G$ to a rooted tree $T$ with root $r$ by iterating through all nodes with multiple ancestors and successively removing the edges to ancestors with the lowest amount of successors. In a tree, we can then define the depth of a concept $x$ as $\mathfrak{d}_T(x) = d_T(r, x)$. This simple, greedy approach make very few cuts in practice and does not strongly affect results.

A simple approach is presented by Rada *et al.* in [22, p. 20], where the semantic distance between two concepts $x$ and $y$ is defined as the graph distance $d_G(x, y)$ between one concept and the other in $G$.

To make similarities comparable between different taxonomies, it may be desirable to take the overall depth of the hierarchy into account. Resnik presents such an approach for trees in [23], considering the maximal depth of $T$ and the *least common ancestor* $\mathcal{L}(x, y)$. $\mathcal{L}$ is the uniquely defined node in the shortest path between two concepts $x$ and $y$ that is an ancestor to both [9, p. 61]. The similarity between $x$ and $y$ is then given as [23, p. 3]:

$$2 \cdot \max_{z \in C} \mathfrak{d}_T(z) - d_T(x, \mathcal{L}(x, y)) - d_T(y, \mathcal{L}(x, y)). \tag{1}$$

## 2.3   Feature-based Similarities

The following approaches use a set-theoretic view of semantics. The set of *features* $\phi(x)$ of a concept $x$ is usually defined as the set of ancestors $A(x)$ of $x$ [9]. We also include $x$ itself, such that $\phi(x) = A(x) \cup \{x\}$ [28].

Inspired by the Jaccard coefficient, Maedche and Staab propose a similarity measure defined as the intersection over union of the concept features of $x$ and $y$ respectively [17, p. 4]. This similarity is bounded by $[0, 1]$, with identical concepts always resulting in 1.

Sanchez *et al.* present a dissimilarity measure that represents the ratio of distinct features to shared features of two concepts. It is defined by [28, p. 7723]:

$$\log_2\Big(1 + \frac{|\phi(x)\backslash\phi(y)| + |\phi(y)\backslash\phi(x)|}{|\phi(x)\backslash\phi(y)| + |\phi(y)\backslash\phi(x)| + |\phi(y)\cap\phi(x)|}\Big). \tag{2}$$

### 2.4 Information-based Similarities

Semantic similarity is also defined using the notion of informativeness of a concept, inspired by information theory. Each concept $x$ is assigned an *Information Content* (IC) $\mathcal{I}(x)$ [23, 25]. This can be defined using only properties of the taxonomy, *i.e.,* the graph $G$ (intrinsic IC), or using the probability of observing the concept in corpora (extrinsic IC) [9, p. 54].

We use an intrinsic definition presented by Zhou *et al.* in [36], based on the descendants $D(x)$:

$$\mathcal{I}(x) = k\cdot\Big(1 - \frac{|D(x)|}{|C|}\Big) + (1-k)\cdot\Big(\frac{\log(\mathfrak{d}_T(x))}{\log(\max_{z\in C}\mathfrak{d}_T(z))}\Big). \tag{3}$$

With a definition of IC, we can apply an information-based similarity measure. Jiang and Conrath propose a semantic distance in [11] using the notion of *Most Informative Common Ancestor* $\mathfrak{M}(x,y)$ of two concepts $x, y$. It is defined as the element in $(A(x)\cap A(y))\cup(x\cap y)$ with the highest IC [9, p. 65]. The distance is then defined as [11, p. 8]:

$$\mathcal{I}(x) + \mathcal{I}(y) - 2\cdot\mathcal{I}(\mathfrak{M}(x,y)). \tag{4}$$

## 3 Visual Similarity

Assessing the similarity of images is not a trivial task, mostly because the term "similarity" can be defined in many different ways. In this section, we look at two common interpretations of visual similarity, namely perceptual metrics and feature-based similarity measures.

