Does language help generalization in vision models?

Benjamin Devillers[1], Bhavin Choksi[2], Romain Bielawski[1], Rufin VanRullen[1,2]

[1] Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
{firstname.lastname}@univ-tlse3.fr

[2] CerCO, CNRS UMR5549, Toulouse
{firstname.lastname}@cnrs.fr

Abstract

Vision models trained on multimodal datasets can benefit from the wide availability of large image-caption datasets. A recent model (CLIP) was found to generalize well in zero-shot and transfer learning settings. This could imply that linguistic or "semantic grounding" confers additional generalization abilities to the visual feature space. Here, we systematically evaluate various multimodal architectures and vision-only models in terms of unsupervised clustering, few-shot learning, transfer learning and adversarial robustness. In each setting, multimodal training produced no additional generalization capability compared to standard supervised visual training. We conclude that work is still required for semantic grounding to help improve vision models.

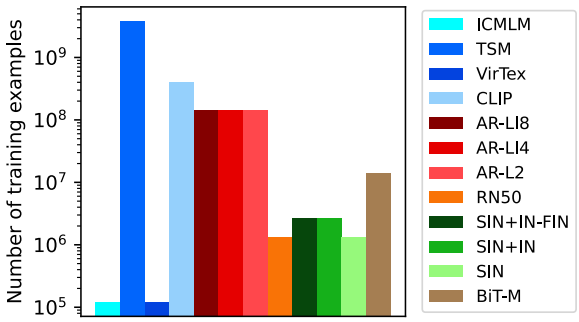

Figure 1: Size of the training dataset used by the models (number of images, in log scale). ICMLM and VirTex are trained on CoCo, TSM on HowTo100M, CLIP on a (not publicly available) scrape of the internet, RN50 is trained on ImageNet-1k, the AR models and SIN models are trained on augmented versions of ImageNet-1k and BiT-M is trained on ImageNet-21k.

1 Introduction

Learning vision models using language supervision has gained popularity (Quattoni et al., 2007; Srivastava and Salakhutdinov, 2012; Frome et al., 2013; Joulin et al., 2016; Pham et al., 2019; Desai and Johnson, 2021; Hu and Singh, 2021; Radford et al., 2021; Sariyildiz et al., 2020) for two main reasons: firstly, vision-language training allows to build massive training datasets from readily available online data, without manual annotation; secondly, language provides additional semantic information that cannot be inferred from vision-only datasets, and this could help with semantic grounding of visual features.

Recently (Radford et al., 2021) introduced CLIP, a language and vision model that shows outstanding zero-shot learning capabilities on many tasks, and compelling transfer-learning abilities. A recent report (Goh et al., 2021) showed that CLIP produces neural selectivity patterns comparable to "multimodal" concept cells observed in the human brain (Quiroga et al., 2005; Reddy and Thorpe, 2014). From these results, it is tempting to assume that CLIP's generalization properties stem from semantic grounding provided by the joint vision-language training.

Here, we show that CLIP and other vision-language models do not perform better than vision-only, fully supervised models on a number of generalization settings and datasets. Representation similarity (Kriegeskorte et al., 2008) analysis reveals that the multimodal representations that emerge through vision-language training are different from *both* linguistic and visual representations–and thus possibly unsuitable for transfer-learning to new visual tasks. In conclusion, additional work on linguistic grounding is still needed, if it is to improve generalization capabilities of vision models.

We provide our code for reproducibility[1].

2 Models

We use a number of publicly available vision, text or multimodal pretrained models, and compare their representations and generalization abilities.

[1] https://github.com/bdvllrs/generalization-vision

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 171–182
November 10–11, 2021.

To facilitate interpretation and comparisons between the models, Figure 1 reports the training dataset size for each of the visual models (including the vision-language models). They are all based on the same backbone (a ResNet50 architecture).

In CLIP, the authors train the joint embedding space of a visual network (hereafter called simply CLIP) and a language network (hereafter called CLIP-T) using contrastive learning on 400M image-caption pairs. Note that in the present paper, the visual backbone of CLIP is a ResNet50, even though the visual-transformer-based CLIP model could reach higher performance; this choice allows for a fair comparison with the other visual models that are all based on the ResNet50 architecture. In addition, we also consider TSM (Alayrac et al., 2020), another multimodal network trained with a contrastive loss on video, audio and text inputs from the HowTo100M dataset (Miech et al., 2019) (containing more than 136M video clips with captions. For training, the authors effectively used 120M video clips of 3.2s sampled at 10 fps). The effects of CLIP's and TSM's contrastive training paradigm can be compared with VirTex and ICMLM—two other recent multimodal networks. In VirTex, the visual feature representations are optimized for an image captioning task (Desai and Johnson, 2021), and for a text-unmasking task in ICMLM (Sariyildiz et al., 2020). Such text-based objectives aim to provide a form of linguistic grounding using significantly fewer images than CLIP (VirTex and ICMLM models are trained on the COCO dataset (Lin et al., 2014) with approximately 120K captioned images).

