How CNNs and ViTs perceive similarities between categories

Katarzyna Filus[0000−0003−1303−9230] and Joanna Domańska[0000−0002−1935−8358]

Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Bałtycka 5, Gliwice 44-100, Poland kfilus@iitis.pl

Abstract. Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) trained for supervised tasks are the leading networks used in practical computer vision. Despite using different techniques, they both perfect their object recognition skills. In this race, it is overall accuracy that matters at most. But is it enough? Should not we care about the correct perception of inter-class similarities? We believe we should, as similarity is a fundamental aspect of categorization and the structure of the world is highly correlated. Models should reasonably assess similarities for more nuanced perception, and we should examine it for more transparency and trust. That is why, we analyzed what state-of-the-art object recognition networks perceive as similar. We proposed a framework to visually and numerically examine and compare the perception of different trained models. We used it to answer a series of similarity-related questions based on experiments on a large population of 42 models.

Keywords: Explainability · Computer Vision · Deep Learning · Supervised Learning · Semantic Similarity

1 Introduction

Is a Poodle similar to a Husky? Are sharks and scuba divers related? Answering such questions is a standard human ability. In cognitive psychology, different concepts are named semantic units [7]. Relations between them are called semantic relations with a narrower group - semantic similarities. Goldstone and Son stated "assessments of similarity are fundamental to cognition because similarities in the world are revealing. The world is an orderly enough place that similar objects and events tend to behave (or look - our postscript) similarly" [11], while Rosch et al. noted that real-world objects exhibit highly correlational structure [31]. Also, maximum information with least cognitive effort is obtained when categories map the world structure as closely as possible allowing to optimally use the finite resources [31]. Therefore, natural correlations should be reflected by robust and accurate categorization systems [31], such as deep vision networks. For a more human-like and robust categorization, computer vision algorithms should not only differentiate objects, but also reasonably structure them, especially that visual and semantic similarities are often related [11].

Correct similarity assessment is also important for improving explainability and trust in Artificial Intelligence, as well as ongoing discussions and efforts of standardization organizations (e.g. European Telecommunications Standards Institute). Showing people that deep models perceive similarity reasonably and not that far from how they do it (expressed via human-created semantic relations), would be somehow comforting. Moreover, humans and computer vision algorithms tend to make mistakes mostly among categories they perceive similar [1, 5], so models with more reasonable perception could also return more reasonable errors, which would be easier for us to understand and even accept [5]. To do this, computer vision researchers try to force networks to reflect human similarity judgments [27]. However, the current rush for new learning approaches practically ignores examination of how modern models trained in a supervised manner (without enforcement of similarity judgments) perceive the world structure, while these are dominant models in real-life vision systems. Some limited works considered this aspect for early Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) being underexplored. Therefore, only now, with heterogeneous CNN, ViT and hybrid models, we are finally able to build representative populations of networks and perform a proper examination.

Motivated by this literature deficiency and possibility, we propose the framework with a core metric - Semantic Similarity Alignment Degree (SSAD). We aim to enable systematic analysis of network similarity perception and comparisons between different networks (also with another metric: Network Similarity Alignment Degree, NSAD). The key feature of our methods is that they do not require any images, and thus offer efficiency. We performed extensive empirical analysis and delivered thorough findings for the most common vision benchmark - ImageNet and object recognition. We examined how 42 state-of-the-art networks perceive inter-class similarities and answered the questions: **(1)** Is similarity perceived by ImageNet-trained CNNs and ViTs related to semantic similarity? (Sec. 4.1); **(2)** Is there a relationship between the networks' ability to align their similarity perception to semantic similarity with their size and ImageNet accuracy? (Sec. 4.2); **(3)** Do networks perceive other semantic relations besides similarity? Which ones? (Sec. 4.3); **(4)** Do different networks share similarity perception? (Sec. 4.4). We provide our implementation to enable future research at `https://github.com/kafilus/DeepNetworksSimilarity`.

2 Related work

As humans possess remarkable ability to categorize objects and assess their similarity, it also became important in computer vision [9, 10, 27, 28]. Researchers focused on the relation between visual and semantic similarities [7] and other types of similarity [30]. They also noticed that CNN error patterns show some kind of hierarchy [1, 5, 17, 26]. Although stimuli-based analysis [19] with templates/confusion matrices can reveal some approximate inter-class relations, it is computationally-intensive and its studies can very likely lead to blind alleys [2]. The alternative is to use class templates [8, 26]. Similarity in the deep learning

domain nowadays is used usually to create new learning approaches [3]. In the advanced schemes, human similarity judgments are used as a reference to align the neural representations [27]. While this is an important path, it would also be important to finally and thoroughly examine how modern networks trained on traditional class label prediction without any similarity perception enforcement perceive similarity (the purpose of our study), because works such as [27] rely on incomplete and potentially partially outdated studies on how networks perceive similarity (e.g. [1]), as they were conducted for early CNNs (homogeneous, with much lower accuracy than the current models) and ViTs remain underexplored. Moreover, attention should be put on other semantic relations than semantic similarity to better understand how networks perceive similarity. Standard lexical terms should be used to systematize the relations, because they can provide a consistent and shared vocabulary for describing similarity sources. To the best of our knowledge, no work numerically compares the similarity perception of networks with semantic similarity for such a large set of models. Because real-world objects exhibit high correlational structure, and thus semantic and visual similarities often coexist [31], it is a large literature deficiency, and a thorough analysis is vital with recent works highlighting this necessity [16, 27].

3 Methods

In this section, we introduce our framework by formulating all the necessary data structures and metrics.

Semantic similarity and relations (our reference). Semantic similarity is a relation between items with a similar meaning [18], and is one of semantic relations [18]. While semantic similarity is limited to synonymy, hyponymy, and hypernymy (is-a relation, e.g. a dog is a hypernym of a Poodle, a Poodle is a hyponym of a dog), semantically related concepts can be semantically dissimilar concepts connected by any type of relation, such as **meronymy** (A is part of B), **function** (A is used to perform B), spatial relations, e.g. **proximity** (A is near B) or **containment** (A is within B) etc. To measure semantic similarity via WordNet [25] one can use methods such as path length (path) [29] or Leacock and Chodorow [20]. They are based on path lengths between concepts or information content of their least common subsumer [29]. As path similarity outperforms the majority of other measures by a large margin in terms of correlation with human judgment of semantic relatedness [18], we use it in our study. The advantages of using this and other WordNet-based measures as a similarity perception reference is its clear formulation, consistent similarity scores due to derivation from a fixed and comprehensive lexical database. Also, compared to human judgments it does not require large-scale polls, already covers more than 150k concepts and was created via objective and systematic approach to defining word relationships. By computing the pair-wise similarities between all WordNet nodes in ImageNet-1k (classes), we obtain the **WordNet Class Similarity Matrix (WNCSM)**. This matrix is used in our analysis as a semantic similarity perception reference. See its visualization in Fig. 1.