### 3.1 Perceptual Metrics

Perceptual metrics are usually employed to quantify the distortion or information loss incurred by using compression algorithms. Such methods aim to minimize the difference between the original image and the compressed image and thereby maximize the similarity between both. However, perceptual metrics can also be used to assess the similarity of two independent images.

An image can be represented by an element of a high-dimensional vector space. In this case, the Euclidean distance is a natural candidate for a dissimilarity measure. With the rule $1/(1 + d)$ from section 2.1, the distance is transformed into a visual similarity measure. To normalize the measure *w.r.t.* image dimensions and to simplify calculations, the mean squared error (MSE)

is used. Applying the MSE to estimate image similarity has shortcomings. For example, shifting an image by one pixel significantly changes the distances to other images, and even its unshifted self. An alternative, but related measure is the mean absolute difference (MAD), which we also consider in our experiments.

In [34], Wang *et al.* develop a perceptual metric called *Structural Similarity Index* to adress shortcomings of previous methods. Specifically, they consider properties of the human visual system such that the index better reflects human judgement of visual similarity.

We use MSE, MAD and SSIM as perceptual metrics to indicate visual similarity in our experiments. There are better performing methods when considering human judgement, *e.g.,* [35]. However, we cannot guarantee that humans always treat visuals and semantics as separate. Therefore, we avoid further methods that are motivated by human properties [33, 3] or already incorporate semantic knowledge [16, 8].

### 3.2   Feature-based Measures

Features are extracted to represent images at an abstract level. Thus, distances in such a feature space of images correspond to visual similarity in a possibly more robust way than the aforementioned perceptual metrics. Features have inherent or learned invariances *w.r.t.* certain transformations that should not affect the notion of visual similarity strongly. However, learned features may also be invariant to transformations that do affect visual similarity because they are optimized for semantic distinction. This behavior needs to be considered when selecting abstract features to determine visual similarity.

GIST [21] is an image descriptor that aims at describing a whole scene using a small number of estimations of specific perceptual properties, such that similar content is close in the resulting feature space. It is based on the notion of a *spatial envelope*, inspired by architecture, that can be extracted from an image and used to calculate statistics.

For reference, we observe the confusions of five ResNet-32 [10] models to represent feature-based visual similarity on the highest level of abstraction. Because confusion is not a symmetric function, we apply a transform $(M + M^T)/2$ to obtain a symmetric representation of the confusion matrix.

## 4   Evaluating the Relationship

Visual and semantic similarity are measures defined on different domains. Semantic similarities compare concepts and visual similarities compare individual images. To analyze a correlation, a common domain over which both can be evaluated is essential. We propose to calculate similarities over all pairs of classes in an image classification dataset, which can be defined for both visual and semantic similarities. These pairwise similarities are then tested for correlation. The process is clarified in the following:

1. **Dataset.** We use the CIFAR-100 dataset [13] to verify our hypotheses. This dataset has a scale at which all experiments take a reasonable amount of time. Our computation times grow quadratically with the number of classes as well as images. Hence, we do not consider ImageNet [26] or 80 million tiny images [32] despite their larger coverage of semantic concepts.

2. **Semantic similarities.** We calculate semantic similarity measures over all pairs of classes in the dataset. The taxonomic relation **is-a** is taken from Word-Net [19] by mapping all classes in CIFAR-100 to their counterpart concepts in WordNet, inducing the graph $G(C, \textbf{is-a})$. Some measures are defined as distances or dissimilarities. We use the rule presented in section 2.1 to derive similarities. The following measures are evaluated over all pairs of concepts $(x, y) \in C \times C$ (see section 2):

(S1) Graph distance $d_G(x, y)$ as proposed by Rada *et al.*, see [22, p. 20].

(S2) Resnik's maximum depth bounded similarity, see eq. (1) and [23, p. 3].

(S3) Maedche and Staab similarity based on intersection over union of concept features [17, p. 4].

(S4) Dissimilarity proposed by Sanchez *et al.* using distinct to shared features ratio, see eq. (2) and [28, p. 7723].