To understand the potential effects of linguistic training, we compare the multimodal networks to vision-only networks. We include a baseline architecture (ResNet50) trained on ImageNet-1K (He et al., 2016) (1.3M labelled images). Second, we consider a similar architecture (ResNet50 backbone) called BiT-M (Kolesnikov et al., 2019), trained on ImageNet-21K, a much larger dataset (14M labelled images).

While generalization and robustness properties can often be derived from access to large labelled image datasets (as in BiT-M), obtaining such labels is costly. An alternative is to train models with additional datapoints based on assumptions about the real-life data distribution–as done, e.g., with adversarial training. In this study, we use the Adversarially Robust (AR) ResNet50 models provided by (Engstrom et al., 2019b), trained on the 1.3M ImageNet training set plus 110 adversarial attacks of each image (i.e. more than 140M images overall). The different model variants (AR-L2, AR-LI4, AR-LI8) correspond to distinct adversarial attacks (refer to (Engstrom et al., 2019b) for more details). This adversarial training was found to produce more perceptually aligned features and to improve generalization (e.g. transfer learning) in some settings (Salman et al., 2020). Another such technique was used for StylizedImageNet (SIN) models (Geirhos et al., 2019), where a variant of the ImageNet dataset (1.3M images) was designed via style-transfer to specifically reduce the network's reliance on texture information. The authors provide weights for models that are (i) only pretrained for SIN images (SIN), (ii) trained on SIN and ImageNet (SIN+IN) combined, or where (iii) a SIN+IN model is finetuned on ImageNet (SIN+IN-FIN).

For the vanilla ResNet50, SIN, AR and BiT-M models, we use activations after the final average pooling operation as feature representations. Although all these models share a ResNet50 backbone, there are minor differences in their implementations. We assume that such small architectural differences would not dramatically affect the feature spaces learned by these models.

Finally, we also use two text-only language models, GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019), in our feature-space comparisons. As these models are not designed to process visual inputs, they cannot be tested on visual generalization; but we can use their representations of class *labels* (or sentence captions) as a basis for comparison with visual or multimodal network representations. In a similar way, the language stream of the CLIP model (CLIP-T) can be treated as a third language model for our comparisons.

3 Generalization tasks

In (Radford et al., 2021), CLIP was systematically tested in a zero-shot setting. However, this requires a language stream to describe the different possible targets, which is not available in standard vision models. To compare the generalization capabilities of multimodal and vision-only models, we thus focus on few-shot, transfer and unsupervised learning. In each case, we evaluate performance on MNIST (LeCun et al., 1998), CIFAR10, CIFAR100 (Krizhevsky et al., 2009), Fashion-MNIST (Xiao et al., 2017), CUB-200-2011 (CUB) (Wah et al.,

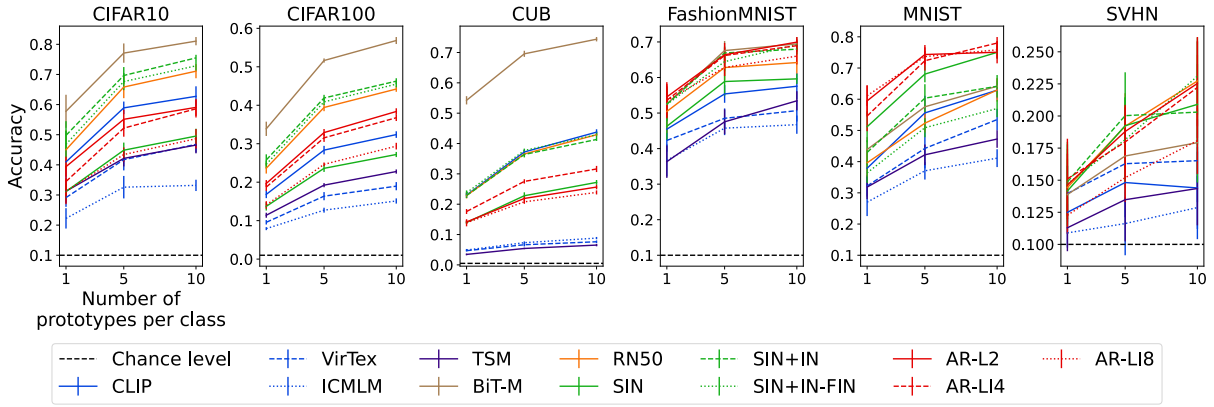

Figure 2: 1-shot, 5-shot and 10-shot accuracy over our evaluation datasets. Multimodal networks (ICMLM, VirTex, CLIP, TSM, in blue) have typically worse performance than the other models for all datasets.

