On the Use of Language and Vision Models for Cognitive Science: The Case of Naming Norms

Zhirui Chen (zhirui.chen@student.uva.nl)
Institute for Logic, Language and Computation, Amsterdam, Netherlands

Andreas Mädebach (a.maedebach@gmail.com)
Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain

Eleonora Gualdoni (eleonora.gualdoni@upf.edu)
Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain

Gemma Boleda (gemma.boleda@upf.edu)
Department of Translation and Language Sciences, Universitat Pompeu Fabra / ICREA, Barcelona, Spain

Abstract

Computational models have long been used in Cognitive Science, but to date most research has used language models trained on text. With recent advances in Computer Vision, new research is expanding to visually informed models. In this paper, we explore the potential of such models to account for human naming behavior as recorded in naming norms (where subjects are asked to name visually presented objects). We compare the performance of three representative models on a set of norms that include stimuli in the form of line drawings, colored drawings, and realistic photos. The state-of-the-art Language and Vision model CLIP, trained on both text and images, performs best. It generalizes well across different types of stimuli and achieves good overall accuracy. CLIP affords both linguistic (text-based) and visual (image-based) representations for names, and we find that textual representations outperform visual representations. This is good news, as textual representations are easier to obtain than visual representations. All in all, our results show promise for the use of Computer Vision and Language and Vision models in Cognitive Science.

Keywords: object naming; naming norms; computer vision; psycholinguistics

Introduction

The last decade has seen a leap in the capabilities of AI models. Already before the current deep learning wave, computational models were being used in Cognitive Science. For instance, Shriberg and Stolcke (1996) used language models to model word predictability in humans. However, this earlier work, as well as most current work using deep neural networks (Goodkind & Bicknell, 2018), used models based on text, as those are the most developed in AI. This means that for other modalities relevant to cognition, like visual stimuli, obtaining data (such as typicality ratings) from human subjects remains the prevailing method. With the advent of deep learning in Computer Vision, the situation is changing, and recent work showcases the potential of these models for Cognitive Science (Günther, Marelli, Tureski, & Petilli, 2023; Gualdoni, Brochhagen, Mädebach, & Boleda, 2023; Brochhagen, Boleda, Gualdoni, & Xu, 2023).

In this paper, we test the potential of Computer Vision and Language and Vision models to model naming behavior in people. Note that our goal is to assess computational models for use in Cognitive Science, as opposed to answering a specific question about cognition. As data, we use naming norms, i.e. collections of names for visually presented stimuli, which are a standard tool to investigate processes related to lexical production. Naming norms are created by asking subjects to name carefully curated sets of images (more details below). As illustrated in Figure 1, naming norms have been collected for different types of images, from black-and-white drawings to more realistic stimuli. We test models on the alignment between images in the norms, on the one hand, and their names, on the other, where the representation of both the images and the names are built from the models.[1] Specifically, we test them in terms of how well they match a given object with the most frequent name produced for it (in Figure 1, that would be "penguin"). This is a general proxy measure for the ability of models to account for naming behavior; we leave more specific applications for future work. We deliberately focus on off-the-shelf models that are easily available,[2] as opposed to adapting models or training models from scratch. This is because we want to ascertain whether current models are mature enough for use by cognitive scientists as is (which would lower the bar in terms of required skills, making them usable by a larger set of researchers), and to test their ability to generalize across different kinds of stimuli.

Related Work

Naming norms. Naming norms are sets of images for which naming data are collected, usually selected so as to be easily identifiable. The images are therefore usually highly stylized and prototypical exemplars of the concept expressed by the target name. In early work, the stimuli were not very realistic; for instance, in the seminal paper that introduced naming norms (Snodgrass & Vanderwart, 1980), the objects were presented as line drawings. The field has moved to more and more naturalistic images (see Figure 1 for example images for the name "penguin" in the norms that we will use

[1] Data and scripts are available at `https://osf.io/7dfgx`, as is an appendix with further results.

[2] E.g. through websites and packages such as HuggingFace `https://huggingface.co`.