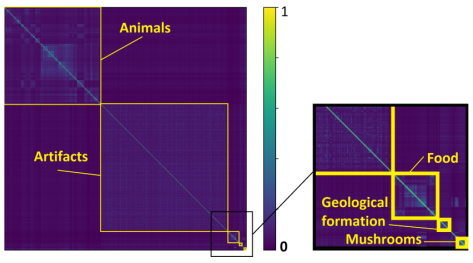
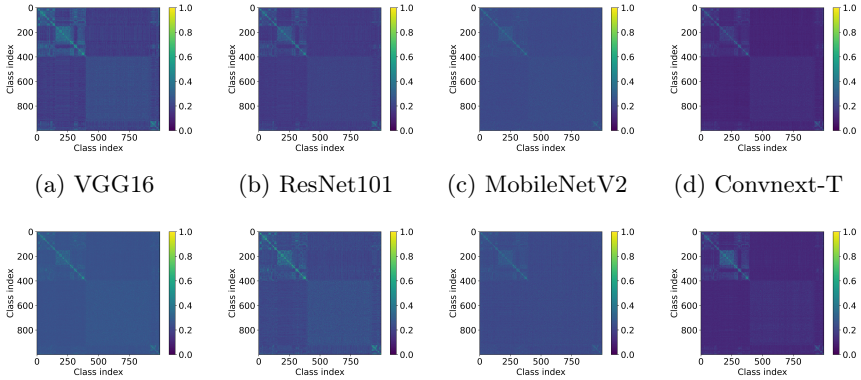


Fig. 1. WNCSM for ImageNet.

Network Class Similarity Matrix is computed based on the similarity of weights in the final classifier of a deep learning model [8, 10, 26, 28]. It is an image-free alternative to using confusion matrices/extracted features to approximate similarity. Each neuron c of the classification layer corresponds to one of the considered classes. Weights connecting this neuron to the neurons in the penultimate layer can be treated as a class template (vector representation) of class c [26, 28] (we attach a graphical representation of this method in our repository). We denote them as w_c with elements w_{ci} corresponding to the weight connecting the neuron representing the c-th class to the i-th neuron in the previous layer. The dimensionality of w_c matches the number of neurons in the penultimate layer and encapsulates the learned representation of class c in the feature space defined by this layer. To compute the similarity between templates of two classes (k and l), cosine similarity (CS) is used: $\mathrm{CS}(k,l) = \frac{w_k^T w_l}{||w_k||||w_l||}$ [26,28]. Computing the similarities between all classes, results in the **Network Class Similarity Matrix (NCSM)** for each examined network (each element in the k-th row and l-th column in this matrix takes value $CS(k,l)$). It can be used for visual (structural) comparison with WordNet Class Similarity Matrix (WNCSM). It can also be utilized for numerical comparison with WNCSM (which results in **Semantic Similarity Alignment Degree - SSAD** - described in the next section). After sorting each row, it can be used to manually inspect which pairs of classes the examined network perceives as the most similar (see Tab. 2 with 5 most similar classes for example classes and networks). While perfectly, a few human subjects would manually evaluate the closest neighborhood of each class in the sorted Network Class Similarity Matrix to examine the similarities, such an approach is not practical as it requires the analysis of $N^2 - N$ class pairs for the dataset with N classes. As an alternative, we propose to analyze a structure we named **Closest Neighbor Pair Ranking (CNPR)**. It is generated by sorting the pairs of the closest neighbors (1st two elements of sorted CSMs) via their similarities values (from the most similar to the least similar pairs): $CNPR = \mathrm{sort}(k,l)\,(\max_{l \neq k} CS(k,l))$. Now, we can manually analyze all or top K class pairs, which is the head of the CNPR (we use $K = 50$ in our manual experiments, which in the case of ImageNet reduces the number of pairs more than 99.9%).

Semantic & Network Similarity Alignment. Semantic Similarity Alignment Degree (SSAD) is a measure that computes to what degree the network perception of similarity and semantic similarity are related. To do this, the correlation or similarity is computed between NCSM and WNCSM. For Cosine similarity it becomes $SSAD_{Cosine} = \mathrm{CS}(NCSM, WNCSM)$, for Spearman Correlation - $SSAD_{Spearman} = \rho(NCSM, WNCSM)$, and Kendall - $SSAD_{Kendall} =$

$\tau(NCSM, WNCSM)$. These measures can be used to examine networks separately or to compare them (higher values imply that the network similarity perception lies closer to the semantic similarity - these perceptions are better aligned). Values of $SSAD$ can be used to examine the populations of networks, e.g. how the degree of alignment is related to the accuracy on ImageNet or network size (measured as Pearson/Spearman/Kendall correlation between accuracy/size and $SSAD_{Cosine}$, $SSAD_{Spearman}$ and $SSAD_{Kendall}$). Also, a similar measure, but taking as arguments 2 Network Class Similarity Matrices can be defined and named **Network Similarity Alignment Degree (NSAD)** to enable comparisons between networks. For Cosine similarity it becomes $NSAD_{Cosine} = \text{CS}(NCSM_1, NCSM_2)$, $NSAD_{Spearman} = \rho(NCSM_1, NCSM_2)$ for Spearman, and $NSAD_{Kendall} = \tau(NCSM_1, NCSM_2)$ for Kendall Correlation. $NCSM_1$ and $NCSM_2$ denote NCSMs of network 1 and 2 used for comparison.



(a) VGG16 (b) ResNet101 (c) MobileNetV2 (d) Convnext-T

(e) CVT-21 (f) LeViT-192 (g) MobileViT-S (h) Swin-B

Fig. 2. Example Network Class Similarity Matrices (NCSMs). All networks perceive similarity in a similar manner, which is exhibited by NCSMs' close structure.