(S5) Jiang and Conrath's distance [11, p. 8], eq. (4), using intrinsic Information Content from [36], see eq. (3).

3. **Visual similarities.** To estimate a visual similarity between two classes $x$ and $y$, we calculate the similarity of each test image of class $x$ with each test image of class $y$ and use the average as an estimate. Again we apply the rule from section 2.1 for distances and dissimilarities. The process of comparing all images from one class to all from another is performed for the following measures (see section 3):

(V1) The mean squared error (MSE) between two images.

(V2) The mean absolute difference (MAD) between two images.

(V3) Structural Similarity Index (SSIM), see [34].

(V4) Distance between GIST descriptors [21] of images in feature space.

(V5) Observed symmetric confusions of five ResNet-32 [10] models trained on the CIFAR-100 training set.

4. **Aggregation.** For both visual and semantic similarity, there is more than one candidate method, *i.e.,* (S1)-(S5) and (V1)-(V5). For the following steps, we need a single measure for each type of similarity, which we aggregate from (S1)-(S5) and (V1)-(V5) respectively. Since each method has its merits, selecting only one each would not be representative of the type of similarity. The output of all candidate methods is normalized individually, such that its range is in $[0, 1]$. We then calculate the average over each type of similarity, *i.e.,* visual and semantic, to obtain two distinct measures (S) and (V).

5. **Baselines.** A basic assumption of machine learning is that "the domains occupied by features of different classes are separated" [20, p. 8]. Intuitively, this should apply to the images of different classes as well. We can then expect to predict at least some of the visual similarity between classes just by knowing whether the classes are identical or not. This knowledge is encoded in the *semantic baseline* (SB), defined as 1 for identical concepts and zero otherwise (see
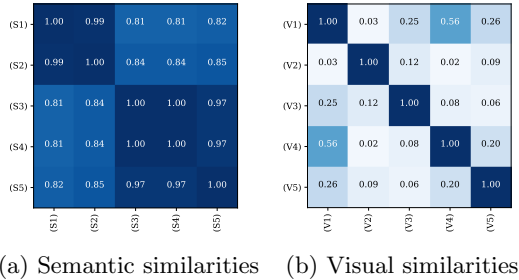
(a) Semantic similarities    (b) Visual similarities

**Fig. 2.** Rank correlation coefficient between different similarities, grouped by semantic and visual. $p < 0.05$ for all correlations.

also section 2.1). We propose a second baseline, the *semantic noise* (SN), where the aforementioned pairwise semantic similarity (S) is calculated, but the concepts are permuted randomly. This baseline serves to assess the informativeness of the taxonomic relationships.

6. **Rank Correlation** The similarity measures mentioned above are useful to define an order of similarity, *i.e.,* whether a concept $x$ is more similar to $z$ than concept $y$. However, it is not reasonable in all cases to interpret them in a linear fashion, especially since many are derived from distances or dissimilarities and all were normalized from different ranges of values and then aggregated. We therefore test the similarities for correlation *w.r.t.* ranking, using Spearman's rank correlation coefficient [30] instead of looking for a linear relationship.

## 5    Results

In the following, we present the results of our experiments defined in the previous section. We first examine both types of similarity individually, comparing the five candidate methods each. Afterwards, the hypotheses proposed in section 1 are tested. We then investigate cases of (dis-)agreement between both types of similarity.

### 5.1    Semantic Similarities

We first analyze the pairwise semantic similarities over all classes. Although we consider semantic similarity to be a single measure when verifying our hypotheses, studying the correlation between our candidate methods (S1)-(S5) is also important. While of course affected by our selection, it reflects upon the degree of agreement between several experts in the domain. Figure 2a visualizes the correlations. The graph-based methods (S1) and (S2) agree more strongly with each other than with the rest. The same is true of feature-based methods (S3) and (S4), which show the strongest correlation. The *inter-agreement R*, calculated by taking the average of all correlations except for the main diagonal, is **0.89**.

