CONCEPTUAL BEHAVIOR AND HUMAN-LIKENESS IN VISION-AND-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning conceptual structures requires acquiring knowledge of how members of a class share a set of semantic properties. The challenge is that some properties are more efficiently learned by perceptual experience (e.g., an image of a dog that shows its texture, shape and color) while others benefit from language input (e.g., "a dog is a mammal"). Unimodal machine learning systems, as opposed to human brains, are therefore fundamentally limited in this respect. In contrast, systems integrating multimodal information should be able to learn a more human-like representational space since they can leverage both types of complementary sources of information. Multimodal neural network models offer a unique opportunity to test this hypothesis. We evaluate this proposal through a series of experiments on architecturally diverse vision-and-language networks trained on massive caption image datasets. We introduce an analytic framework that characterizes the semantic information behind the discrimination of concepts (i.e., lexicalized categories) through image-text matching tasks and representational similarity. We further compare how this discrimination (i.e., the model's "conceptual behavior") differs from that of humans and unimodal networks, and to what extent it depends on the multimodal encoder mechanism. Our results suggest promising avenues to align human and machine representational invariants via multimodal inputs.

1 INTRODUCTION

Efficient concept learning in humans requires integrating multimodal information. Human concepts are lexicalized categories (i.e., "named" representations that share a set of invariances). Lexicalization enables the grouping of concept features that are available through different senses under a single name (Sloutsky & Deng, 2019). For example, the visio-perceptual invariances that humans learn to associate with a cheetah (e.g. black spots, four legs) can be linked to knowledge about that concept that is best transmitted through language (e.g. "is a carnivore"). This results in a more comprehensive and structured semantic knowledge base about the entities in the world (i.e. commonsense-knowledge), which is essential for higher-level cognitive abilities like reasoning, planning or inference (Yee, 2019).

Recently, there has been increasing attention to the development of multimodal neural networks, with the aim of advancing models that possess such enhanced knowledge capable of generalizing to a variety of downstream tasks (Du et al., 2022). Among these are Vision-and-Language models (V+L), which jointly process images and text. Although there is recent work on the alignment performance and cross-modal integration of V+L networks, we still lack an understanding of whether these networks learn more human-like semantic representations and whether they use these for downstream tasks that require conceptual knowledge (e.g., image generation, concept detection, visual question answering, commonsense reasoning). For this, it is important to analyze the *conceptual behavior* (i.e., the ability to discriminate between concepts) of these models. This approach is especially relevant to the interpretability and understanding of deep learning models, where we care to explain the invariances that have been learned to perform a certain task.

In this work we set out to investigate these questions. We introduce an evaluation framework based on image-text matching tasks and representational similarity analysis that: **(a)** Examines the conceptual behavior and learning of V+L networks and measures their similarity to that of humans; and **(b)** Probes the extent to which V+L models are able to recognize the semantic features related to a

particular concept, and crucially measures whether the recognition of these properties is involved in the detection of such concept.

2 Related Work

Most studies aiming to understand the inner workings of V+L models have examined if and how these models combine cross-modal information (Aflalo et al., 2022; Cafagna et al., 2021; Cao et al., 2020; Frank et al., 2021; Hessel & Lee, 2020; Salin et al., 2022; Shekhar et al., 2017). They found that V+L models may not integrate multimodal information in an optimal manner (Hessel & Lee, 2020), and their performance seems to rely predominantly on one or the other modality depending on the task at hand (Cao et al., 2020; Frank et al., 2021; Salin et al., 2022). Although we indirectly examine multimodal integration via changes in conceptual behavior, the main focus of the present work is to determine how multimodality affects the encoding of conceptual knowledge and to what extent the latter is structured in a human-like way.

Another set of studies has focused on measuring the grounding capabilities of V+L regarding linguistic information, such as the recognition of the verb implied by an image (Hendricks & Nematzadeh, 2021), the number of entities shown (Parcalabescu et al., 2021), the semantic congruence between the caption and the image (Lindström et al., 2021), and the influence of negation in understanding the scene semantics (Dobreva & Keller, 2021). Unlike these studies, we are not concerned with how well V+L models can recognize if the events showcased in an image match a textual description. Instead, we focus our analyses on quantifying how these models recognize and represent general semantic knowledge about the properties of concepts depicted in a set of images, beyond what is observable in a particular instance (e.g. can the network recognize that a photo of a cheetah shows "something that is fast" even if the animal is in a sitting position?).

Studies in the field of interpretability of large language models share some of our aims, although they differ markedly due to their lack of multimodality. These set out to investigate how BERT and RoBERTa encode knowledge about the semantic properties of common concepts (Apidianaki & Soler, 2021; Weir et al., 2020) and found good overall retrieval but the performance depended on the type of feature. Collell Talleda & Moens (2016)'s work is of particular relevance, since it investigated how semantic features can be decoded from both linguistic and vision embeddings separately, and compared the informativeness of these representations for different types of features. Our work overcomes some of its limitations, as we are able to investigate how multimodal training changes the representation of semantic features by analyzing joint language-vision embeddings.