4 Experiments

In our experiments, we examine ImageNet-1k [32] models for object recognition due to it being the most important vision benchmark and due to the suitability of ImageNet [6] to study semantic relations, as it was created based on the semantic hierarchy of WordNet [25]. ImageNet-1K, offers a uniform categorization (leaf-level categories only) ideal for studying how vision networks represent complex information hierarchies. When it comes to the ImageNet-trained models, the last decade brought colossal changes, and we are finally able to create network populations that are diverse enough to properly examine them from the perspective of similarity perception. We build the network population (42 networks) with CNNs (24) and ViTs (18) - see all networks in Tab. 1. We perform the experiments listed below on PC with AMD Ryzen 7 5800X3D to awnser the questions stated in the introduction (no GPU needed) and 64GB RAM:

1. We generate the NCSMs for all 42 networks and visually compare them with the WNCSM.
2. We compare numerically NCSMs with the WNCSM via SSAD and manually on the basis of their structure.
3. We measure the correlation (Pearson, Spearman, Kendall) between SSAD and network size and ImageNet accuracy (for the whole population and for CNNs and ViTs separately).
4. We manually search in the CNPRs for other semantic relations that cause the perceived similarity (homophony, hypernymy, hyponymy, synonymy, sister terms, meronymy, holonymy, containment, physical proximity).
5. We examine to which extent networks share similarity perception via NSAD and histograms of correlations between all NCSMs.

4.1 Is similarity perceived by networks related to semantic similarity?

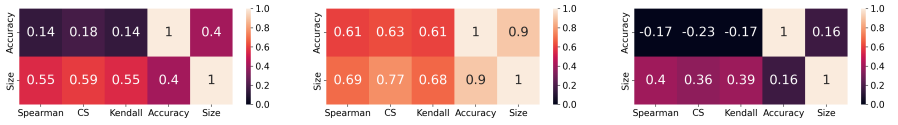
Table 1. $SSAD_{Cosine/Spearman/Kendall}$ values sorted by $SSAD_{Cosine}$. It can be observed that smaller models generally achieve lower SSAD values than larger ones.

Position	Name	Cosine ↑	Spearman	Kendall
1	MobileViT-small [24]	0.818	0.079	0.055
2	MobileViT-xx-small [24]	0.819	0.097	0.067
3	MobileNetV2 [33]	0.822	0.108	0.075
4	EfficientNetV2-B1 [37]	0.833	0.19	0.132
5	CvT-21 [40]	0.834	0.182	0.127
6	EfficientNetV2-B0 [37]	0.834	0.19	0.132
7	LeViT-256 [12]	0.835	0.15	0.104
8	CvT-13 [40]	0.835	0.144	0.1
9	DeiT-tiny-patch16-224 [39]	0.835	0.172	0.12
10	LeViT-384 [12]	0.836	0.168	0.117
11	ResNet152V2 [14]	0.836	0.192	0.134
12	InceptionV3 [36]	0.836	0.19	0.132
13	LeViT-128 [12]	0.836	0.177	0.123
14	InceptionResNetV2 [35]	0.836	0.181	0.126
15	ResNet101v2 [14]	0.837	0.195	0.136
16	LeViT-128S [12]	0.837	0.171	0.119
17	ResNet50v2 [14]	0.837	0.197	0.137
18	EfficientNetV2-B2 [37]	0.837	0.213	0.148
19	LeViT-192 [12]	0.839	0.198	0.137
20	EfficientNetV2-B3 [37]	0.839	0.235	0.164
21	DeiT-S-patch16-224 [39]	0.841	0.202	0.141
22	ConvNeXt-S [23]	0.841	0.232	0.162
23	DeiT-B-patch16-224 [39]	0.842	0.184	0.127
24	ResNet152 [13]	0.843	0.255	0.178
25	Xception [4]	0.843	0.228	0.159
26	ConvNeXt-B [23]	0.844	0.275	0.192
27	ResNet101 [13]	0.844	0.266	0.186
28	ResNet50 [13]	0.844	0.265	0.185
29	DenseNet201 [15]	0.845	0.265	0.186
30	DenseNet169 [15]	0.845	0.264	0.185
31	DenseNet121 [15]	0.846	0.267	0.187
32	Swinv2-B-p4-w16-256 [21]	0.846	0.243	0.169
33	NASNetMobile [41]	0.846	0.256	0.179
34	NASNetLarge [41]	0.846	0.258	0.181
35	Swinv2-S-p4-w16-256 [21]	0.848	0.27	0.189
36	Swin-S-p4-w7-224 [22]	0.849	0.284	0.199
37	Swin-T-p4-w7-224 [22]	0.850	0.298	0.209
38	ConvNeXt-T [23]	0.850	0.322	0.225
39	Swinv2-T-p4-w16-256 [21]	0.852	0.302	0.212
40	Swin-B-p4-w7-224 [22]	0.857	0.295	0.208
41	VGG16 [34]	0.857	0.371	0.262
42	VGG19 [34]	0.857	0.375	0.265

Although the source of inspiration of ImageNet - WordNet - is naturally hierarchical (see Figure 1 for the WNCSM), no information regarding the semantic similarity of classes was used during the training of the examined networks. Despite that, all 42 networks (both the CNNs and transformers) used in the analysis were able to relate classes with each other. In Figure 2, we provide example NCSMs for 8 networks: 4 CNNs and 4 ViTs. The clearly visible block diagonal structure of all NCSMs exhibits high resemblance to the WNCSM (the Class Similarity Matrix created with semantic similarity). This structure is weaker for

the mobile models, though still visible. Models from the ConvNeXt and Swin transformer (hierarchical transformer) families build less-noisy class similarity landscape (high contrast of NCSMs). It indicates the potential superiority of these models to other ones, which exhibit a lot of outside group noise in their NCSMs.

A similar structure of the WNCSM and all NCSMs undoubtedly shows that the similarity perceived by all networks (CNNs and ViTs) is related to the semantic similarity. Let us now quantify this phenomenon by computing 3 variants of Semantic Similarity Alignment Degree (SSAD). The numerical results sorted by the increasing $SSAD_{Cosine}$ value have been presented in Table 1. We also included a larger table with the number of parameters, the most similar pairs and the ImageNet testing accuracy in Appendix B, Table B.1. Although the value ranges differ significantly for different SSAD variants, the overall ordering of networks is very similar. All measures show a positive correlation between network size and semantic similarity. The lowest $SSAD$ values have been obtained for small, mobile models (Mobile-ViTs, MobileNetV2), and the highest ones for the largest (and the oldest) models with quite modest accuracy - VGGs (it may be due to their different classifier structure, consisting of a Flattening and a few Dense layers). On the other hand, although the MobileViT-S' accuracy is quite high, network's semantic relation is not as developed as the one of other networks.