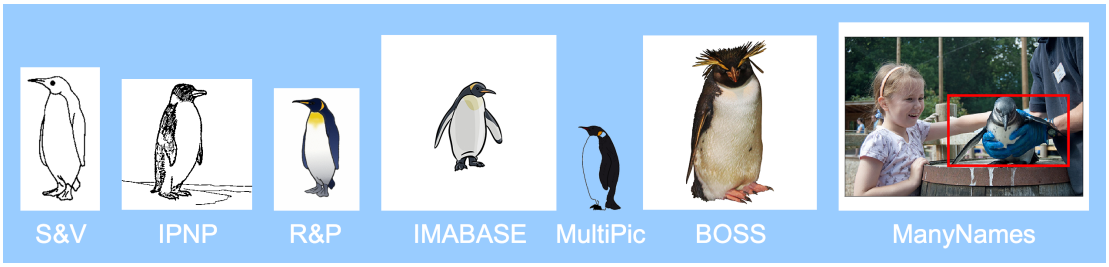

Figure 1: Images named "penguin" from each of the naming norms we use.

in this paper). However, the focus on prototypical exemplars has remained, except for a set of naming norms that was created within the computational linguistics community, MANYNAMES (Silberer, Zarrieß, & Boleda, 2020; Silberer, Zarrieß, Westera, & Boleda, 2020).

Using Computer Vision models for Cognitive Science. A very recent series of studies have shown the promise of Computer Vision models to address research questions in Cognitive Science. Much of this work has targeted image similarity and categorization (Jozwik, Kriegeskorte, Storrs, & Mur, 2017; Peterson, Abbott, & Griffiths, 2018; Zhang, Isola, Efros, Shechtman, & Wang, 2018; Battleday, Peterson, & Griffiths, 2020; Singh, Peterson, Battleday, & Griffiths, 2020); the latter is related to but not the same as the phenomenon of naming that we tackle here. This work has collectively shown that deep visual representations (especially those in the outmost layers of the models) mirror human representations remarkably well, at least if measured by comparing human perceptual similarity judgments to similarity scores assigned by models.

More related to the present article is work by Gualdoni et al. (2023) and Günther et al. (2023), who build visual-semantic representations for relatively large vocabularies using deep learning-based Computer Vision models. In both studies, the authors build representations for individual images as well as concepts or names, like we do here. We adopt part of their methodology, and examine a large set of naming norms including different types of images as well as a wider range of models.

Representations in Computer Vision. The models used by Gualdoni et al. (2023) and Günther et al. (2023) are object classification models, trained to associate images to ground truth labels, such as DOG or CHAIR. ResNet (He, Zhang, Ren, & Sun, 2015) and BottomUp (Anderson et al., 2018) are well-known examples; we use the latter here. To be successful at the task, these models learn to encode the images in vector representations, different for each class label. As a result of this training regime, these models perform very well on data that do not differ from the ones that they have seen in the training phase.

In part to overcome this lack of generalization abilities, multi-modal models have been proposed, with the goal of learning perceptual properties from supervision contained in both images and natural language (as found in textual data). The most successful available Language and Vision model to date is CLIP (Radford et al., 2021), which features both a text encoder and an image encoder.[3] Both encoders are transformers, the state of the art architecture in both computational linguistics and Computer Vision (Vaswani et al., 2023; Dosovitskiy et al., 2021). A key feature of CLIP is its training regime: during training, the model has to match images with descriptive captions (produced by annotators), thus learning to associate a wide variety of visual concepts with the language that is used to talk about them.

On the other side of the spectrum, self-supervised models have been proposed (Devlin, Chang, Lee, & Toutanova, 2019; Radford et al., 2019; Caron et al., 2021), with the idea of leveraging the information contained in the training data itself, without the need for annotations (e.g., manually provided object classes or captions). These models learn representations for images by learning to reconstruct the original image from several perturbations (e.g. from rotated, cropped, or blurred versions). They are known to learn robust and general information, although they typically perform worse for a given task than a supervised system that has been trained for that task. A representative model of this class is the one we use here, the Vision Transformer (Dosovitskiy et al., 2021).