Finally, since our study tests models in their ability to recognize semantic properties that are not commonly described in caption-imaging datasets, it complements existing work on the robustness and out-of-distribution generalization of V+L models (Li et al., 2020; Zhou et al., 2022).

3 Evaluation Framework

3.1 Models

We examined three V+L models that are representative of common multimodal architecture choices, and differ in how early and under which mechanisms vision and language inputs are integrated. Comparing these models is thus an indirect way of testing if any particular multimodal training mechanism leads to more human-like conceptual behavior.

CLIP (Radford et al., 2021). The first examined model consists of a dual encoder architecture that "superficially" integrates multimodal information by projecting the final embeddings of a visual and text encoder to the same semantic space via a dot product. Different versions of CLIP exist with varying types of image encoders, with a fixed choice of a GPT-2 text encoder (Radford et al., 2019). To be similar to the other models probed in this work, we chose to examine the version composed of the visual transformer architecture ViT-B/32. The model was trained using a dataset of 400 million image-text pairs that is not publicly available.

ViLT (Kim et al., 2021). In contrast to CLIP, ViLT is a so called "fusion" single-stream model that integrates cross-modal information from its first stages of processing. It consists of a stack of

transformer modules that are initialized with the weights of a pre-trained ViT-B/32, and process the concatenation of independently extracted visual and text embeddings. ViLT was trained with an image-text matching (ITM) and a masked-language-modeling task, using as inputs approximately 5 million image-caption pairs that are extracted from a combination of public datasets.

ALBEF (Li et al., 2021). The last examined model combines the dual stream and fusion approach. ALBEF first processes image and text separately through a vision and a language encoder and aligns their final embeddings using a contrastive loss. The visual backbone is initialized with the weights of a ViT-B/16, while the language encoder is based on the first 6 layers of a BERTbase model (Devlin et al., 2018). After alignment, a multimodal encoder, initialized with the weights of the last six layers of BERTbase, fuses the visual and language embeddings of the unimodal streams by means of crossmodal attention. ALBEF is trained using the same datasets and tasks as ViLT, with the additional supervision of a momentum model.

3.2 DATASETS

To probe the conceptual behavior and semantic knowledge of V+L models, we combined a linguistic and visual dataset by mapping a series of images depicting a concept with a series of captions describing the semantic properties of said concept.

McRae feature norms. The database of McRae et al. (2005) contains human-annotated semantic properties of 541 concepts. Based on previous studies (Devereux et al., 2014), we categorized these features into those describing: **(1)** taxonomic information (e.g. "a dog is an animal"), **(2)** visual information (color, parts, or surface properties of the concept; e.g. "a dog has ears"), **(3)** non-visual perceptual information (sound, smell, taste, etc.; e.g. "a dog barks"), **(4)** functional information (related to the functions of the concept; e.g. "a dog can be used for protection"), **(5)** encyclopaedic information (common knowledge about the concept which cannot be included in the other categories; e.g. "a dog is domestic"). We avoid sense ambiguity by mapping these concepts to their corresponding synsets. After filtering the concepts of this database by those that overlap with the image dataset, the average number of features per concept was 13.83 (min. value of 6; max. value of 39).

THINGS images. The original dataset of Hebart et al. (2019) contains a sub-selection of images from ImageNet depicting 1854 object concepts. We chose this dataset because: **(1)** for each image the object occupies most of the canvas, with very limited and partial information in the background; and **(2)** when fed to a CNN, the selected images maximize the intra-class and minimize the inter-class similarity, while still demonstrating variability in their low-level properties; and **(3)** the synsets corresponding to each depicted object are annotated. After computing the overlap between the THINGS object concepts and those provided by McRae et al. (2005), our final image-property dataset was composed of 342 concepts and 4928 images depicting them (average value of 14.4 images per concept; min. value of 12; max. value of 27).

3.3 DIAGNOSTIC METHODS

Inspired by previous work in cognitive neuroscience (Kriegeskorte et al., 2008) and AI research (Morcos et al., 2018), our study provides insights on the similarity between V+L models, human cognition and unimodal networks, by comparing the distances between their concept-related representational structure. We specifically set out to answer the following questions:

Is the discriminability of concepts in V+L models different from their unimodal counterparts? If multimodal learning has changed the semantic invariances learned to identify a concept, the representational structure of categories in V+L models should be different from that of unimodal networks. To answer this question, we pair the stream of every V+L model M_i with its unimodal counterparts $(U_j, ...)$, $\mathcal{P} = \{(M_i, U_j), ...\}$, and compare their representational space (see Table 4 in the appendix for details on how models were paired). For every transformer block l of every model $m \in (M_i, U_i)$, we compute the cosine distance:

$$\mathbf{1} - \frac{\boldsymbol{c}_i \cdot \boldsymbol{c}_j}{\|\boldsymbol{c}_i\|\|\boldsymbol{c}_j\|} \tag{1}$$