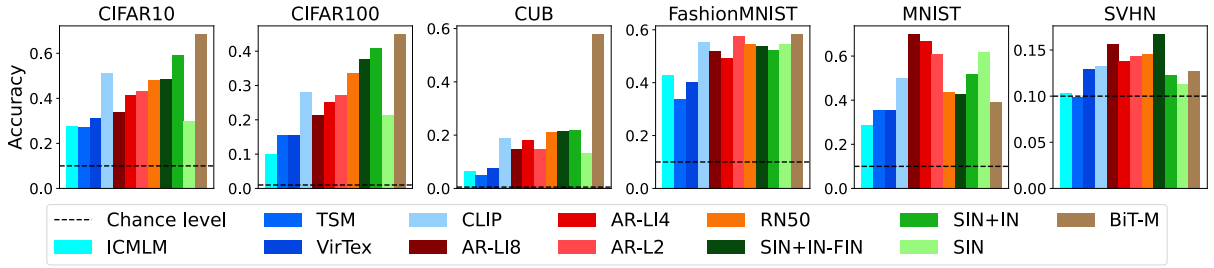

Figure 3: Unsupervised clustering accuracy over our evaluation datasets. Clustering is obtained using Scikit-learn Spectral Clustering algorithm. Multimodal networks (ICMLM, VirTex, CLIP, TSM, in blue) are worse than vision-only models (in various colors) on average.

2011) and SVHN (Netzer et al., 2011)[2]. These datasets test generalization capabilities for natural images of various classes.

3.1 Few-shot learning

As a first generalization experiment, we compare few-shot learning accuracy. For this experiment, we directly pass N randomly selected samples for each class (N-shot learning) through the pretrained models to obtain a feature representation for each sample. Then, we define a class prototype by averaging the feature representations of all the samples in each class. We measure the performance of vision-only and text-vision models for $N = 1, 5$ and 10. Each time, the reported performance is averaged over 10 trials with different class prototypes (i.e., different random selection of samples). Figure 2 shows the performance of each model on each dataset. For CIFAR10, CIFAR100 and CUB (all the natural images datasets), BiT-M has the best accuracy. On the other hand, ICMLM, VirTex, CLIP and TSM do not perform better than the vision-only models.

[2]For more details on these datasets, see appendix A.

Figure 5 shows the average performance of each model across datasets, in the leftmost 3 panels.

3.2 Unsupervised clustering

Our second generalization test is an unsupervised clustering task over the same datasets. For this, we apply an out-of-the-box spectral clustering algorithm (Pedregosa et al., 2011) using the cosine of two feature vectors as a metric. We provide the number of required clusters (number of classes) to the clustering algorithm: this ensures that all classes are represented by a cluster. The clusters are computed only on the test-sets.

To compute the accuracy on the prediction, we need to assign labels to each cluster. To do so, we use a greedy algorithm: we first choose the cluster containing the most elements in common with a given class and assign it the corresponding label. We then continue with the second cluster that has the most elements in common with another class, and so on until all clusters have been labelled.

Figure 3 shows the unsupervised clustering performance on individual datasets. It shows a similar ranking to the few-shot learning task where BiT has the best performance overall and the visio-

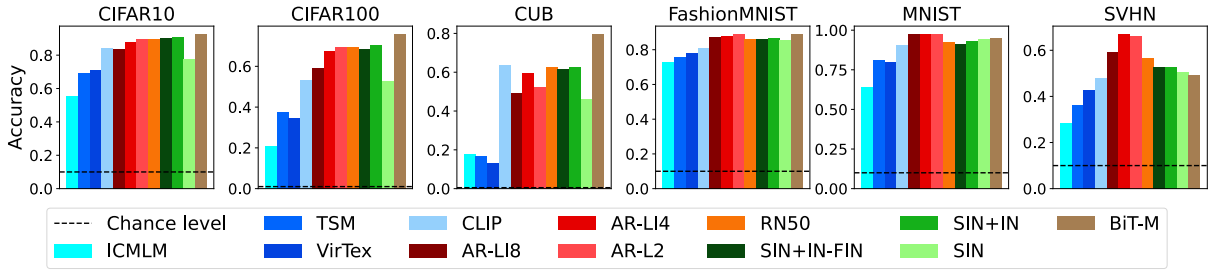

Figure 4: Transfer learning accuracy over our evaluation datasets. For each dataset and model, we train a linear layer to classify the models' visual features. Multimodal networks (ICMLM, VirTex, CLIP, TSM, in blue) have worse performance accuracy than vision-only models (in various colors).

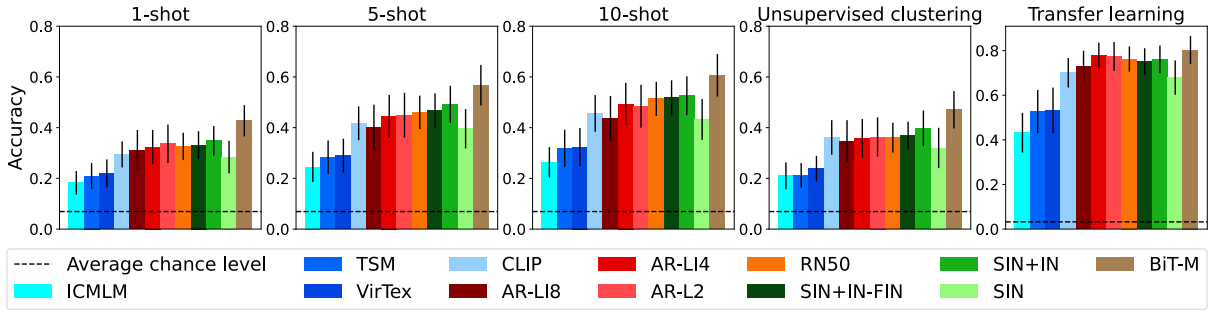

Figure 5: Average performance of the models across datasets, with standard error of the mean, for the various generalization tasks (few-shot learning, unsupervised clustering, transfer learning). Multimodal networks (ICMLM, VirTex, CLIP, TSM in blue) have worse generalization accuracy across all tasks.