This is a strong agreement and suggests that the order of similarity between concepts can be, for the most part, considered representative of a universally agreed upon definition (if one existed). At the same time, one needs to consider that all methods utilize the same WordNet hierarchy.

*Baselines* Our semantic baseline (SB, see section 4) encodes the basic knowledge that different classes look different. This property should also be fulfilled by the average semantic similarity (S, see section 4). We thus expect there to be at least some correlation. The rank correlation between our average semantic similarity (S) and the semantic baseline (SB) is 0.17 with $p < 0.05$. This is a weak correlation compared to the strong inter-agreement of 0.89, which suggests that the similarities (S1)-(S5) are vastly more complex than (SB), but at the same time have a lot in common. As a second baseline we test the semantic noise (SN, see section 4). It is not correlated with (S) at $\rho = 0.01, p > 0.05$, meaning that the taxonomic relationship strongly affects (S). If it did not, the labels could be permuted without changing the pairwise similarities.

### 5.2    Visual Similarities

Intuitively, visual similarity is a concept that is hard to define clearly and uniquely. Because we selected very different approaches with very different ideas and motivations behind them, we expect the agreement between (V1)-(V5) to be weak. Figure 2b shows the rank correlations between each candidate method. The agreement is strongest between the mean squared error (V1) and the GIST feature distance (V4). Both are L2 distances, but calculated in separate domains, highlighting the strong nonlinearity and complexity of image descriptors. The inter-agreement is very weak at $R = \mathbf{0.17}$. The results confirms our intuitions that visual similarity is very hard to define in mathematical terms. There is also no body of knowledge that all methods use in the visual domain like WordNet provides for semantics.

### 5.3    Hypotheses

To give a brief overview, the rank correlations between the different components of $\mathcal{H}_1$-$\mathcal{H}_3$ are shown in fig. 3. In the following, we give our results *w.r.t.* the individual hypotheses. They are discussed further in section 5.5.

$\mathcal{H}_1$: *There is a link between visual and semantic similarity.* Using the definitions from section 4 including the semantic baseline (SB), we can examine the respective correlations. The rank correlation between (V) and (S) is $\mathbf{0.23}$, $p < 0.05$, indicating a link. Before we consider the hypothesis verified, we also evaluate what fraction of (V) is already explained by the semantic baseline (SB) as per our condition given in section 4. The rank correlation between (V) and (SB) is 0.17, $p < 0.05$, which is a weaker link than between (V) and (S). Additionally, (V) and (SN) are not correlated, illustrating that the wrong semantic knowledge can be worse than none. Thus, we can **verify** $\mathcal{H}_1$.

|       | (V)      | (S)     | (SN)  | (SB)  |
|-------|----------|---------|-------|-------|
| (V)   | 0.17     | 0.23    | (-0.01) | 0.17  |
| (S)   | 0.23     | 0.89    | (0.01)  | 0.17  |
| (SN)  | (-0.01)  | (0.01)  | 1.00  | 0.17  |
| (SB)  | 0.17     | 0.17    | 0.17  | 1.00  |

**Fig. 3.** Rank correlation coefficient between different types of similarities, grouped by semantic and visual. $p < 0.05$ except for numbers in parentheses. Main diagonal represents inter-agreement $R$. Similarities: (V) – visual, (S) – semantic, (SN) – semantic noise, (SB) – semantic baseline.

$\mathcal{H}_2$: *There is a link between visual similarity and model confusion.* Since model confusion as (V5) is a contributor to average visual similarity (V), we consider only (V-), comprised of (V1)-(V4) for this hypothesis. The rank correlation between (V-) and the symmetric model confusion (V5) is **0.21**, $p < 0.05$. Consequently, $\mathcal{H}_2$ is also **verified**.