where $\boldsymbol{c_i}, \boldsymbol{c_j} \in \mathbb{R}^n$ are the activation vectors of the ith and jth concepts in $\mathcal{C}$, and n is the number of neurons in a token of l. We examine the [CLS] or [EOS] tokens that provide an image or sentence-level representation and serve as the inputs of the ITM heads or the projection to the contrastive embedding space. In the case of the image streams of CLIP and ALBEF, and the unimodal vision networks ViT-B/16 and ViT-B/32, $\boldsymbol{c}_i = \frac{1}{p}\sum_{k=1}^{p} \boldsymbol{v}_i^k$. That is, $\boldsymbol{c}_i$ is obtained by averaging the activations $\boldsymbol{v}_i$ over the p images depicting the ith concept in $\mathcal{C}$. For the text streams of CLIP and ALBEF, $\boldsymbol{c}_i$ is the activation vector of the word representing the ith concept. To obtain meaningful categorical representations from BERT and GPT-2, we computed $\boldsymbol{c}_i$ by averaging the activation vectors of prompt-engineered sentences containing the word that references the ith concept (e.g. for the concept *dog*: "What is a dog?", "A photo of a dog", "This is my dog").

We thus obtained a vector of distances $\boldsymbol{d}_m^l \in \mathbb{R}^n$, where $n = \binom{|\mathcal{C}|}{2}$, for each model m (unimodal or multimodal) and layer l . To compare the representation of concepts in each model, we compute a Spearman's rank correlation coefficient $r_{k,q}^{i,j}$ between every pair of layers (l_k, l_q) for each model pair $(m_i, m_j) \in \mathcal{P}$, with $i \neq j$:

$$r_{k,q}^{i,j} = \frac{\mathrm{cov}(R(\boldsymbol{d}_i^k), R(\boldsymbol{d}_j^q))}{\sigma_{R(\boldsymbol{d}_i^k)}\sigma_{R(\boldsymbol{d}_j^q)}} \tag{2}$$

where R is the rank of $\boldsymbol{d}$ and σ is standard deviation.

Do V+L models learn to lexicalize categories? If V+L models are able to learn "named" categories, they should be able to recognize the word humans use to reference the object depicted in an image. We thus quantified how well each of the 4928 images matched the words associated with the concepts in $\mathcal{C}$. This procedure gave 4928×342 independent scores to analyze per model $m \in \mathcal{M} = \{\text{CLIP}, \text{ViLT}, \text{ALBEF}\}$. The matching scores of CLIP were obtained by computing the dot product between the final image and text embeddings, while the matching scores for ViLT and ALBEF were computed using the pre-trained ITM heads of the models. Since we are interested in the match between the images depicting c_i and the word referring to c_j for every ith and jth concept pair in $\mathcal{C}$, we examined pairwise logits.

To what extent is the conceptual behavior of V+L models explained by the invariances learned by unimodal models? To investigate to what degree the conceptual behavior of V+L models is accounted for by what can be learned through unimodal systems, we analyzed how similar the discriminability of concepts in V+L models is to the distances $\boldsymbol{d}_m$ between categories in the representational space of unimodal models (computed in step 1). The discriminability vector $\boldsymbol{s}_h \in \mathbb{R}^n$, where $n = \binom{|\mathcal{C}|}{2}$ and $h \in \mathcal{M} = \{\text{CLIP}, \text{ViLT}, \text{ALBEF}\}$, was obtained by computing the match scores (logits) between the various images depicting c_i and the word referring to c_j, averaging over images, and concatenating these n averages (for every ith and jth concept pair in $\mathcal{C}$). We then computed a Spearman's rank correlation coefficient (eq. 2) between each $\boldsymbol{s}_h$ and every $\boldsymbol{d}_m^l$, where each m is a unimodal counterpart of h and l is the layer of model m.

How similar to humans is the conceptual behavior of V+L models? To quantify the similarity between the conceptual behavior of V+L models and that of humans, we first constructed vectors representing the conceptual behavior of humans as follows. We format the McRae feature norms as a matrix where the columns are semantic features and the rows are the concepts; cell i, j contains the number of people who named feature j as belonging to concept i. For every semantic feature type t (e.g., taxonomic, visuo-perceptual) we select the corresponding feature set (columns), compute the cosine distance (eq. 1) between every pair of concepts (rows), and concatenate them into a vector $\boldsymbol{u_t}$. For every pair of different feature sets t_i, t_j we have $t_i \cap t_j = \emptyset$. Finally we computed the Spearman's rank correlation coefficient (eq. 2) between $\boldsymbol{s}_h$ and $\boldsymbol{u}_t$ for every model $h \in \mathcal{M}$ and feature type t.

To quantify the extent to which the similarity between the conceptual behavior of V+L models and humans can be explained by the multimodal training over and above what can be learned by

unimodal models, we carried out a variance partitioning analysis. More specifically, we computed the r^2 of an Ordinary Least Squares (OLS) regression using unimodal vectors $\boldsymbol{d}_m^l$ and multimodal vectors $\boldsymbol{s}_h$ as independent variables, and a semantic vector $\boldsymbol{u}_t$ as the dependent variable, where l represents the layer in $\boldsymbol{d}_m$ with the highest correlation to $\boldsymbol{u}_t$. To obtain the unique variance between $\boldsymbol{s}_h$ and $\boldsymbol{u}_t$, we then subtracted from that result the r^2 of an OLS regression using the unimodal vectors as the only predictors for the dependant variable $\boldsymbol{u}_t$. We repeated this analysis for every feature type t, and model $h \in M$ where m is a unimodal counterpart of h.