In our experiments, we use a representative example of each type of model.

Method

Naming Norms

We use a representative set of naming norms for English, jointly comprising four image types: black-and-white line drawings, colored drawings, photorealistic isolated images (with no background), and realistic photos with the object in context. Table 1 contains descriptive statistics for the naming norms that we test, and Figure 1 provides an example image of each. Each set of norms provides the name that was produced by the most subjects for each image (henceforth, **top name**).[4]

[3] The encoder is the part of the model that transforms the input data into a vector representation.

[4] Some of them also provide the remaining names that were produced by participants (other than the top names), but this is not con-

Table 1: Norms used in the experiments. "#imgs" and "#names" contain the number of provided images and top names, respectively; "subj/img" indicates the number of subjects by which each image was named (for ManyNames, this is an average).

Image Type	Naming Norms	#imgs	#names	subj/img
line	S&V (Snodgrass & Vanderwart, 1980)	260	260	42
	IPNP free (Székely et al., 2004)	244	243	NA
colored	R&P (Rossion & Pourtois, 2004)	260	260	20
	IMABASE (Bonin et al., 2020)	313	312	30
	MultiPic (Duñabeitia et al., 2018)	750	693	100
photo	BOSS (Brodeur et al. 2010, 2014)	1468	1194	39, 42
real. photo	ManyNames (Silberer, Zarrieß, & Boleda, 2020)	25315	7970	31

Line drawings. The Snodgrass and Vanderwart (1980) naming norms (henceforth, **S&V**) were the first to be created, and they are still widely used as visual stimuli in Cognitive Science. Snodgrass and Vanderwart defined the task as well as the method to gather naming data and assess naming agreement. Subjects were instructed to write the first name that came to mind for each object. The International Picture-Naming Project (Székely et al., 2004, **IPNP**) expanded on the S&V norms by gathering data for 520 common objects in seven languages. 176 of the 520 images were from the S&V norms, and 244 are available as freeware for research purposes. Here we use the IPNP freeware subset (so there is no overlap with S&V), and the naming data for English.

Colored drawings. Rossion and Pourtois (2004, **R&P**) added gray-level texture, surface details and color to the black-and-white line drawings of Snodgrass and Vanderwart and collected data from French-speaking subjects. The pictures for the other two sets of naming norms for colored drawings, **IMABASE** (Bonin et al., 2020) and **MultiPic**, were created anew. While IMABASE and R&P collected French data, they also provide English translations for the top name of each image, and we use those. Compared to names directly collected from English speakers, we expect only marginal differences for the common objects depicted in these databases. MultiPic provides 750 images with naming norms for six languages including British English, which is what we use.

Photorealistic isolated. The 1.5K images in the Bank of Standard Stimuli project (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010; Brodeur, Guérard, & Bouras, 2014, **BOSS**) are photos that were edited so as to have a white background by the creators of the dataset. They were selected to be highly prototypical. Naming data was collected from both French and English speakers; we use the English data.[5]

Photorealistic in context. ManyNames (Silberer, Zarrieß, & Boleda, 2020; Silberer, Zarrieß, Westera, & Boleda, 2020, **ManyNames**) is a large-scale dataset providing 31 names on average for each of 25K objects in real-world images. It includes common objects from different domains (e.g. animals, buildings, etc.). The images were retrieved from the VisualGenome dataset (Krishna et al., 2017), and target objects were highlighted in a red bounding box. ManyNames was created to study naming in a naturalistic context. The objects were not pre-selected to be prototypical or of high quality, beyond general aspects such as having a minimal size. The instructions were the same as in naming norms.

There is no image overlap in stimuli across these naming norms except for the fact that the R&P colored drawings are processed on the basis of S&V black-and-white line drawings. All naming norms provide downloadable image files except S&V. We retrieved the latter by taking screenshots of their pdf file. To enable fair comparison across datasets, and after a preliminary exploration of this setting, for all norms except for ManyNames we used a white margin background around the target object of size 1/4 of the target object. That is, for an object that is e.g. 400×200 pixels in size, the image size would become (100+400+100)×(50+200+50).