$\mathcal{H}_3$: *There is a link between semantic similarity and model confusion.* Here we evaluate the relationship between (S) and the symmetric confusion matrix (V5) as defined in section 4. (S) should offer more information about where confusions occur than the baseline (SB) to consider $\mathcal{H}_3$ verified. The rank correlation between (V5) and (S) is **0.39**, $p < 0.05$, while (V5) and (SB) are only correlated at $\rho = 0.21$, $p < 0.05$, meaning that $\mathcal{H}_3$ is **verified**, too.

See section 5.5 for a discussion of possible consequences.

### 5.4   Agreement and Disagreement

To further analyze the the correlation, we examine specific cases of very strong agreement or disagreement. Figure 4 shows these extreme cases. We determine agreement based on ranking, so the most strongly agreed upon pairs (see fig. 4a) still show different absolute similarity numbers. Interestingly, they are not cases of extreme similarities. It suggests that even weak disagreements are more likely to be found at similarities close to the boundaries. When investigating strong disagreement as shown in fig. 4b, there are naturally extreme values to be found. All three pairs involve `forest.n.01`, which was also a part of the second least semantically similar pair. Its partners are all animals which usually have a background visually similar to a forest, hence the strong disagreement. However, the low semantic similarity is possibly an artifact of reducing a whole image to a single concept.
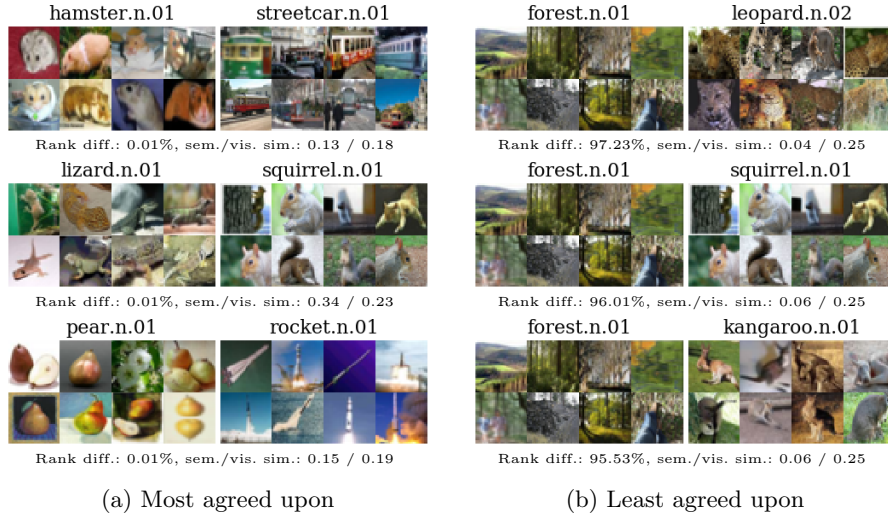
| hamster.n.01 | streetcar.n.01 |
| --- | --- |

Rank diff.: 0.01%, sem./vis. sim.: 0.13 / 0.18

| lizard.n.01 | squirrel.n.01 |
| --- | --- |

Rank diff.: 0.01%, sem./vis. sim.: 0.34 / 0.23

| pear.n.01 | rocket.n.01 |
| --- | --- |

Rank diff.: 0.01%, sem./vis. sim.: 0.15 / 0.19

(a) Most agreed upon

| forest.n.01 | leopard.n.02 |
| --- | --- |

Rank diff.: 97.23%, sem./vis. sim.: 0.04 / 0.25

| forest.n.01 | squirrel.n.01 |
| --- | --- |

Rank diff.: 96.01%, sem./vis. sim.: 0.06 / 0.25

| forest.n.01 | kangaroo.n.01 |
| --- | --- |

Rank diff.: 95.53%, sem./vis. sim.: 0.06 / 0.25

(b) Least agreed upon

**Fig. 4.** CIFAR-100 classes selected by highest and lowest ranking agreement between visual and semantic similarity measures as defined in section 4.