How human-like is the conceptual learning of V+L models? Studies from the field of cognitive psychology have shown that lexicalization abilities are impacted by certain psycholinguistic properties of the concepts being learned (Balota et al., 2007). To examine if V+L models exhibit the same pattern, we quantified the similarity between the concept recognition accuracies of the V+L models, and human-annotated information about the frequency (Gimenes & New, 2016), familiarity (Brysbaert et al., 2019), prevalence (Brysbaert et al., 2019), average age of acquisition (Kuperman et al., 2012), concreteness (Brysbaert et al., 2014), imageability (Scott et al., 2019) and perceptual strength (Lynott et al., 2020) of the concepts in our database. We also directly quantified how similar the lexicalization accuracy of V+L models was with the performance of humans in a lexical decision task (Balota et al., 2007). More specifically, we correlated (see eq. 2) the vector $\boldsymbol{c}_m \in \mathbb{R}^{|\mathcal{C}|}$ of the top-1 concept-detection accuracies obtained in the image-text matching task, with a vector $\boldsymbol{o}_i \in \mathbb{R}^{|\mathcal{C}|}$ containing the ratings of the ith psycholinguistic variable, for every model $m \in \mathcal{M}$.

Can V+L models recognize the semantic features of concepts in an image? Since the conceptual behavior of V+L models can resemble that of humans without representing the same semantic invariances, we directly examined how V+L models can recognize the semantic properties of a concept depicted in an image. We computed the match scores (logits) between the various images depicting c_i and each of the semantic features of the McRae dataset $f \in \mathcal{F}$, for every ith concept in $\mathcal{C}$, and every model $m \in \mathcal{M}$.

Is the recognition of semantic features related to the ability to recognize a concept? To characterize the relationship between the recognition of semantic features and the recognition of concepts, we carried out three types of analysis:

(a) To examine if concepts that V+L models recognize as sharing more semantic properties are also recognized as more similar in the lexicalization task, we first constructed vectors representing the semantic knowledge of V+L networks for every concept $c_i \in \mathcal{C}$ by computing the match scores (logits) between the various images depicting c_i and every semantic feature $f \in \mathcal{F}$, and averaging over the images and concatenating to obtain a concept-vector $\boldsymbol{a}_i \in \mathbb{R}^{|\mathcal{F}|}$. We then computed the cosine distances between every concept pair $\boldsymbol{a}_i, \boldsymbol{a}_j \in \mathcal{C}$ and concatenated them into a vector $\boldsymbol{w}_m$ that we then correlated (see eq. 2) with the discrimination vector $\boldsymbol{s}_m$ for each model $m \in \mathcal{M}$.

(b) To study if the recognition of semantic properties that humans considered related to a concept are also the most predictive of the matching scores of such concept in the lexicalization task, we computed the mutual information between every pair of matching distributions $\boldsymbol{c}_i, \boldsymbol{f}_j \in \mathbb{R}^n$ where $\boldsymbol{c}_i$ is the vector of logits for the ith concept in $\mathcal{C}$, $\boldsymbol{f}_i$ is the vector of logits for the jth feature in $\mathcal{F}$, and n is the number of images in our dataset.

(c) To visualize how semantic and conceptual recognition abilities are related, we chose the 200 concept-feature pairs c_i, f_j with the strongest dependence as measured by mutual information, and created two sets of images: one containing the images where f_j was recognized as "present" (i.e. with logit values higher than zero), and the other containing the images where f_j was "absent". We then averaged the matching scores of c_i separately for each of these sets of images and visualized the results.

4 EXPERIMENTS

The conceptual representations of V+L models and their unimodal counterparts have a different geometry. As seen in Figure 5 in the appendix, our study found evidence that multimodal training changes the representation of categorical information. Of the networks studied, the image streams of CLIP and ALBEF retained more similarity with the conceptual spaces of their unimodal

counterparts, in comparison with the text streams. Multimodal streams, on the contrary, seem to share a similar amount of resemblance to the representational spaces of both vision and language unimodal networks, although of lesser magnitude than the visual streams discussed. In the case of ALBEF, this similarity sharply decreases for the last layers of the network, which could be interpreted as the encoder creating a unique conceptual space more abstracted from its unimodal inputs. In the case of ViLT (see Figure 1), the bimodal similarity is seen until the last layers of the model, observing higher correlation values with the middle layers of the unimodal visual encoder and with the later layers of the unimodal text encoder.