Models and Representations

As mentioned above, we use three models that are representative of the three major types of models that are available nowadays: **ViT** (Caron et al., 2021, self-supervised), **Bottom-Up** (Anderson et al., 2018, object classifier), and **CLIP** (Radford et al., 2021, multi-modal). As we aim at testing whether off-the-shelf, easily available models can be of use for cognitive scientists, we do not do any further adaptation, or fine-tuning (Chen, Kornblith, Norouzi, & Hinton, 2020; Kornblith, Shlens, & Le, 2019), of the models.

To obtain visual representations for names from the models, we follow the methodology of Gualdoni et al. (2023) and Günther et al. (2023): the visual representation of a name is defined as the centroid (average) of the visual embeddings (vector representations) of a set of exemplar images tagged with that name in a given image dataset. For Bottom-Up, we use the name representations made available by Gualdoni et al. (2023), the rest we compute ourselves. For ViT, it's the embedding of the [CLS] token. Since CLIP maps two different modalities (language and vision) to the same space, it affords two ways of estimating name representations, one based on visual data (images), and one on linguistic input (text). We explore both types of representations. We extract visual representations from the visual encoder (henceforth, **CLIP-V**), and linguistic representations from the linguistic encoder (henceforth, **CLIP-L**, details below).

sistent across norms and we do not use these data here. They also provide other data, such as naming agreement among subjects.

[5]We use the union of BOSS1 and BOSS2, two data collection efforts within the project in which naming data were collected for different images, and where each image was named by 39 and 42 subjects, respectively.

voc.	#names	repr.	model	section
VGMN	874	visual	B-Up	Analysis 1
VGMN	874	visual	CLIP	Analysis 1
VGMN	874	visual	ViT	Analysis 1
VG	2016	visual	CLIP	Analysis 2
VG	2016	linguistic	CLIP	Analysis 2
THINGS	1854	visual	CLIP	Analysis 2
THINGS	1823	linguistic	CLIP	Analysis 2
UNION	3231	linguistic (x6)	CLIP	Analysis 2

Table 2: Name spaces, with vocabulary label ("voc."), size ("#names"), type of representation ("repr."), model (where B-Up is Bottom-Up), and section where they are used.

Image Datasets for Space Construction

To build name representations, we sampled image exemplars from two different image datasets, namely VisualGenome (Krishna et al., 2017) and THINGS (Hebart et al., 2019).

VisualGenome contains over 108K images of natural scenes where different objects were marked and given a description by human annotators. The ManyNames image in Figure 1 is an example (images in ManyNames were sampled from VisualGenome, and one object per image selected for naming). We use the head noun of the description (provided by VisualGenome) as the object name. The objects in VisualGenome bounding boxes are exemplars of their names, but since they are part of realistic photos, they can be relatively small in size, incomplete, or from an atypical viewpoint. VisualGenome thus has the advantage of being realistic and large, but the disadvantage of being noisy.

THINGS provides a set of 1854 object concepts with corresponding exemplar images (12-35 depending on the concept). The images were gathered through a semi-automatic method involving web searches and ImageNet. Images were checked for size, quality and naturalness of background, and cropped so that they are square. THINGS is still largely naturalistic, but curated and smaller than VisualGenome. 58 of the 1854 concepts correspond to polysemous or homonymous names (27 names in total). For instance, there are two concepts corresponding to the name "bat" (for the animal and the sports tool), each with its own set of image exemplars.

Name Spaces

We build 15 different name spaces (vector representations for names obtained from model embeddings) for the analysis. These spaces were obtained by varying the vocabulary of names taken into account in the evaluation (VGMN, VG, THINGS, UNION, explained below), the type of representation (visual or linguistic embeddings), and the computational model used to obtain them (ViT, Bottom-Up, or CLIP). Table 2 summarizes the different spaces that we worked with and the section of the corresponding analyses.