### 5.5 Discussion

$\mathcal{H}_1$: *There is a link between visual and semantic similarity.* The relationship is stronger than a simple baseline, but weak overall at $\rho = 0.23$ vs $\rho = 0.17$. This should be considered when employing methods where visuals and semantics interact, *e.g.,* in knowledge transfer. Failure cases such as in fig. 4b can only be found when labels are known, which has implications for real-life applications of semantic methods. As labels are unknown or lack visual examples, such cases are not predictable beforehand. This poses problems for applications that rely on accurate classification such as safety-critical equipment or even research in other fields consuming model predictions. A real-world example is wildlife conservationists relying on statistics from automatic camera trap image classification to draw conclusions on biodiversity. That the semantic similarity of randomly permuted classes is not correlated with visual similarity at all, while the baseline is, suggests that wrong semantic knowledge can be much worse than no knowledge.

$\mathcal{H}_2$: *There is a link between visual similarity and model confusion.* Visual similarity is defined on a low level for $\mathcal{H}_2$. As such, it should not cause model confusion by itself. On the one hand, the model can fail to generalize and cause an avoidable confusion. On the other hand, there may be an issue with the dataset. The test set may be sampled from a different distribution than the training set. It may also violate the postulate that different classes look different by containing the same or similar images across classes.

$\mathcal{H}_3$: *There is a link between semantic similarity and model confusion.* Similar to $\mathcal{H}_1$, it suggests that semantic methods could be applied to our data, but

maybe not in general because failure cases are unpredictable. However, it implies a stronger effectiveness than $\mathcal{H}_1$ at $\rho = 0.39$ vs. the baseline at $\rho = 0.21$. We attribute this to the model's capability of abstraction. It aligns with the idea of taxonomy, which is based on repeated abstraction of concepts. Using a formulation that optimizes semantic similarity instead of cross-entropy (which would correspond to the semantic baseline) could help in our situation. It may still not generalize to other settings and any real-world application of such methods should be verified with at least a small test set.

*Qualitative* Some failures or disagreements may not be a result of the relationship itself, but of its application to image classification. The example from fig. 1 is valid when the whole image is reduced to a single concept. Still, the agreement between visual and semantic similarity may increase when the image is described in a more holistic fashion. While "deer" and "forest" as nouns are taxonomically only loosely related, the descriptions "A deer standing in a forest, partially occluded by a tree and tall grass" and "A forest composed of many trees and bushes, with the daytime sky visible" already appear more similar, suggesting that more complex tasks stand to benefit more from semantic methods.

## 6   Conclusion

We present results of a comprehensive evaluation of semantic similarity measures and their correlation with visual similarities. We measure against the simple prior knowledge of different classes having different visuals. Then, we show that the relationship between semantic similarity, as calculated from WordNet [19] using five different methods, and visual similarity, also represented by five measures, is more meaningful than that. Furthermore, inter-agreement measures suggest that semantic similarity has a more agreed upon definition than visual similarity, although both concepts are based on human perception.

The results indicate that further research, especially into tasks different from image classification is warranted because of the semantically reductive nature of image labels. It may restrict the performance of semantic methods unneccessarily. It is likely that the relationship between semantic and visual similarity is much stronger when the semantics of an image are better approximated.

Further work should focus on the experimental setup. Viable alternatives to the simple visual similarity measures that we currently use are ones based on image gradients as well as cross-correlation. Larger datasets with well established semantic groundings such as ImageNet could be used together with a stochastic approach to approximating similarity to make them computationally feasible. Additonally, not only nouns should be considered, but also adjectives, decompositions of objects into parts. Datasets like Visual Genome [12] offer more complex annotations mapped to WordNet concepts.