(a) All networks (b) Transformers (c) CNNs

Fig. 3. Spearman correlation between size and accuracy for three variants of SSAD. ViTs exhibit positive correlations between all versions of SSAD and both the accuracy and size.

The qualitative results in Tab. 2 visualize it. The table presents top 5 similar classes to example 4 classes from the animal, objects and a geological formation semantic groups according to example networks and WordNet. Even in the close similarity neighborhood of example classes for MobileViT-S, we obtain unrelated classes, such as ski - chain saw, sleeping bag - pencil box. Hierarchical transformers obtained high SSAD results. They are followed by ConvNeXts, ResNets, NASNets and DenseNets. Other networks - pure Transformer (DeiT) and transformer-convnet hybrids (CvT, LeViT) were placed below the aforementioned networks along with EfficientNets and ResNetsV2. It is visible that a family membership strongly impacts SSAD (e.g. see how DenseNets take places after each other in the table, and members of other families lie close to each other). Moreover, other examples in Table 2 show that the closest neighborhood of classes that are natural creations (tiger shark and Alps) largely coincides with the results returned by WordNet. WordNet similarity returns only loosely con-

nected categories for artificial objects such as ski - lighter, while networks return rather closely related categories (not particularly semantically similar).

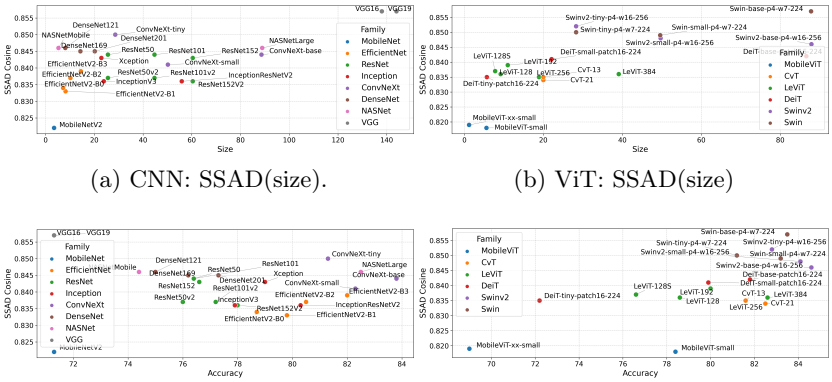
Table 2. 5 most similar classes to 4 examples according to WordNet and 5 networks.

Class	Similarity	1st neighbor	2nd neighbor	3rd neighbor	4th neighbor	5th neighbor
tiger	WordNet	hammerhead	white shark	electric ray	stingray	barracouta
shark	ConvnextTiny	white shark	hammerhead	dugong	scuba diver	stingray
	Swin base	white shark	hammerhead	scuba diver	stingray	dugong
	VGG16	white shark	hammerhead	dugong	stingray	sturgeon
	MobileViT-S	hammerhead	white shark	dugong	stingray	scuba diver
	ResNet101	white shark	hammerhead	scuba diver	sturgeon	stingray
ski	WordNet	lighter	pick	remote control	oil filter	pier
	convnext tiny	snowmobile	dogsled	alp	ski mask	puck
	swin base	snowmobile	dogsled	alp	bobsled	snowplow
	vgg16	snowmobile	dogsled	bobsled	alp	paddle
	MobileViT-S	snowmobile	snowplow	alp	ski mask	chain saw
	ResNet101	snowmobile	alp	ski mask	bobsled	snowplow
sleeping	Wordnet	mailbag	backpack	purse	plastic bag	pot
bag	Convnext tiny	mountain tent	quilt	pajama	punching bag	stretcher
	Swin base	mountain tent	quilt	stretcher	bath towel	studio couch
	vgg16	quilt	mountain tent	studio couch	stretcher	sweatshirt
	MobileViT-S	mountain tent	stretcher	studio couch	quilt	pencil box
	ResNet101	mountain tent	stretcher	studio couch	bath towel	punching bag
alp	WordNet	volcano	promontory	cliff	seashore	coral reef
	convnext tiny	valley	promontory	volcano	cliff	ski
	swin base	valley	promontory	cliff	volcano	mountain bike
	vgg16	valley	cliff	volcano	promontory	mountain tent
	MobileViT-S	valley	ibex	cliff	mountain bike	ski
	ResNet101	valley	ski	mountain bike	promontory	volcano

4.2 How do network size and accuracy relate with Semantic Similarity Alignment Degree?

The relationship between the size of the model and its semantic alignment, which we noticed visually in Tab. 1, prompted us to investigate it numerically, as well as the relationship between SSAD and model accuracy. Fig. 3 presents the Spearman correlation for SSAD and size/ImageNet accuracy. Moderate positive correlations between size and SSAD suggest that larger networks' perceive similarity closer to the semantic similarity, which supports our qualitative finding. Although it occurs for the whole population, the correlation is significantly higher for ViTs than CNNs. These results are supported by scatter plots of SSAD(size) for CNNs and ViTs presented in Fig. 4. The scatter plots reveal a clear, positive (non-linear) relationship for ViTs, and existent, but less evident for CNNs. For the SSAD-accuracy correlation, Spearman correlation results imply a low positive correlation for all networks. By analizing the networks separately, we can see that ViTs exhibit a moderate positive correlation, while CNNs - a small negative correlation.

To analyze these correlations in more detail, we provide the scatter plots of SSAD(accuracy) in Fig. 4. It is visible that while for ViTs the positive relationship between these two can be observed, for CNNs no obvious relation exists.



(a) CNN: SSAD(size). (b) ViT: SSAD(size)

(c) CNN: SSAD(accuracy). (d) ViT: SSAD(accuracy).

Fig. 4. SSAD and accuracy/size Scatter plots. A visible relation between ViTs' SSAD and both: accuracy and size can be noticed.

Moreover, by analyzing the positions of the networks in the scatter plot for ViTs, we observe that networks are not always ordered by size. This indicates that the positive relationship between accuracy and SSAD is not confounded by network size. The results indicate that not only ViTs' accuracy scales with size but also their capacity to align better with semantic similarity. Consequently, their degree of Semantic Similarity Alignment correlates with accuracy. Also, the highest $SSAD$-accuracy correlation results for $SSAD_{Cosine}$ imply its best suitability for performance analysis. In contrast, due to the fact that CNNs' accuracy is not correlated with SSAD, it suggests that it can be used as an additional criterion for the model selection. For models with the same accuracy, the model with higher SSAD can be selected, because its perceived similarities can be better explained with semantic similarities and lexical ontologies.