VGMN and **VG**. We define two vocabularies based on VisualGenome. The first, VGMN, was defined in Gualdoni et al. (2023) and is included for comparison purposes only. It contains the 874 names that occurred as top names for images in ManyNames and had at least 30 exemplars in VisualGenome. The second, VG, is a more comprehensive vocabulary: it includes all the 2016 names in VisualGenome that have at least 30 exemplar images meeting our quality criteria (namely, having width and height no less than 10 pixels, and a bounding box area no less than 1% of the original VisualGenome picture). We built three name spaces for each of these vocabularies: one with ViT, the other two based on CLIP but using either visual or linguistic representations (see Analysis 2 below for details). The image exemplars that went into name representations were taken from VisualGenome: For each name, we used up to 200 images for prototype construction (with random subsampling to 200 if needed), totaling 74,910 object exemplars from 48,041 images.[6]

THINGS. The vocabulary of the THINGS dataset consists of 1,823 names and 1,854 concepts (some names are homonyms; different sets of image exemplars for each concept are associated to a given homonym). As with VG, we obtained two visual spaces (with ViT and CLIP-V), and one linguistic space with CLIP-L. The visual spaces contains separate name representations for each of the 1,854 THINGS concepts. All the provided exemplars are used for prototype computation. Instead, the linguistic space contains representations for the 1,823 names.[7]

UNION. Finally, the UNION vocabulary is the union of the three vocabularies VGMN, MN, and THINGS (total: 3,231 names), which we use the experiments with linguistic representations. Linguistic representations are not bounded by the amount and type of images in any dataset, as they are built based on textual input, which affords the use of a larger vocabulary. Details about the spaces with the UNION vocabulary will be given in Analysis 2 below.

Evaluation

We compare the different representations in terms of how well they match the images in the naming norms with their top name. Specifically, for the evaluation of a particular name space on a particular set of norms, we 1) select the images in the norms whose top name is included in the space; 2) feed them to the ViT, Bottom-Up, or CLIP model to obtain their visual representation; and 3) compute the cosine between each image embedding and each name embedding in the space. We thus obtain a ranked vocabulary for each image, and we use MRR (mean reciprocal rank, i.e., mean of the reciprocal of the rank) of the gold name as the main metric.[8] For fairness,

[6] Gualdoni et al. (2023) excluded from the space all object images that appear in ManyNames, to avoid circularity in their evaluation, which had different goals from ours. We did not exclude the ManyNames images from the construction of the VG space because we wanted to build as general a space as possible, for reuse in future research. 775 (1%) of the object exemplars and 13,944 (29%) of the images are included in ManyNames.

[7] For instance, BAT1 and BAT2 were merged into "bat".

[8] Note that an MRR of 1 means that the gold label is always ranked first; an MRR of 0.5 means that, on average, the gold label is ranked second; 0.33 third, and so on. We choose MRR instead

Img Type	Norms	N	ViT	B-Up	CLIP-V
line	S&V	130	0.04	0.03	**0.39**
	IPNP free	95	0.06	0.04	**0.35**
color	RP	130	0.18	0.18	**0.44**
	IMABASE	130	0.10	0.10	**0.48**
	MultiPic	261	0.09	0.14	**0.42**
photo	BOSS	406	0.18	0.29	**0.35**
real. photo	ManyNames	25K	0.18	**0.40**	0.29
	average		*0.12*	*0.17*	***0.39***

Table 3: Mean Reciprocal Ranks for representations generated with the ViT, Bottom-up and CLIP-V models on the VGMN data (higher is better). *N*: number of images of each norm included for evaluation (all those whose top name is in VGMN); *average*: macro-average (each dataset is one datapoint).

the models are compared on the same set of names (which varies in each experiment).

Analysis 1: Visual vs. Multi-modal Representations

The first analysis shows that, as expected, CLIP, which was explicitly trained to align visual and linguistic information, generalizes better to different kinds of stimuli in naming norms than the Bottom-Up model and ViT.