linguistic models lag behind the vision-only models. Figure 5 panel 4 (from left) shows the performance of the unsupervised clustering algorithm averaged over all datasets.

3.3 Transfer learning

To further evaluate the models' generalization properties, we use a transfer learning setting as described in (Salman et al., 2020). We use the same datasets as in the other tasks, each time training a linear probe using the Adam optimizer. We train each linear probe for 20 epochs with a learning rate of 1e-3 and a weight decay of 5e-4.

Fig 4 shows the performance of the models on this task, separately for each dataset, and Fig 5 (rightmost panel) reports the average across datasets. Multimodal networks fail again to improve generalization.

3.4 Robustness to adversarial attacks

Another important test for generalization is the robustness to input perturbations (a form of out-of-distribution generalization). Here, we compare the adversarial robustness of different models against untargeted and targeted random projected gradient descent (RPGD) attacks (Madry et al., 2018). We use L_2 and L_∞ norms to distinguish any norm-specific effects. Figure 6 shows the success rate of the 100-step RPGD attacks on 1000 images taken from the ImageNet validation set. We use the foolbox API (Rauber et al., 2017) to perform all the attacks with configurations provided by (Engstrom et al., 2019a).

3.5 Summary

Overall, models trained with multimodal information (CLIP, VirTex, ICMLM, TSM) do not achieve better performance than the visual-only ResNet-based models. This systematic observation across multiple image datasets and generalization tasks (including few-shot, transfer and unsupervised learning, as well as adversarial robustness) goes against the assumption that linguistic grounding should help generalization in vision models.

Among the multimodal networks, CLIP does indeed appear to be more generalization-efficient than VirTex, ICMLM and TSM. As mentioned in (Radford et al., 2019), directly predicting highly variable text captions (as done in VirTex or ICMLM) is a difficult task that does not scale well. CLIP (and TSM) avoid generating text, relying instead on a contrastive loss between visual

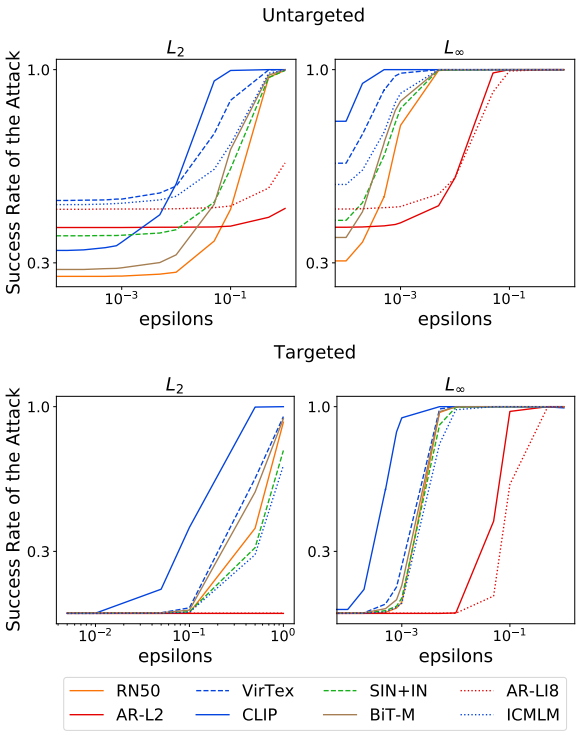

Figure 6: Robustness of some of the models to untargeted (top) and targeted (bottom) random projected gradient descent (RPGD) attacks for varying epsilons, with L_2 (left) or L_∞ norm (right). AR models are robust by design. Multimodal networks (CLIP, VirTex) are less robust than vision-only models (RN50, SIN+IN, BiT-M).

and linguistic embeddings. However, even with the potential benefits provided by this contrastive loss, CLIP (and TSM) do not outperform the vision models.

Finally, BiT-M, a simple vision-only model trained on a very large labelled dataset, turns out to be the overall best performing model for few-shot learning, unsupervised clustering and transfer learning, and on par with the standard ResNet50 for adversarial robustness.

Although these results are fairly consistent across datasets, there are still some differences.

For the CUB dataset, BiT-M largely outperformed the other models. This result is to be expected as the bird species in CUB are also part of ImageNet-21K labels. Then, among visio-linguistic models, CLIP is the only one competitive with the remaining visual models on this dataset.