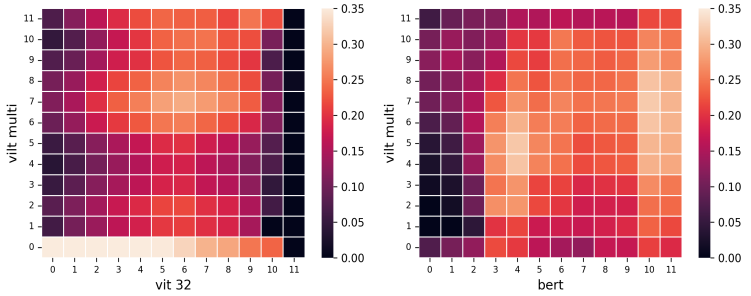

Figure 1: *Representational similarity between the conceptual spaces of ViLT and their unimodal counterparts*. See appendix for the comparison between all models.

V+L models are able to lexicalize categories with varying degrees of accuracy. CLIP achieved the best performance in the lexicalization task (Top-1 accuracy: 83%), followed by ALBEF (Top-1 accuracy: 81%), and ViLT (Top-1 accuracy: 51%). In more than 80% of the trials, the correct concept could be found within the top 10 highest-valued logits (Top-10 accuracy of CLIP: 98%; ViLT: 83%; ALBEF: 96%). This shows that all V+L models are able to represent the concepts in our database, and enables the characterization of these conceptual spaces carried out by the following analyses in this study.

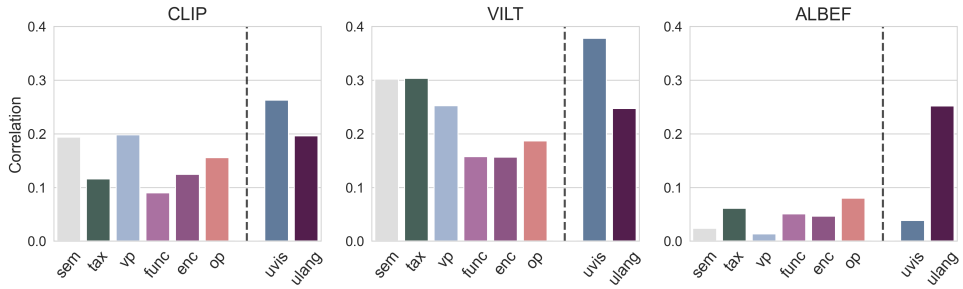

Figure 2: *Similarity between the conceptual behavior of V+L models and that of humans and unimodal networks*. sem: all semantic; tax: taxonomic; vp: visio-perceptual; func: function; enc: encyclopaedic; op: other-perceptual; uvis: unimodal visual model; ulang: unimodal language model.

The conceptual behavior of V+L can be partially explained by the invariances learned by unimodal models. It can be seen from Figure 2 that the conceptual behavior of ViLT, and to a lesser degree of CLIP, is moderately similar to the conceptual representations of vision and language unimodal encoders. Both networks show a higher resemblance to the intermediate layers of a visual system (layer 6 of ViT-B/32), in comparison to the similarity scores with the later layers of a language network (most similar for ViLT was layer 11 in BERT, while for CLIP was layer 9 in GPT-2), which relates to the findings comparing the representational structures of these networks. Surprisingly, ALBEF only showed moderate similarity values with the first layers of a language encoder (the most similar was layer 1 in BERT). This could be associated with the creation of a unique con-

ceptual space in the last layers of the multimodal encoder that was observed in the representational similarity analysis.

ViLT shows the conceptual behavior and learning most similar to humans. As shown in Figure 2, the conceptual behavior of ViLT shares the strongest resemblance to the conceptual behavior of humans, across all semantic feature types. Of these, taxonomic information exhibited the greatest similarity, closely followed by visio-perceptual information. We also found moderate degrees of human-like conceptual behavior in CLIP, especially in regards to visio-perceptual properties. On the contrary, the conceptual decisions of ALBEF showed little resemblance to that of humans. If anything, the semantic features with the highest similarity scores are those related to information preferentially transmitted via language, which could be explained by the higher correlation values between ALBEF and the conceptual spaces of language networks.

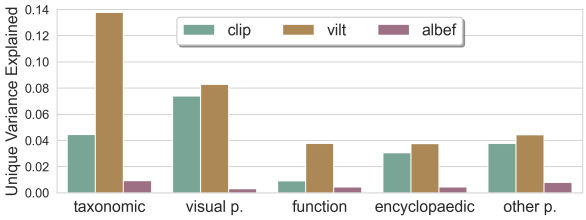

Figure 3: *Relationship between the conceptual behavior of V+L models and that of humans, after variance partitioning analysis.*

After conducting the variance partitioning analysis, we found evidence that ViLT creates conceptual spaces that go beyond the representational capabilities of unimodal networks, potentially the result of integrating multimodal information. As seen in Figure 3, CLIP also shows promising results for some types of semantic properties. On the contrary, the already low relationship between ALBEF and the conceptual behavior of humans was further decreased.

The analysis of conceptual learning gave supporting evidence for the similarity between concept processing in ViLT and humans. Of all V+L models, ViLT was the only whose concept-accuracy scores significantly correlated with those of humans in a lexical decision task, and with the ratings of concept frequency, prevalence, age of acquisition and perceptual strength (see 1). The imageability of concepts had a significant correlation with the concept predictions of both ViLT and CLIP, while concreteness ratings were significantly related to the conceptual decisions of ViLT and ALBEF. Concreteness and imageability are constructs that are predictors of recognition memory performance (i.e. how well a concept can be remembered) in humans (Khanna & Cortese, 2021), which has been linked to an increased capacity for generating and storing a mental image for such concept in the brain. Future work could explore if these correlations reflect how "strongly" some concepts are represented in a neural network.