4.3 What other semantic relations are perceived?

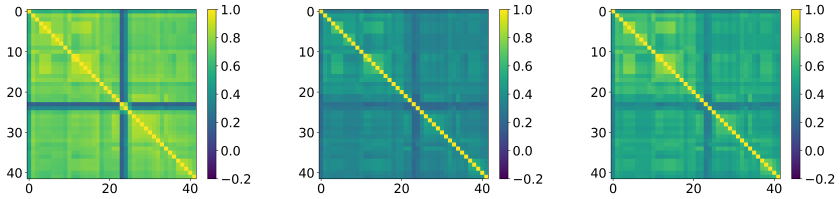
In Fig. 5 we include examples of the most similar categories to example categories from Tab. 2. We can notice some co-occurring (presumably often) objects in the same image, which is the most probable cause of some similarities in this table. They are not connected to semantic similarity, but other semantic relations. In the next fragments, let us provide some concrete examples from the TOP 50 pairs of CNPRs obtained for the tested networks. For each example, we name a semantic relation that presumably resulted in the emergent similarity, having a direct impact on visual features of the image.

Homophones/partial homophones Our analysis of the CNPR allowed us to indicate some pairs in ImageNet that due to the same name (such words are called homophones) or almost the same name either (1) include in their training folders some incorrectly labeled images or (2) the folders' content overlap. The

(a) (b) (c) (d) (e) (f)

Fig. 5. Co-occurence of concepts makes them similar/related for networks: (a) sharks are photographed with divers; (b) ibex lives in Alps; (c, d) bikes/skis are used in mountains; (e) a skier with skis and a mask; (f) a sleeping bag with a tent. All images are from the Imagenet-1k dataset.

example of (1) includes MobileViT-S placing **tiger - tiger cat** at the 47th place in the ranking. Although tigers and tiger cats do exhibit some similarities, this very high similarity value is most probably caused by the confusion of two WordNet nodes – tiger cat (WN 3.0: 02123159-n) and Felis tigrina, tiger cat (WN 3.0: 02126465-n) – the first node is a hyponym of a domestic cat and the other one – of a wildcat. As a result of category names being homophones, some labeling issues occurred while creating the training set of ImageNet: the folder representing a domestic cat includes many tigers. The example of (2) can be Swin-Base placing **sunglasses - sunglass** at the 19th place in the ranking. While the term 'sunglasses' (WN 3.0: 04356056-n) is obvious, sunglass (WN 3.0: 04355933-n) is defined in WordNet as "lens that focuses the rays of the sun; used to start a fire" and it should definitely be separated from sunglasses, while the content of sunglass training folder has been created with sunglasses and is only a misleading duplicate of the sunglasses category.



(a) Cosine Similarity (b) Kendall Corr. (c) Spearman Corr.

Fig. 6. Similarity/correlation (corr.) matrices computed for all NCSMs ($NSAD_{Cosine}$, $NSAD_{Kendall}$, $NSAD_{Spearman}$). Fig. 7 shows that visible clusters of networks represent the models from the same family – truly similar models via architecture.

Hypernyms/hyponyms Although ImageNet classes are WordNet leaves, some hypernymity relations can still be found. Our analysis of the CNPRs allowed us to indicate some examples: **mushroom - agaric** placed at the 43th (DenseNet121) and **tub - bathtub** placed at the 7th (CVT-13) place in CNPR. In the case of mushroom - agaric, the definition of mushroom (WN 3.0: 07734744-n) has been extended from "fleshy body of any of numerous edible fungi" to "ed-

ible or poisonous fungi" (the training folder contains also poisonous mushrooms, such as flybanes). After extending the definition, agaric (WN 3.0: 12998815-n) becomes mushroom's hyponym. A similar example is tub - bathtub (in ImageNet, tub is just a category with bathtub and additional other tubs, such as hot tub etc.). Yet another example is **assault rifle - rifle** placed at the 35th place in the EfficientNetV2-B0's CNPR.

Synonyms CNPRs allowed us to indicate synonyms within ImageNet classes that can be considered duplicates/redundant. Two examples are: **missile - projectile** and **laptop - notebook** placed at the 1st (by the majority of networks) and the 27th (DeiT-Tiny) places. While these terms differ slightly in WordNet, they are treated as synonyms in ImageNet: missile (WN 3.0: 03773504)/projectile (WN 3.0: 04008634-n) represent rocket explosives, while the second pair – portable computers. Some images are duplicated within the folders.

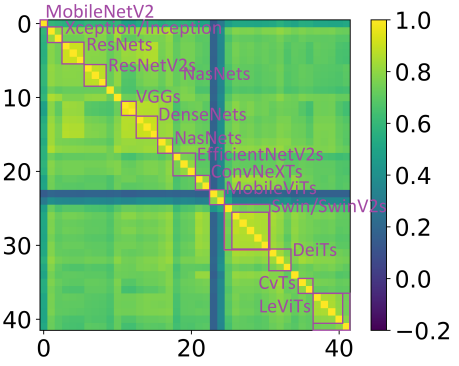


Fig. 7. $NSAD_{Cosine}$ from Fig. 6 shows that networks cluster in families (similar networks), showing the its usefulness for image-free network comparisons.

Siblings/sister terms This is the broadest relation group including categories that share (1) visual/functional similarities, (2) inter-species relation (e.g. ancestor-descendant, a common ancestor), (3) within-species gender relation. We provide a few examples: (1) **cassette player - tape player** placed 11th by Xception and **barbell - dumbbell** placed 43th by ResNet101V2 in the CNPR; (2) **Indian elephant - tusker** (19th place, VGG16), **brown bear - American black bear** (43th place, ResNet50), **tiger beetle - ground beetle** (41th place, ResNet50); (3) **hen - cock** (ranked 47th in NASNet-L's CNPR).

Meronyms/holonyms Another relation that can be found via interpreting the results in the Closest Neighbor Pair Ranking is meronymy or holonymy. It occurs, when one concept is a physical part of another concept. A few examples that our analysis helped us to recognize are: **screen - monitor** placed at the 11th place by InceptionV3, **typewriter keyboard - space bar** placed at the 33rd place by EfficientNetV2-B1 and **breastplate - cuirass** (breastplate, WN 3.0: 03146219-n, is the front part of a cuirass, WN 3.0: 02895154-n) placed at the 29th place by MobileViT-S.

Containment High similarity perception can occur when the containment semantic relation exists (it can be perceived as specific type of co-occurrence). It means that one concept is contained by another one (e.g. room/landscape). An example from the ResNet50's CNPR (rank 21) can be a pair **barber chair - barber shop**. The other examples are **ibex - Alps**, **skis/mountain bike - Alps** from top5 neighbors of example classes (Tab. 2, Fig. 5).