MNIST and SVHN require classification of digits. According to (Radford et al., 2019), CLIP should be able to generalize to this task, as its training set included numerous images with text and digits. Indeed we observe that CLIP can perform

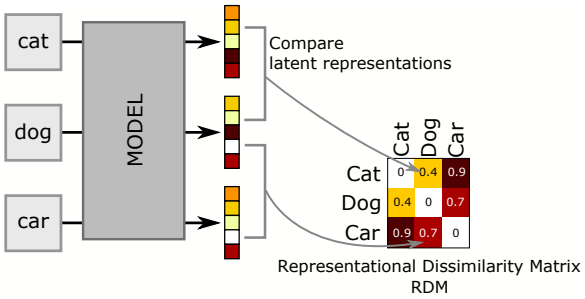

Figure 7: How to compute representational Dissimilarity Matrices (RDMs). RDMs are built from the model's embedding space. The RDMs can then be used for a Representational Similarity Analysis by comparing them using a Pearson Correlation.

as well as some of the vision models for these datasets. However, SIN and AR models perform generally better than other models.

For datasets with more natural images (CIFAR, FashionMNIST, CUB), vision models are generally better than their visio-linguistic counterparts.

4 Model comparison

To better understand the similarities and differences between the feature spaces learned by the various models, we now compare them using Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008).

Method RSA is a comparison method originally used to compare fMRI data. It allows us to compare different models (with different latent space dimensions, norms, ...) which share the same structure.

This works by comparing the models' *Representational Dissimilarity Matrix* (RDM). RDMs are obtained by computing the 2 by 2 distances for each class of the latent representations (see figure 7). More specifically, for each visual model, we define for each class the set $\mathcal{F}_c$ containing the feature vectors of all the images of class c, its average $\bar{f}_c$ and its standard deviation σ_c. The RDM matrix is then defined as $[RDM_{i,j}]$ where

$$RDM_{i,j} = \left\| \frac{\bar{f}_i - \bar{f}_j}{\sqrt{\frac{\sigma_i^2}{|\mathcal{F}_i|} + \frac{\sigma_j^2}{|\mathcal{F}_j|}}} \right\|_2 \quad (1)$$

for each pair of class (i, j).

We use the norm of the unequal variance t-test (Welch, 1947) as our distance metric between the latent representations, because it allows to normalize

the distances between class centroids with respect to their variances. Indeed, each class is represented by a cluster of latent vectors of different sizes.

In the case of language models (all transformer-based), we use as latent representations, the encoding of the sentence "a photo of x." where we replace "x" by the corresponding label. We then use the contextualization of the label as the text feature vector. Compared to the vision models, there is only one representation per class (only one sentence per class) hence a lack of variance associated with the feature vector of each class. As a result, the distance used in the RDM matrix becomes an L_2 norm.

The RDM matrix obtained with this method contains the respective distances between pre-defined concepts (in our case the 1000 classes of ImageNet). RDMs can therefore be considered as a standardized representation of latent spaces. This means that we can compare our models' representations by computing the Pearson correlation between their respective RDMs. The corresponding comparison matrix, for all pairs of models, is illustrated in Fig 8.

Results Figure 9 shows the results of a hierarchical clustering (a) or t-SNE (Van der Maaten and Hinton, 2008) embedding (b) of the RDMs using Pearson correlation as a distance. Looking at the dendrogram, all the vision-only models are very close to one another with a maximum distance <0.2. Then, multimodal models stand a bit further (CLIP, TSM, VirTex, ICMLM); and finally, CLIP-T and the language models (BERT, GPT2) are the furthest away. This indicates that the language supervision (contrastive embedding, text-generation or text-unmasking objectives) has changed the structure of the ResNet latent space for CLIP, TSM, VirTex and ICMLM models (respectively). Yet these multimodal models are not truly linguistic either, as they are very distant also from the standard language models.

This conclusion is also supported by the t-SNE plot, showing a cluster of BiT-M, RN50 and SIN vision models, a second cluster with the AR models, and further along the same direction, the multimodal networks (CLIP, VirTex, ICMLM, TSM). Note that, although this arrangement might suggest that multimodal networks possess adversarial robustness properties in common with AR models, this suggestion was not supported by our tests using actual adversarial attacks (Fig 6). Finally, the language models (BERT, GPT2 and CLIP-T) are separated from the rest, along a distinct direction. Overall, the analysis suggests that multimodal representations are neither visual nor linguistic, but surprisingly, *not really in-between either*[3]. This is surprising as we should expect that representations trained with access to both vision and language would derive information from both modalities, and consequently end up somewhere in-between purely visual and purely textual representations.

5 Performance on linguistic tasks

This suggestion might be further supported by evaluating the usefulness of the learned visual representations on *linguistic* tasks. According to the above findings, visual representations obtained via multimodal training may fare no better than vision-only representations. To test this, for each vision model, we collect the ImageNet features for each image class, and train a standard word embedding (Skip-Gram method) while constraining the class label words to these visual feature vectors. The resulting linguistic space will thus capture some of the structure of the vision model's latent space.