Table 1: Correlation of concept detection accuracy scores with psycholinguistic variables.

	CLIP		ViLT		ALBEF	
PL Variable	*corr*	*p-val*	*corr*	*p-val*	*corr*	*p-val*
Lexical Decision Acc	-0.012	0.826	**0.200**	**0.000**	0.038	0.487
Frequency	-0.064	0.247	**0.342**	**0.000**	-0.005	0.927
Prevalence	-0.008	0.888	**0.181**	**0.001**	-0.031	0.579
Age of Acquisition	-0.033	0.550	**-0.317**	**0.000**	-0.030	0.583
Concreteness	0.102	0.064	**0.121**	**0.028**	**0.149**	**0.007**
Imaginability	**0.245**	**0.000**	**0.279**	**0.000**	0.088	0.159
Perceptual Strength	0.091	0.099	**0.180**	**0.001**	0.088	0.113

V+L are able to moderately recognize semantic features, to different extents depending on the model and the type of semantic feature. Similar to the concept detection task, CLIP achieved the

best scores in the recognition of semantic features, followed by ALBEF and later ViLT (see Table 2). CLIP's advantage was observable across all types of semantic properties. ALBEF and ViLT have a similar performance in the recognition of taxonomic, visuo-perceptual, and other-perceptual features. However, ALBEF was significantly better at recognizing functional and encyclopaedic information. More generally, taxonomic properties are the easiest to learn for all V+L networks, while other-perceptual ones are the hardest. The reason behind the taxonomic preference could be explained by the pre-training of the visual backbone of these networks since widely used benchmarks like ImageNet provide image labels that in some cases reflect taxonomic information. The Coverage Error scores were high for all models, indicating that these networks struggle with the detection of some subset of properties.

Table 2: Semantic Feature recognition performance.

	Recall	*CE*	*Tax*	*VP*	*Func*	*Enc*	*OP*
CLIP	**0.451**	**0.597**	**0.650**	**0.314**	**0.550**	**0.366**	**0.237**
ViLT	0.232	0.650	0.454	0.255	0.206	0.196	0.155
ALBEF	0.318	0.685	0.479	0.224	0.370	0.335	0.175

ViLT exhibits the most human-like relationship between semantic-feature recognition and conceptual discrimination. When examining if concepts that V+L models recognize as sharing more semantic properties are also recognized as more similar in the lexicalization task, ViLT exhibited the strongest relationship ($r = 0.438$), followed by CLIP ($r = 0.245$), and ALBEF displaying the lowest scores ($r = 0.078$). ViLT also had the highest mutual dependence scores between the values of the logits associated with the presence of a semantic feature, and the logits of the concepts that possess such property as rated by humans (see Table 3 and Figure in apendix). In other words, the recognition values of the features associated with a concept (e.g. "is an animal" for "dog"), are more predictive of how well the concept will be recognized by the model in the lexicalization task, in comparison with unrelated features. As Table 3 shows, both CLIP and ALBEF exhibited a similar pattern of results. However, ViLT's differences in the mutual information scores between semantic related and unrelated features were higher for all types of properties, except for "other perceptual" information for which CLIP exhibited a stronger difference. In all cases, taxonomic properties have the strongest mutual dependence scores with concept detection, while the order of differences in the other properties varies across models.

Table 3: Difference in mutual information between concept-related and unrelated semantic features.

	All	*Tax*	*VP*	*Func*	*Enc*	*OP*
CLIP	0.080	0.184	0.062	0.090	0.064	**0.076**
ViLT	**0.113**	**0.334**	**0.079**	**0.115**	**0.094**	0.063
ALBEF	0.028	0.076	0.016	0.036	0.030	0.020

Taking together, these findings suggest that, especially for ViLT, the semantic invariances learned to predict features are related to the invariances learned to predict concepts. However, this discovery doesn't explain the directionality of the effect. Figure 4 sheds some light on this aspect and shows how only for ViLT the recognition of a feature can both be useful to recognize the presence or absence of a concept (note the distribution centered around 0). For example, recognizing that there "is an animal" in an image showing an object in the sky, decreases the chances that an airplane will be detected. This behavior seems to be closer to how humans use feature knowledge for recognizing concepts. However, ViLT appears to still lack the ability to aid the recognition of concepts by the absence of a feature (e.g. "this is *not* an animal" and thus cannot be a bird).