# References

1. Barz, B., Denzler, J.: Hierarchy-based image embeddings for semantic image retrieval. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 638–647. IEEE (2019)
2. Bilal, A., Jourabloo, A., Ye, M., Liu, X., Ren, L.: Do convolutional neural networks learn class hierarchy? **24**(1), 152–162. https://doi.org/10.1109/TVCG.2017.2744683
3. Van den Branden Lambrecht, C.J., Verscheure, O.: Perceptual quality measure using a spatiotemporal model of the human visual system. In: Digital Video Compression: Algorithms and Technologies 1996. vol. 2668, pp. 450–462. International Society for Optics and Photonics (1996)
4. Brust, C.A., Burghardt, T., Groenenberg, M., Kding, C., Khl, H., Manguette, M., Denzler, J.: Towards automated visual monitoring of individual gorillas in the wild. In: International Conference on Computer Vision Workshop (ICCV-WS) (2017)
5. Chen, G., Han, T.X., He, Z., Kays, R., Forrester, T.: Deep convolutional neural network based species recognition for wild animal monitoring. In: Image Processing (ICIP), 2014 IEEE International Conference on. pp. 858–862. IEEE (2014)
6. Deselaers, T., Ferrari, V.: Visual and semantic similarity in imagenet. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 1777–1784. IEEE (2011)
7. Freytag, A., Rodner, E., Simon, M., Loos, A., Khl, H., Denzler, J.: Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In: German Conference on Pattern Recognition (GCPR). pp. 51–63 (2016)
8. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 2121–2129. Curran Associates, Inc. (2013)
9. Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: Semantic similarity from natural language and ontology analysis. Synthesis Lectures on Human Language Technologies **8**(1), 1–254 (2015)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (CVPR) (2016)
11. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: arXiv preprint arXiv:cmp-lg/9709008 (1997)
12. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) **123**(1), 32–73 (2017)
13. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 4, University of Toronto (2009)
14. Kumar, A.: Computer-vision-based fabric defect detection: A survey. IEEE transactions on industrial electronics **55**(1), 348–363 (2008)
15. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
16. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern recognition **40**(1), 262–282 (2007)
17. Maedche, A., Staab, S.: Comparing ontologies-similarity measures and a comparison study. Tech. rep., Institute AIFB, University of Karlsruhe (2001)

18. Malamas, E.N., Petrakis, E.G., Zervakis, M., Petit, L., Legat, J.D.: A survey on industrial vision systems, applications and tools. Image and vision computing **21**(2), 171–188 (2003)
19. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM **38**(11), 39–41 (Nov 1995). https://doi.org/10.1145/219717.219748
20. Niemann, H.: Pattern Analysis. Springer Series in Information Sciences, Springer Berlin Heidelberg (2012), https://books.google.de/books?id=mdOoCAAAQBAJ
21. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision **42**(3), 145–175 (2001)
22. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE transactions on systems, man, and cybernetics **19**(1), 17–30 (1989)
23. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: arXiv preprint arVix:cmp-lg/9511007 (1995)
24. Rohrbach, M., Stark, M., Schiele, B.: Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 1641–1648. IEEE (2011)
25. Ross, S.M.: A first course in probability. Macmillan (1976)
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)
27. Salem, M.A.M., Atef, A., Salah, A., Shams, M.: Recent survey on medical image segmentation. In: Computer Vision: Concepts, Methodologies, Tools, and Applications, pp. 129–169. IGI Global (2018)
28. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: A new feature-based approach. Expert systems with applications **39**(9), 7718–7728 (2012)
29. Silla, C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery **22**(1), 31–72 (Jan 2011)
30. Spearman, C.: The proof and measurement of association between two things. The American journal of psychology **15**(1), 72–101 (1904)
31. Thevenot, J., López, M.B., Hadid, A.: A survey on computer vision for assistive medical diagnosis from faces. IEEE journal of biomedical and health informatics **22**(5), 1497–1511 (2018)
32. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. Transactions on Pattern Analysis and Machine Intelligence (PAMI) **30**(11), 1958–1970 (2008)
33. Tversky, A.: Features of similarity. Psychological review **84**(4),  327 (1977)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (April 2004)
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: arXiv preprint arXiv:1801.03924
36. Zhou, Z., Wang, Y., Gu, J.: A new model of information content for semantic similarity in wordnet. In: Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on. vol. 3, pp. 85–89. IEEE (2008)