5.1 Method

Architecture We train Skip-Gram models (Mikolov et al., 2013a) on Wikipedia using the Gensim library (Řehůřek and Sojka, 2010). Before training, some of the embedding vectors (corresponding to the ImageNet class labels) are set to the latent representations of a vision model, and frozen during training. This training procedure forces the word embedding space to adopt a similar structure to the vision model's latent space (at least for the frozen words, i.e. the class labels).

Visual words We denote 'visual word embeddings' (resp. visual words) as the word embeddings (equivalent to the visual feature vectors) obtained from the vision models (resp. the associated word token) on ImageNet classes. Some of the classes are composed of multiple words (e.g. "great white shark"). We leverage the WordNet (Miller, 1998) structure of ImageNet classes to only keep the hypernym of the class that contains only one word (e.g. "great white shark" becomes "shark").

[3] Of course, we describe multimodal networks as *neither visual nor linguistic*, but this is to be understood in relative terms–they are *relatively* far from both visual models and linguistic models. In absolute terms, there is always a reasonable amount of similarity between multimodal networks and certain visual or linguistic models.

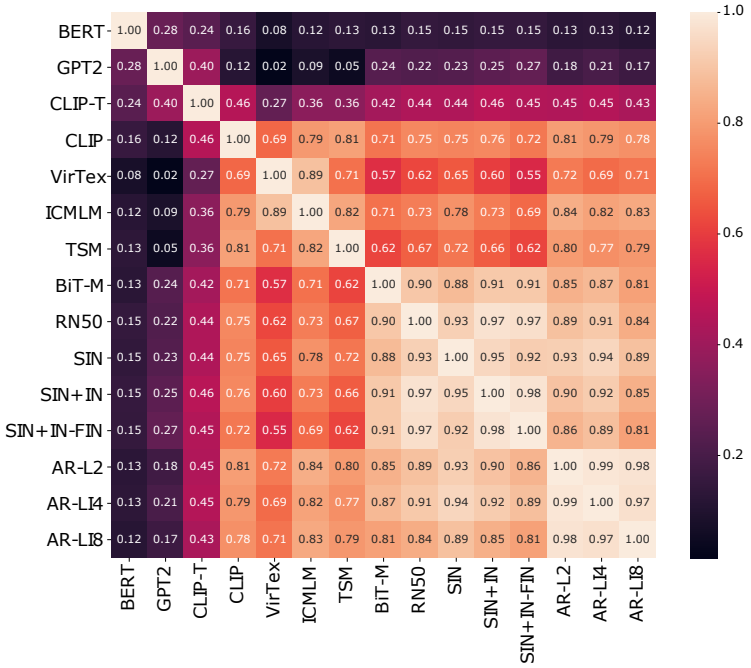

Figure 8: Correlations of the RDMs of our evaluation models. The RDMs are computed as explained in Fig 7 using the ImageNet dataset.

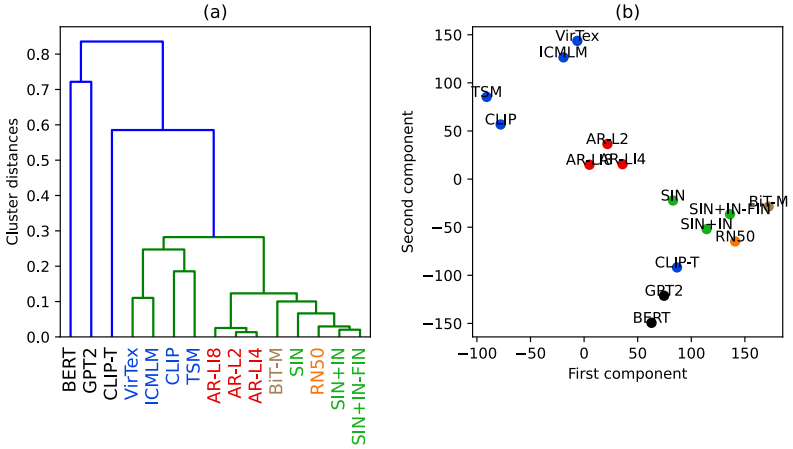

Figure 9: (a) Dendrogram of a hierarchical clustering of the RDMs. (b) t-SNE of the RDMs.

All of the ImageNet categories that have the same one-word hypernym are grouped together into one unique hyperclass. For instance, the "shark" hyperclass contains the classes "great white shark" and "tiger shark". Finally, to obtain the visual word embeddings, we average the visual representation of all the images of each hyperclass from the ImageNet validation set. This gives a total of 824 visual words.

Besides, we choose a vocabulary of 20,000 words (taken from the most frequent tokens in Wikipedia). Only 368 visual words are among the 20,000 most frequent words, so we extend our vocabulary to also contain the 456 other visual words, resulting in a total vocabulary of 20,456 words.

Embedding dimension Since the vision models do not all share the same feature dimensions, in order to compare all Skip-Gram models, we reduce the dimensionality of the feature spaces of all vision models to 300 dimensions using a PCA. The PCA is computed using the visual features of all images in the ImageNet validation set. Consequently, the Skip-Gram word embeddings are trained with 300 dimensions.