5 Conclusions

Our work introduced an analytic framework that enabled the analysis of the conceptual and semantic knowledge encoded by V+L models, and how similar it is to that of humans. We demonstrated that although multimodal training changes the representation of concepts in deep learning models, this

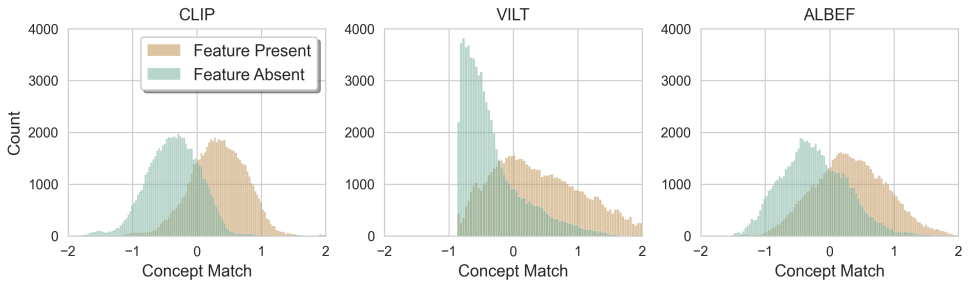

Figure 4: *The effects of feature detection on concept matching scores.*

transformation does not necessarily lead to a conceptual behavior that is more similar to humans in comparison to unimodal systems. In fact, we critically showed that models exhibiting the most accurate performance in concept or feature detection tasks are not the most similar to how humans process these kinds of information. This gives supporting evidence to the idea that artificial neural networks may be able to perform tasks with a human-like level of accuracy, but achieve such scores by learning invariances different from those encoded by the brain. We also found that certain types of semantic information are easier to learn for all models, but the reasons behind this remain speculative. Our findings thus highlight the need for understanding in depth what V+L models are learning to represent, and how different kinds of representations may affect the performance of artificial networks in a variety of cognitive tasks.

In addition, our study suggests that V+L models' architecture or training objective has a considerable impact on how these networks are able to represent concepts in a human-like manner. In particular, we found that a system that fuses early information from different modalities, with little to no unimodal processing, exhibits a conceptual behavior that is the most similar to humans. We also demonstrated that this type of network can exhibit complex relationships between the representations of concepts and semantic properties. Given that it is still debated how brains depend on the encoding and detection of semantic information for the perceptual and cognitive processing of concepts (and vice versa), this finding can be of particular importance for work at the intersection of neuroscience and AI.

6 Limitations

This study is empirical in nature and only probed three V+L models, and thus cannot make general statements about the inner workings and representational capabilities of multimodal deep neural networks. In addition, we cannot make conclusive claims on the reasons behind the differences between models in their performance or human similarity, since we do not provide a causal analysis. However, our main aim was to **(a)** investigate how information is represented in widely-used V+L models that currently form the basis of numerous applications and fine-tuned models, and **(b)** provide an analytic framework that can easily be applied to probe any available multimodal model. In addition, by investigating different types of architectures and training objectives, this study makes suggestions on which aspects of V+L models may be relevant for the encoding of semantic information in a human-like manner, which future work could explore in depth.

Our work also lacked an analysis of how the differences in conceptual or semantic representations of V+L models may impact the networks' performance on downstream tasks, especially those tackling higher-level cognitive abilities. We leave for future work this examination.

7 Reproducibility Statement

The open-source implementation of our work is publicly available at (anonymous). The repository contains a digital notebook to easily reproduce the reported results (figures and tables). Every step of the evaluation framework can be executed by running the source-code in a local environment. The results obtained at every intermediate step of the analysis (e.g. image-text matching tasks) can

be accessed here: (anonymous). All models and datasets probed in this study are open-source and can be freely examined.

REFERENCES

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21406–21415, 2022.

Marianna Apidianaki and Aina Garí Soler. All dolphins are intelligent and some are friendly: Probing bert for nouns' semantic properties and their prototypicality. *arXiv preprint arXiv:2110.06376*, 2021.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. The english lexicon project. *Behavior research methods*, 39(3):445–459, 2007.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911, 2014.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51(2):467–479, 2019.

Michele Cafagna, Kees van Deemter, and Albert Gatt. What vision-language modelssee'when they see scenes. *arXiv preprint arXiv:2109.07301*, 2021.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pp. 565–580. Springer, 2020.

Guillem Collell Talleda and Marie-Francine Moens. Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 2807–2817. ACL, 2016.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Radina Dobreva and Frank Keller. Investigating negation in pre-trained vision-and-language models. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 350–362, 2021.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

Manuel Gimenes and Boris New. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior research methods*, 48(3):963–972, 2016.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! *arXiv preprint arXiv:2010.06572*, 2020.

Maya M Khanna and Michael J Cortese. How well imageability, concreteness, perceptual strength, and action strength predict recognition memory, lexical decision, and reading aloud performance. *Memory*, 29(5):622–636, 2021.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, pp. 4, 2008.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990, 2012.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020.

Adam Dahlgren Lindström, Suna Bensch, Johanna Björklund, and Frank Drewes. Probing multimodal embeddings for linguistic properties: the visual-semantic case. *arXiv preprint arXiv:2102.11115*, 2021.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52(3):1271–1291, 2020.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559, 2005.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. In *AAAI 2022*, 2022.

Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, 51(3):1258–1270, 2019.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017.