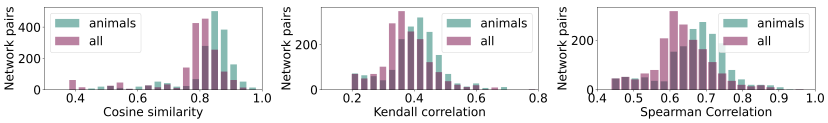


Fig. 8. Distribution of network similarities computed for the whole class space and a single domain – animals (NSAD: Network Similarity Alignment Degree with variants Cosine Similarity, Kendall and Spearman). Models are more similar in a single domain.

Physical proximity Another reason why a network perceives concepts as similar is their frequent co-occurence (physical proximity) in the training images. The example of such a relation can be **academic gown - mortarboard** from the 41st place (Xception) in the Closest Neighbor Pair Ranking. The other ones can be **Tiger Shark - scuba diver**, **skis - ski mask** and **sleeping bag - mountain tent** from top5 neighbors of example classes (Tab. 2, Fig. 5).

4.4 Do networks share similarity perception?

In Fig. 6, we present the matrices obtained for the comparison of NCSM of all examined networks with different $NSAD$ variants ($NSAD_{Cosine}$, $NSAD_{Kendall}$, $NSAD_{Spearman}$). Each value in these matrices reflects the similarity in how a specific pair of networks perceives relationships among classes. Network similarity perception is the most similar within network families (see a block diagonal structure of the matrices). It shows the impact of architectural choices on learning class similarities. We obtained the highest differences for mobile models (MobileViTs, MobileNetV2) compared to all the other models. It manifests itself as stripes belonging to index 0 and a distinctive cross in Fig. 6. We show in Fig. 7 that $NSAD$ values **cluster the models from the same family together (thus architecturally similar models), showcasing the usefulness of our methods for image-free model comparison.**

We also present the histograms of the pair-wise similarity/correlation values ($NSAD$) between networks in Fig. 8 (red histogram). The distributions are roughly Gaussian with relatively high mean. Inspired by the clear box structure of CSMs for the animal group (visible in individual Class Similarity Matrices), we decided to compute the correlations/similarities between network CSMs and generate histograms for only these classes (we drop others, therefore we use a smaller set of class representations) and see how it impacts the distributions (Fig. 8 – green histograms). These distributions have higher central values compared to those computed based on the whole class set, indicating more homogeneous similarity perceptions among networks within the animal domain. This result **suggests that model similarity is greater within single-domain class groups than across broader, multi-domain categories.**

5 Discussion

Our analysis and the proposed tool set helped us to answer the questions defined in the Introduction. **(1) Is similarity perceived by ImageNet-trained CNNs and ViTs related to semantic similarity?** All the examined networks developed similarity perception related to semantic similarity. It is supported by their CSMs similar to those created with semantic similarity via WordNet and the numerical analysis with SSAD. Although perceived and semantic similarities are related, they are not equivalents, which suggests that network perception encompasses other semantic relations than similarity.; **(2) Is there a relationship between the networks' ability to align their similarity perception to semantic similarity with their size and ImageNet accuracy?** Network size and semantic alignment are positively correlated (stronger correlation for ViTs than for CNNs). ViTs exhibit a positive correlation between their SSAD and accuracy, while CNNs do not. This suggests that SSAD can be used as an additional criterion for model selection. At similar accuracy, a model with higher SSAD can be chosen, as its perceived similarities can be more easily explained with semantics, having a positive impact on explainability. In our future work, we will dig deeper into the differences in similarity perception of CNNs and ViTs. We will examine why attention-based models better align with semantics, explore links between representation geometry and self-attention, and test whether model size causally influences the alignment. We will also focus on evaluating the robustness of SSAD to determine the significance of its variations, as the score differences are often small.; **(3) Do networks perceive other semantic relations besides similarity?** Deep networks perceive not only the semantic similarity, but also other semantic relations, such as meronymy, containment or physical proximity. While the analysis of CNPRs is a manual effort it can improve our understanding of the causes of some developed similarities.; **(4) Do different networks share similarity perception?** Yes, the networks largely share similarity perception. It is proven by highly similar structure of their Class Similarity Matrices, high alignment of all Network CSMs with WordNet CSMs (the same reference used for all networks) and high similarities between the CSMs of different networks. The practical implication is that because this perception is not identical and because it is shared mostly within model families (therefore implicitly similar models), the network alignment can be used to compare the representational similarity of models at a category level.

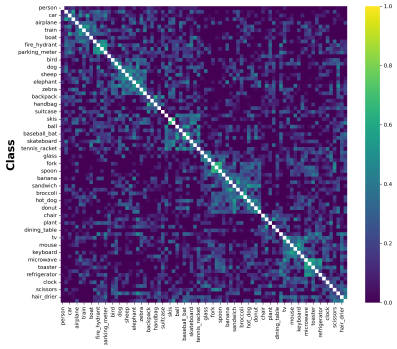


Fig. 9. CSM for COCO 2017 and DETR-ResNet50.

Limitations: As we focused on ImageNet, it could introduce biases in class representation affecting the validity of our conclusions. While it provided a strong foundation for investigating visual representations, we also include a CSM obtained for COCO 2017 object detection in Fig. 9 to improve generalizability. We provide more CSMs for different tasks (COCO detection, segmentation), as well as for a Self-Supervised Learning (SSL) model in Appendix A. The box-diagonal structures present across all matrices (sorted via WordNet nodes) underscore that networks across various tasks and datasets align their perception with semantics, reflecting the inherent correlations of the visual world. Identifying non-similarity relations requires manual inspection, limiting scalability. We will explore how to automate it in our future work, however manual analysis still best captures their context-dependent nature. While our approach, based on the final-layer weights, offers simplicity and efficiency, other approaches to obtain representations could be used to form a more comprehensive similarity view, e.g. confusion matrices, intermediate-layer features, weights, and bias contributions could be used. A limitation can be the reliance on WordNet similarity, as its linguistically driven structure may not align with visual similarity. Also, visual similarity may sometimes be more relevant for certain tasks. However, WordNet's well-defined organization is beneficial for modeling and often aligns with visual similarities, e.g. when stemming from shared functionalities or evolution.