Training We train the Skip-Gram models for 5 epochs, using the standard negative sampling strategy. We use window sizes of 5 words and a learning

rate of 1e-3. We use the "`vectors_lockf`" feature of the Gensim library to freeze certain word embeddings during training.

For the dataset, we use a recent dump of Wikipedia and we split it into two sets containing 80% and 20% of the articles for the training and validation sets.

5.2 Evaluation

We evaluate our Skip-Gram embeddings on two tasks: word analogies and word pair similarities.

Word Analogy This standard task (Mikolov et al., 2013b) for evaluating the quality of word embeddings consists of quadruplets $\{A, B, C, D\}$ (e.g. "man", "king", "woman","queen") supporting the relation "A is to B as C is to D". The task consists in finding the 4th one given the first three, by solving the equation in the latent space: $D = B - A + C$. The more accurate the model, the better its representation. We evaluate the word embeddings on the full dataset provided by (Mikolov et al., 2013b) that we split in two different sets: *morphology analogies* (such as "write", "writes", "work", "works"), and *semantic analogies* (such as "son", "daughter", "boy", "girl"). If vision-language training helps "ground" the visually-derived word embeddings, we expect this grounding to be more helpful in the resolution of semantic, rather than morphology analogies.

Word Pair similarity Another task for evaluating the quality of word embeddings is to ask humans to rate the semantic similarity of pairs of words (e.g. on a scale of 0 to 10, how close is "queen" to "king"? How close is "queen" to "woman"? etc.) (Finkelstein et al., 2001) and then compute the same similarity evaluations in the latent spaces of the models. The higher the (Pearson) correlation between the pairwise similarities of a model and human pairwise similarity judgments, the better the representation of the model.

5.3 Results

The baseline Skip-Gram produces the best word embeddings overall (black bars in Fig 10). This is to be expected since the embeddings are learned freely, without any additional constraint during training. Interestingly, this baseline advantage is weakest in the case of the semantic analogy task (Fig 10, leftmost panel), where some of the vision and visio-linguistic models are on par with the baseline. This shows that the frozen vectors do not necessarily impede the performance when the analogies are defined semantically (and might thus be presumed to contain some visual component). However, even for these semantic analogies, vision or vision-language word embeddings never significantly surpassed the baseline performance.

In the word pair similarity task, networks show variable performance levels, but without a clear distinction between vision-only and vision-language models. Among the visio-linguistic networks, CLIP and TSM, which are trained contrastively on a large amount of data (see Figure 1) have embeddings that correlate well with human word similarity judgements. However, when compared with the vision-only models, we do not observe a clear-cut performance improvement. Indeed, the best vision-only model (BiT-M) is on par with CLIP and TSM. Interestingly, by comparing the results from the Fig 10 rightmost panel to the data plotted in Fig 1, we observe that among our twelve models, the top six for the word pair similarity task (TSM, CLIP, BiT, and the three AR models) correspond to those models that were trained on the largest datasets.

For the analogy tasks (semantic and morphology), there is no particular trend. However in both cases, the best performing model (excluding the baseline) is a visual one: SIN+IN in the semantic case, and AR-L2 in the morphology case.

In summary, we find that multimodal training of visual features does not improve their usefulness for language tasks either, and we suggest that the amount of training data may be a more important factor for generalization.

5.4 Legitimacy of the visual word embeddings

In the previous results, for training the visually-guided word embedding models, we averaged the visual feature vectors over many examples for each class. This averaging can potentially alter the quality of the embeddings, e.g. by discarding important information about the feature distributions. Thus, we check the validity of these averaged feature vectors[4], by verifying that they remain useful in a vision context. We use these visual feature vectors as class prototypes and evaluate the corresponding nearest-neighbor classification accuracy on the ImageNet validation set[5] with a method similar to

[4]We here test the 300d vectors after the PCA dimensionality reduction.

[5]With the images regrouped into our 824 classes.

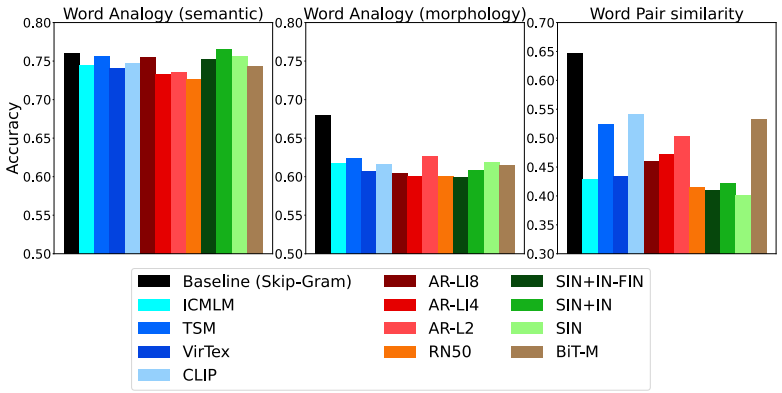

Figure 10: Semantic Word Analogy (such as "son", "daughter", "boy", "girl"), Morphology Word Analogy (such as "write", "writes", "work", "works") and word pair similarity results for the visually constrained Skip-Grams. The Baseline is a vanilla Skip-Gram model (300 dimensions) where all 20,456 word embeddings are free to be learned.

section 3.1. For all models considered, classification accuracy was well above chance ($p<0.01$): this means that the class-specific vectors indeed remain useful as visual representations of their category.