Vladimir M Sloutsky and Wei Deng. Categories, concepts, and conceptual development. *Language, cognition and neuroscience*, 34(10):1284–1297, 2019.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. Probing neural language models for human tacit assumptions. *arXiv preprint arXiv:2004.04877*, 2020.

Eiling Yee. Abstraction and concepts: when, how, where, what and why?, 2019.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. Vlue: A multi-task benchmark for evaluating vision-language models. *arXiv preprint arXiv:2205.15237*, 2022.

A APPENDIX

A.1 COMPARISON BETWEEN DIFFERENT STREAMS OF V+L MODELS AND THEIR UNIMODAL COUNTERPARTS

Table 4: V+L and unimodal comparison

Model 1	*Model 2*
CLIP image stream	ViT-B/32
CLIP text stream	GPT 2
ViLT multimodal stream	ViT-B/32
ViLT multimodal stream	BERT
ALBEF visual stream	ViT-B/16
ALBEF text stream	BERT
ALBEF multimodal stream	ViT-B/16
ALBEF multimodal stream	BERT

A.2 REPRESENTATION SIMILARITY OF CONCEPTUAL SPACES IN V+L MODELS AND THEIR UNIMODAL COUNTERPARTS

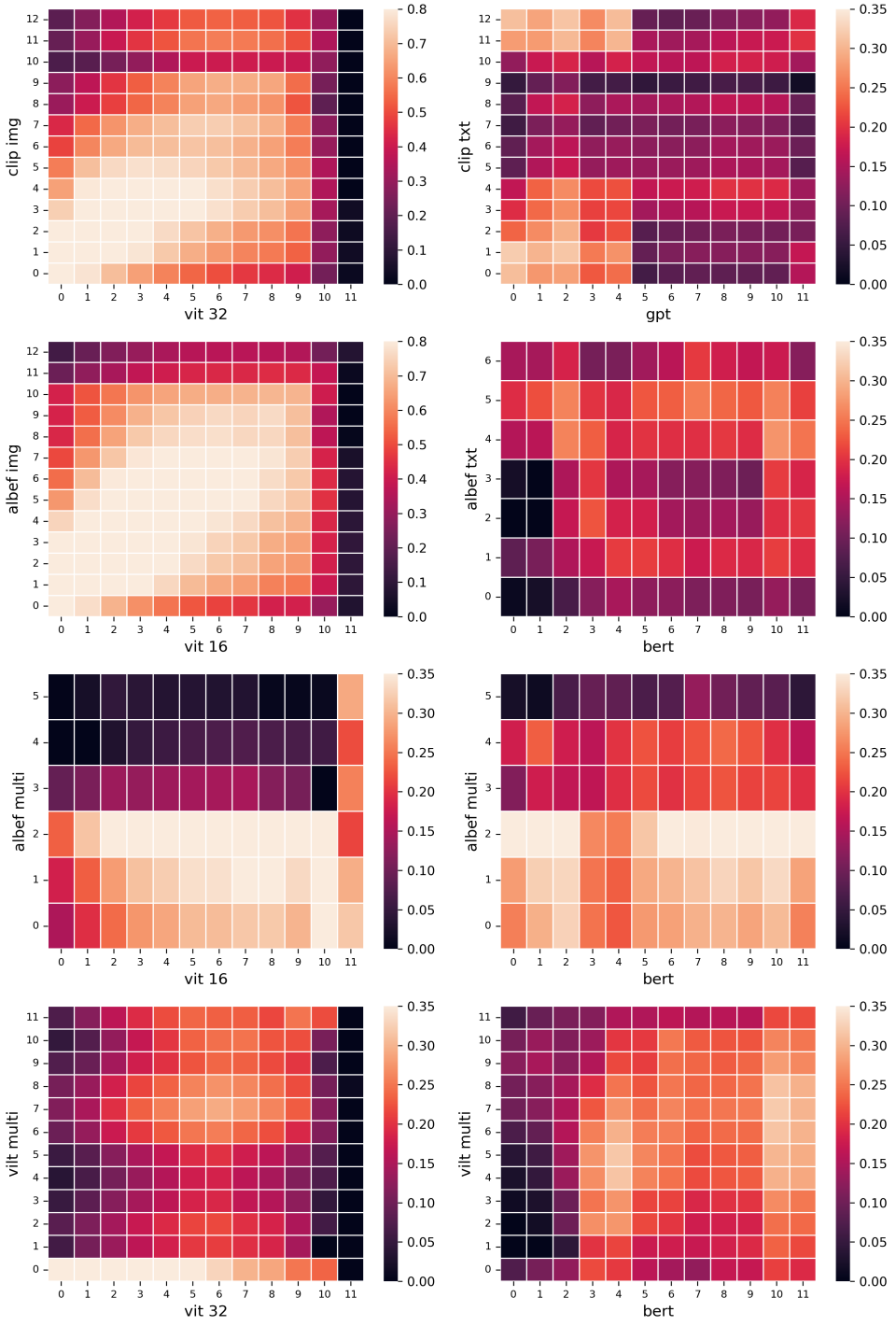

Figure 5: *Representational similarity between the conceptual spaces of V+L models and their unimodal counterparts*