As shown in Table 3, CLIP works much better than Bottom-Up and ViT for all datasets except for ManyNames, for which Bottom-Up works better. This is most probably due to the fact that Bottom-Up was trained exactly on the same kind of data as that contained in ManyNames (namely, images from VisualGenome). That being said, the performance of CLIP is not terrible for this dataset (it obtains 0.29 MRR, compared to 0.40 for Bottom-Up).

For the kind of norms used in Cognitive Science, Bottom-Up works really badly but shows a gradient that makes sense: from really low performance for line drawings (MRR 0.03-0.04, near bottom) through low performance for colored drawings (0.10-0.18) to reasonable performance for photos (0.29).[9] The same gradient is shown by ViT, with very low performances on line drawings (MRR 0.04-0.06) and generally higher –and similar to Bottom-Up– performances on colored pictures (0.09-0.18), scoring 0.18 on photos. Given that ManyNames images were part of the Bottom-Up training set, making this dataset an easy testbed for Bottom-Up, it is remarkable that the overall performances of Bottom-Up on the other image sets are comparable to those of ViT, with the exception of the BOSS dataset. That is, a costly supervised training regime leads to better performances only on stimuli that look much like the training dataset, without gain in generalization. On stimuli of different kinds, a model trained without image labels can achieve the same result.

Overall, CLIP shows a much better and more stable behavior (range: 0.36-0.48). However, there is an intriguing result: it doesn't show the expected gradient (line $<$ colored drawings $<$ photos), but performs best on colored drawings (0.42-0.48) and worst on photos (0.36). We have no explanation for this pattern at present. The average result obtained by CLIP across datasets is 0.39, which means that the top name of an object is ranked, on average, within the top 3 predictions. While this is a notable performance, we next investigate whether using linguistic representations leads to better results.

Analysis 2: Visual vs. Linguistic Representations

In our second analysis, we focus on the best model, CLIP, and compare the performance of its visual and linguistic representations. The visual representations (CLIP-V) are obtained as in the first analysis. The linguistic representations are afforded by the textual encoder, CLIP-L, which works with textual inputs.

There are different ways of obtaining linguistic representations, depending on which input is given to CLIP-L. We tested 8 textual kinds of input: the name (e.g., "penguin"), adding a determiner ("some penguin"),[10] specifying that it is an image ("an image of some penguin"), and specialized templates for each type of stimulus (e.g. "a photo / colored drawing / line drawing of some penguin"). Results on the UNION vocabulary show that there are no major differences between the different prompt templates (range: 0.50-0.56); and the generic template "an image of some X" worked best.[11] We will use this template for the present analysis.

We evaluate the performance of the representations using the VG and THINGS data.[12] Recall that THINGS separates concepts for homonymous and polysemous names such as "bat"; we exclude them from the results for visual representations, and carry out a separate evaluation below.

The results, summarized in Table 4, show that linguistic representations are clearly superior to visual representations for all image types, by a large margin: for VG, CLIP-L obtains on average 0.60 MRR, CLIP-V 0.39; for THINGS, the results are 0.64/0.55, respectively. Remarkably, the results for CLIP-L are quite good also in absolute terms, with an average of 0.60 and 0.64 for VG and THINGS, respectively. This is especially so for the "traditional" psycholinguistic norms, where CLIP-L obtains MRR scores between 0.53-0.76 (VG)

of accuracy because of the differences in the size of the vocabularies, which range from 874 for VGMN to 3,231 for UNION. For completeness, accuracy is reported in the Appendix (see OSF repository). The overall patterns do not change.

[9] Qualitative analysis revealed that Bottom-Up tends to give uniform predictions for drawings: for line drawings, it often predicts "power lines", and for colored drawings, "parachute".

[10] We use "some" instead of "a/an" because our names include mass nouns like "water" and plural nouns like "blinds", for which using "a" would be ungrammatical.

[11] We provide detailed results in the Appendix available in the OSF repository.

[12] VG is from the same source as VGMN, and larger.