Furthermore, we computed the correlation between this visual classification accuracy of the word embedding, and the corresponding word analogy or word-pair similarity accuracy for each model. The resulting Pearson correlation coefficient was r=-0.0821 with the semantic Word Analogy performance, r=0.301 with the morphology Word Analogy performance, and r=0.797 with the Word Pair Similarity.

The significant high correlation of visual classification with word-pair similarity performance might be caused by the visual component of the word similarity judgments performed by human subjects. Indeed, many "similar words" also entail similar visual features (tiger, jaguar, cat, feline), and so the word-pair similarity task may not be a pure language task.

6 Discussion and Conclusion

It is a highly appealing notion that semantic grounding could improve vision models, by introducing meaningful linguistic structure into their latent space, and thereby increasing their robustness and generalization properties. Unfortunately, our experiments reveal that current vision-language training methods do not achieve this objective: the resulting multimodal networks are not better than vision-only models, neither for few-shot learning, transfer learning or unsupervised clustering, nor for adversarial robustness. In addition, compared to vision-only models, the multimodal networks' visual representations do not appear to provide additional semantic information that could serve as a useful constraint for a word-embedding linguistic space.

The present inability of linguistic grounding methods to deliver their full promise does not imply that this cannot happen in the future. In fact, we believe that exploring the current models' performance and representations, as we do here, can help us understand their limitations and adjust our methods accordingly. Specifically, we found that multimodal representations are neither visual nor linguistic, but are not really in-between either (Fig 9). In CLIP and TSM, for instance, the contrastive learning objective encourages the visual and language streams to agree on a joint embedding of images and corresponding captions. However, such agreement, by itself, does not constrain either latent space to remain faithful to its initial domain. As a result, CLIP's (and TSM's) visual representations may discard information that could prove critical for transfer-learning to other visual tasks. If this is true, we predict that adding domain-specific terms to the multimodal loss function (e.g. self-supervision) could be a way to improve visual generalization, while retaining the advantages of multimodal training—possibly including semantic grounding.

Acknowledgements

This research was supported by ANITI ANR grant ANR-19-PI3A-0004, AI-REPS ANR grant ANR-18-CE37-0007-01 and OSCI-DEEP ANR grant ANR-19-NEUC-0004.

References

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Karan Desai and Justin Johnson. 2021. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. 2019a. Robustness (python library).

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. 2019b. Adversarial robustness as a prior for learned representations.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: the concept revisited. In *Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001*, pages 406–414. ACM.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2121–2129.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Ronghang Hu and Amanpreet Singh. 2021. Transformer is all you need: Multimodal multitask learning with a unified transformer.

Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. 2016. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2019. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *Lecture Notes in Computer Science*, page 740–755.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6892–6899. AAAI Press.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2007. Learning visual representations using images with captions. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

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

Jonas Rauber, Wieland Brendel, and Matthias Bethge. 2017. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*.

Leila Reddy and Simon J Thorpe. 2014. Concept cells through associative learning of high-level representations. *Neuron*, 84(2):248–251.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. 2020. Do adversarially robust imagenet models transfer better? In *Advances in Neural Information Processing Systems*, volume 33, pages 3533–3545. Curran Associates, Inc.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning visual representations with caption annotations. In *European Conference on Computer Vision (ECCV)*.

Nitish Srivastava and Ruslan Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 2231–2239.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

A Datasets

We briefly describe all the datasets used in our experiments.

A.1 CIFAR10 and CIFAR 100

These datasets contain images of animals and objects comprising either 10 (CIFAR10) or 100 (CIFAR100) categories. All the 60,000 images – 50,000 train and 10,000 test– are of 32×32 resolution with RGB color channels.

A.2 CUB dataset

Caltech-UCSD Birds, or CUB, dataset consists of 6033 images of 200 species of birds. Apart from the species name, the dataset also provides bounding boxes, approximate bird segmentation and attribute labels for each image, allowing for a finer analysis at a feature level.

A.3 MNIST

Considered one of the simplest datasets, MNIST contains 28×28 black and white images of handwritten digits from 0 to 9. It comprises of 60,000 training images and 10,000 test images.

A.4 Fashion MNIST

Based on article images from Zalando, an e-commerce platform, Fashion MNIST contains 28×28 black and white images of 10 clothing categories. Designed with an aim to act as a "direct drop-in replacement for the original MNIST", it contains the same number of training and testing images as that of MNIST.

A.5 StreetView House Numbers

StreetView House Numbers, or SVHN dataset consists of images of digits from 0 to 9. Compared to MNIST, it is generally considered a more real-world dataset for optimizing neural networks since it contains images of digits in a more natural setting– 600,000 colored images of digits provided by Google Street View images.