6 Conlusions

The framework introduced in the paper can help to better understand deep models (e.g. what they perceive similar, whether their perception aligns with semantics or with the one of other networks) and their training datasets (e.g. labeling issues, overlapping, duplicated classes). Our methods do not require any images for testing, while our insights and results can serve as a reference for future comparisons and benchmarking due to analyzing a large set of networks. As a large part of the visual and semantic similarities naturally intersect, vision networks should be able to discover such links, and they do. The degree of this alignment can be measured by the proposed metrics, and thus can be used as an additional model selection criterion. We showed that our metrics enable image-free comparison of different networks, as they cluster models from the same family (architecturally similar) together, which is important for the growing area of model representational similarity comparisons [19].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

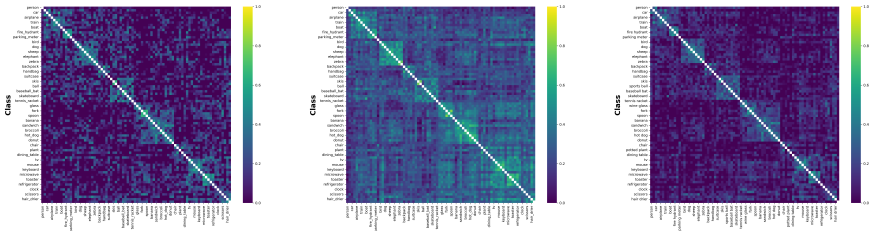
1. Bilal, A., Jourabloo, A., Ye, M., Liu, X., Ren, L.: Do convolutional neural networks learn class hierarchy? IEEE Transactions on Visualization and Computer Graphics **24**(1), 152–162 (2017)

2. Bowers, J.S., Malhotra, G., Dujmović, M., Montero, M.L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J.E., Heaton, R.F., et al.: Deep problems with neural network models of human vision. Behavioral and Brain Sciences **46** (2023)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: International Conference on Computer Vision. pp. 9650–9660 (2021)
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Conference on Computer Vision and Pattern Recognition. pp. 1251–1258 (2017)
5. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: European Conference on Computer Vision (2010)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (2009)
7. Deselaers, T., Ferrari, V.: Visual and semantic similarity in imagenet. In: Conference on Computer Vision and Pattern Recognition. pp. 1777–1784 (2011)
8. Filus, K., Domanska, J.: Netsat: Network saturation adversarial attack. In: IEEE International Conference on Big Data. pp. 5038–5047 (2023)
9. Filus, K., Domańska, J.: Extracting coarse-grained classifiers from large convolutional neural networks. Engineering Applications of Artificial Intelligence **138**, 109377 (2024)
10. Filus, K., Domańska, J.: Similarity-driven adversarial testing of neural networks. Knowledge-Based Systems p. 112621 (2024)
11. Goldstone, R.L., Son, J.Y.: Similarity. Oxford University Press (2012)
12. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. In: International Conference on Computer Vision. pp. 12259–12269 (2021)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision. pp. 630–645 (2016)
15. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)
16. Huang, T., Zhen, Z., Liu, J.: Semantic relatedness emerges in deep convolutional neural networks designed for object recognition. Frontiers in Computational Neuroscience **15**, 625804 (2021)
17. Jere, M., Rossi, L., Hitaj, B., Ciocarlie, G., Boracchi, G., Koushanfar, F.: Scratch that! an evolution-based adversarial attack against neural networks. arXiv preprint arXiv:1912.02316 (2019)
18. Kolb, P.: Experiments on the difference between semantic similarity and relatedness. In: Nordic Conference of Computational Linguistics. pp. 81–88 (2009)
19. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning (2019)
20. Leacock, C., Chodorow, M.: Combining local context and wordnet similarity for word sense identification. WordNet: An electronic lexical database **49**(2) (1998)
21. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Conference on Computer Vision and Pattern Recognition. pp. 12009–12019 (2022)
22. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Conference on Computer Vision and Pattern Recognition. pp. 10012–10022 (2021)

23. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
24. Mehta, S., Rastegari, M.: Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (2022)
25. Miller, G.A.: WordNet: An electronic lexical database. MIT press (1998)
26. Mopuri, K.R., Shaj, V., Babu, R.V.: Adversarial fooling beyond" flipping the label". In: Conference on Computer Vision and Pattern Recognition Workshops (2020)
27. Muttenthaler, L., Linhardt, L., Dippel, J., Vandermeulen, R.A., Hermann, K., Lampinen, A., Kornblith, S.: Improving neural network representations using human similarity judgments. Advances in Neural Information Processing Systems **36** (2024)
28. Nayak, G.K., Mopuri, K.R., Shaj, V., Radhakrishnan, V.B., Chakraborty, A.: Zero-shot knowledge distillation in deep networks. In: International Conference on Machine Learning. pp. 4743–4751 (2019)
29. Pedersen, T., Patwardhan, S., Michelizzi, J., et al.: Wordnet:: Similarity-measuring the relatedness of concepts. (2004)
30. Roads, B.D., Love, B.C.: Enriching imagenet with human similarity judgments and psychological embeddings. In: conference on computer vision and pattern recognition. pp. 3547–3557 (2021)
31. Rosch, E., Lloyd, B.B.: Cognition and categorization (1978)
32. 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 **115**, 211–252 (2015)
33. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
35. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence. vol. 31 (2017)
36. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
37. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning. pp. 10096–10106 (2021)
38. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. Advances in neural information processing systems **29** (2016)
39. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P.: Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 (2020)
40. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: International Conference on Computer Vision. pp. 22–31 (2021)
41. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Conference on Computer Vision and Pattern Recognition. pp. 8697–8710 (2018)

A Generalizability of the findings - other networks

We show CSMs for 3 networks trained on COCO 2017[1] (80 classes, object detection, semantic segmentation) in Fig. A.1. We also provide a CSM for a Self-Supervised Learning (SSL) model (DINOv2[2]) in Fig. A.2 for the mini-ImageNet [38] templates. All matrices show a characteristic box-diagonal structure. The boxes for COCO include classes related via semantic relations, such as the membership to one basic-level category (e.g. animals for dog, sheep etc.) or physical proximity (e.g. forks or bananas can be both often found in a kitchen).



(a) DETR, object detection [3]

(b) YOLOST, object detection [4]

(c) MaskFormerT semantic segmentation [5]

Fig. A.1. Class Similarity Matrices generated for COCO 2017 models.