Image Type	Norms	VG			THINGS		
		N	CLIP-V	CLIP-L	N	CLIP-V	CLIP-L
line	S&V	168	0.41	**0.68**	230	0.47	**0.64**
	IPNP free	146	0.34	**0.53**	184	0.48	**0.58**
color	R&P	168	0.47	**0.76**	227	0.66	**0.75**
	IMABASE	189	0.42	**0.72**	244	0.67	**0.78**
	MultiPic	406	0.39	**0.61**	460	0.60	**0.65**
photo	BOSS	634	0.37	**0.62**	940	0.55	**0.64**
real. photo	ManyN	25K	**0.35**	0.31	23K	0.40	**0.42**
average			*0.39*	***0.60***		*0.55*	***0.64***

Table 4: Mean Reciprocal Ranks for representations generated with CLIP-V and CLIP-L, evaluated on the VG and THINGS data. Information as in Table 3.

and 0.58-0.78 (THINGS).[13] As in Analysis 1, CLIP-L results are worst for ManyNames (0.31 VG, 0.42 THINGS); in fact, for ManyNames in the VG space the representations of CLIP-V are superior to the linguistic ones (0.35). In both modalities, we find the same intriguing pattern of results as in the previous sections, with the best results obtained for colored drawings.

Finally, we present a focused evaluation for ambiguous (homonymous/polysemous) names in the THINGS data. There are 63 images with ambiguous names in the norms; we manually disambiguate the concept corresponding to each image (assigning them to, e.g., BAT1 for animals and BAT2 for the sports tool). It could be expected that CLIP-V would perform better on ambiguous names, due to its ability to distinguish between the two concepts associated to them, as it has different visual prototpes for each (whereas CLIP-L only has a single representation). Instead, CLIP-L still wins (MRR 0.55, vs. 0.51 of CLIP-V; there is, as expected, a drop in performance for both models). This could be due to the fact that the word representations (e.g., for "bat") contain information from the two senses, both linguistic and visual, but further research should probe this hypothesis.

Discussion and Conclusion

In this paper, we have assessed the potential of visually informed computational representations to model human naming behavior, in particular naming responses to visual stimuli. Recall from above that, due to our focus on the use of off-the-shelf models as is, we use three models that, while representative of their respective types, differ in other aspects that are known to affect performance of deep learning models (such as the amount of training data and the number of parameters). Future research should assess the extent to which the differences in model performance found in this paper generalize.

We find, in line with results in language and vision tasks (Radford et al., 2021), that the multi-modal model CLIP greatly outperforms both the object classification model Bottom-Up and the self-supervised model ViT, which were trained in a uni-modal setting (with only images as input) and a less cognitively rich task (object classification or image reconstruction, as opposed to caption-image matching). As expected, the self-supervised model ViT performs worst. Self-supervised models are valuable for Cognitive Science because they afford language-independent representations. It is however encouraging that ViT performs better (though still with low MRR values of around 0.18) for naming datasets using realistic pictures, as Cognitive Science has been steadily moving towards this kind of stimulus.

The superiority of CLIP tentatively suggests that linguistic information provides models with crucial clues as to what is common to different objects that cannot be inferred based purely on visual features. For instance, linguistic information may help capture the fact that different types of coffee makers, though looking different, have the same function. Moreover, we find that using the linguistic encoder to obtain name representations performs much better than using an abstraction over representations obtained through the visual encoder.[14] This is actually good news for applications of these models in Cognitive Science, as linguistic representations are easy to obtain —a simple model query is enough (visual representations do not scale up as easily). We consider the absolute performance of CLIP-L to be very good: around 0.60-0.64 MRR, meaning that it tends to rank the most frequently produced name for a given image between first (MRR 1) and second (MRR 0.5), in vocabularies that contain around 2000 names.

Overall, thus, our results suggest that current Computer Vision and Language and Vision models hold great promise for psycholinguistic research on object naming and, possibly, for other research in Cognitive Science involving visual stimuli. Moreover, depending on the application, off-the-shelf models can be used without further adaptation via e.g. fine-tuning. This means that a much larger pool of cognitive scientists can work with these representations than would be the case if training or adapting a model was needed.

[13] Note that the MRR scores suggest that results are overall better when building name representations with THINGS images as opposed to VG images. However, the results cannot be compared directly because they are computed on different sets of stimuli from the norms (compare the two columns "N" in the table).

[14] Recall that using CLIP enables linguistic name representations and visual image representations to be directly compared, because they live in the same space.

Acknowledgments

This research has been partially funded by grant PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033, funded by the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (Spain).

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6077-6086.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020, oct). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*(1). Retrieved from https://doi.org/10.1038%2Fs41467-020-18946-z doi: 10.1038/s41467-020-18946-z

Bonin, P., Poulin-Charronnat, B., Duplessy, H. L., Bard, P., Vinter, A., Ferrand, L., & Méot, A. (2020). Imabase: A new set of 313 colourised line drawings standardised in french for name agreement, image agreement, conceptual familiarity, age-of-acquisition, and imageability. *Quarterly Journal of Experimental Psychology*, *73*, 1862 - 1878.

Brochhagen, T., Boleda, G., Gualdoni, E., & Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. *Science*, *381*, 431 - 436. Retrieved from https://api.semanticscholar.org/CorpusID:260202674

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (boss), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE*, *5*.

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (boss) phase ii: 930 new normative photos. *PLoS ONE*, *9*.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proc. of ieee int. conf. comp. vis.* (pp. 9650–9660).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale.*

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). Multipic: A standardized set of 750 drawings with norms for six european languages. *The Quarterly Journal of Experimental Psychology*, *71*, 808-816. doi: 10.1080/17470218.2017.1310261

Goodkind, A., & Bicknell, K. (2018, 01). Predictive power of word surprisal for reading times is a linear function of language model quality. In (p. 10-18). doi: 10.18653/v1/W18-0102

Gualdoni, E., Brochhagen, T., Mädebach, A., & Boleda, G. (2023). What's in a name? A large-scale computational study on how competition between names affects naming variation. *Journal of Memory and Language*, *133*, 104459.

Günther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2023). Vispa (vision spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychological Review*, *130*(4), 896.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition.*

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019, 10). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, *14*(10), 1-24. Retrieved from https://doi.org/10.1371/journal.pone.0223792 doi: 10.1371/journal.pone.0223792

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*.

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better imagenet models transfer better? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2661–2671).

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... Li, F.-F. (2017, 05). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). *Evaluating (and improving) the correspondence between deep neural networks and human representations.*

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Icml.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.. Retrieved from https://api.semanticscholar.org/CorpusID:160025533

Rossion, B., & Pourtois, G. (2004). Revisiting snodgrass and vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*, 217 - 236.

Shriberg, E., & Stolcke, A. (1996). Word predictability after hesitations: a corpus-based study. In *Proceeding of fourth international conference on spoken language processing.*

icslp '96 (Vol. 3, p. 1868-1871 vol.3). doi: 10.1109/ICSLP.1996.607996

Silberer, C., Zarrieß, S., & Boleda, G. (2020). Object naming in language and vision: A survey and a new dataset. In *Proceedings of the 12th language resources and evaluation conference (lrec)* (pp. 5792–5801). Marseille, France: European Language Resources Association.

Silberer, C., Zarrieß, S., Westera, M., & Boleda, G. (2020, December). Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1893–1905). Barcelona, Spain (Online): International Committee on Computational Linguistics.

Singh, P., Peterson, J. C., Battleday, R. M., & Griffiths, T. L. (2020). End-to-end deep prototype and exemplar models for predicting human behavior. In *Proceedings of the 42nd annual conference of the cognitive science society.* Cognitive Science Society.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology. Human learning and memory*, *6 2*, 174-215.

Székely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., ... Bates, E. A. (2004). A new on-line resource for psycholinguistic studies. *Journal of memory and language*, *51 2*, 247-250.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2023). *Attention is all you need.*

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). *The unreasonable effectiveness of deep features as a perceptual metric.*