B ImageNet additions

In Table B.1, we extend Tab. 1 with the number of model parameters and ImageNet accuracy. We can see that ImageNet accuracy and network size increase with larger SSAD values. For each network, we also provide the pair of classes perceived as the most similar. Most networks perceive the projectile-missile pair as the most similar classes (connected semantically via a synonymity relation). 3 out of 4 LeViTs perceive the jaguar-leopard pair as the most similar, suggesting that the membership to a given network family also impacts what pair is perceived as the most similar. Besides a few networks returning different

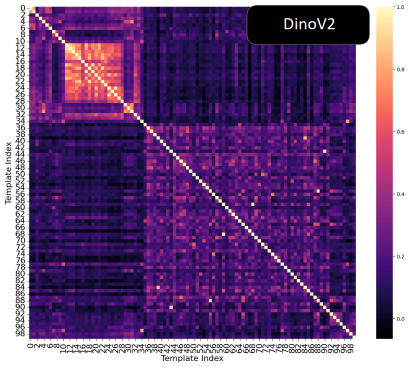


Fig. A.2. DINOv2's CSM.

[1] COCO 2017 dataset.
[2] DINOv2 - HuggingFace model card.
[3] DETR (End-to-End Object Detection) - HuggingFace model card.
[4] YOLOS (tiny-sized) - HuggingFace model card.
[5] MaskFormer - HuggingFace model card.

classes than missile-projectile, all pairs show highly semantically related concepts (bathtub-tub, husky-Eskimo dog).

Table B.1. Models with SSAD, ImageNet accuracy (Acc.) and parameters (Params.).

	Name	Params.	Acc.	Most similar pair	Cosine	Spearman	Kendall
1	MobileViT-small [24]	5.6M	78.4 [24]	projectile - missile	0.818	0.079	0.055
2	MobileViT-xx-small [24]	1.2M	69 [24]	projectile - missile	0.819	0.097	0.067
3	MobileNetV2 [33]	3.5M	71.3 [6]	projectile - missile	0.822	0.108	0.075
4	EfficientNetV2-B1 [37]	8.2M	79.8 [6]	projectile - missile	0.833	0.19	0.132
5	CvT-21 [40]	20M	82.5 [40]	projectile - missile	0.834	0.182	0.127
6	EfficientNetV2-B0 [37]	7.2M	78.7 [6]	missile - projectile	0.834	0.19	0.132
7	LeViT-256 [12]	18.9M	81.6 [12]	leopard - jaguar	0.835	0.15	0.104
8	CvT-13 [40]	20M	81.6 [40]	projectile - missile	0.835	0.144	0.1
9	DeiT-tiny-patch16-224 [39]	5.7M	72.2 [7]	projectile - missile	0.835	0.172	0.12
10	LeViT-384 [12]	39.1M	82.6 [12]	leopard - jaguar	0.836	0.168	0.117
11	ResNet152V2 [14]	60.4M	78 [6]	missile - projectile	0.836	0.192	0.134
12	InceptionV3 [36]	23.9M	77.9 [6]	missile - projectile	0.836	0.19	0.132
13	LeViT-128 [12]	9.2M	78.6 [12]	leopard - jaguar	0.836	0.177	0.123
14	InceptionResNetV2 [35]	55.9M	80.3 [6]	missile - projectile	0.836	0.181	0.126
15	ResNet101v2 [14]	44.7M	77.2 [6]	projectile - missile	0.837	0.195	0.136
16	LeViT-128S [12]	7.8M	76.6 [12]	tub - bathtub	0.837	0.171	0.119
17	ResNet50v2 [14]	25.6M	76 [6]	missile - projectile	0.837	0.197	0.137
18	EfficientNetV2-B2 [37]	10.2M	80.5 [6]	missile - projectile	0.837	0.213	0.148
19	LeViT-192 [12]	11M	80 [12]	jaguar - leopard	0.839	0.198	0.137
20	EfficientNetV2-B3 [37]	14.5M	82 [6]	projectile - missile	0.839	0.235	0.164
21	DeiT-small-patch16-224 [39]	22.1M	79.9 [7]	missile - projectile	0.841	0.202	0.141
22	ConvNeXt-small [23]	50.2M	82.3 [6]	projectile - missile	0.841	0.232	0.162
23	DeiT-base-patch16-224 [39]	86.6M	81.8 [7]	missile - projectile	0.842	0.184	0.127
24	ResNet152 [13]	60.4M	76.6 [6]	projectile - missile	0.843	0.255	0.178
25	Xception [4]	22.9M	79 [6]	missile - projectile	0.843	0.228	0.159
26	ConvNeXt-base [23]	88.6M	83.8 [23]	projectile - missile	0.844	0.275	0.192
27	ResNet101 [13]	44.7M	76.4 [6]	missile - projectile	0.844	0.266	0.186
28	ResNet50 [13]	25.6M	74.9 [6]	missile - projectile	0.844	0.265	0.185
29	DenseNet201 [15]	20.2M	77.3 [6]	projectile - missile	0.845	0.265	0.186
30	DenseNet169 [15]	14.3M	76.2 [6]	projectile - missile	0.845	0.264	0.185
31	DenseNet121 [15]	8.1M	75 [6]	projectile - missile	0.846	0.267	0.187
32	Swinv2-base-p4-w16-256 [21]	87.9M	84.6 [22]	missile - projectile	0.846	0.243	0.169
33	NASNetMobile [41]	5.3M	74.4 [6]	missile - projectile	0.846	0.256	0.179
34	NASNetLarge [41]	88.9M	82.5 [6]	missile - projectile	0.846	0.258	0.181
35	Swinv2-small-p4-w16-256 [21]	49.7M	84.1 [8]	missile - projectile	0.848	0.27	0.189
36	Swin-small-p4-w7-224 [22]	49.6M	83.2 [8]	missile - projectile	0.849	0.284	0.199
37	Swin-tiny-p4-w7-224 [22]	28.3M	81.2 [8]	projectile - missile	0.85	0.298	0.209
38	ConvNeXt-tiny [23]	28.6M	81.3 [6]	projectile - missile	0.85	0.322	0.225
39	Swinv2-tiny-p4-w16-256 [21]	28.3M	82.8 [8]	missile - projectile	0.852	0.302	0.212
41	Swin-base-p4-w7-224 [22]	87.8M	83.5 [8]	husky - eskimo dog	0.857	0.295	0.208
40	VGG16 [34]	138M	71.3 [6]	projectile - missile	0.857	0.371	0.262
42	VGG19 [34]	144M	71.3 [6]	missile - projectile	0.857	0.375	0.265

[6] Keras Applications

[7] Data-efficient Image Transformer (DeIT) - HuggingFace model card

[8] Swin Transformer - Girhub Repository