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.

1 Introduction

Learning vision models using language supervision has gained popularity (Quattoni et al., 2007; Srivastava et al., 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.

2 Models

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

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 1M videos). 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., 2018), 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., 2011) and SVHN (Netzer et al., 2011). These datasets test generalization capabilities for natural images of various classes.

Few-shot learning As a first generalization experiment, we compare few-shot learning accuracy. For each class, we define a class prototype by averaging the feature representations of N prototypes (N-shot learning). We measure the performance of vision-only and text-vision models for $N =$ 1, 5 and 10, each time averaged over 10 trials with different class prototypes in each trial. Figure 1 shows the average performance of each model across datasets, in the leftmost 3 panels.

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. The clusters are computed only on the test-sets. To assign labels to each cluster, we match the most similar predicted and original clusters via a greedy algorithm. Figure 1 panel 4 (from left) shows the performance of the unsupervised clustering algorithm averaged over all datasets.

Transfer learning To further evaluate the models' generalization properties, we use a transfer learning setting as described in (Salman et al.,

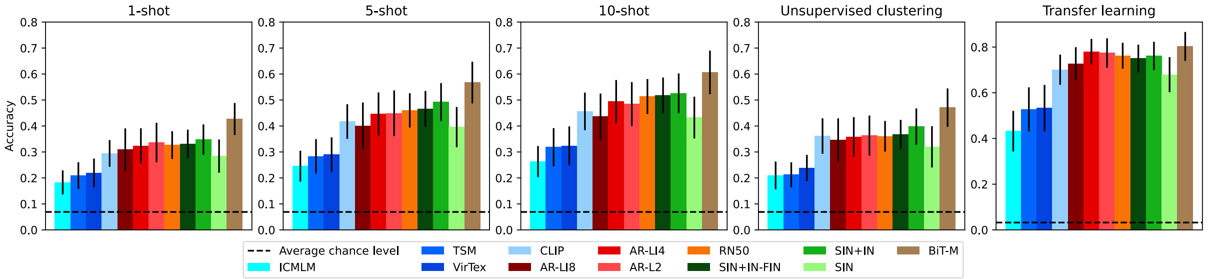

Figure 1: Average performance of the models across datasets, with standard error of the mean, for the various generalization tasks (few-shot learning, unsupervised clustering, transfer learning).

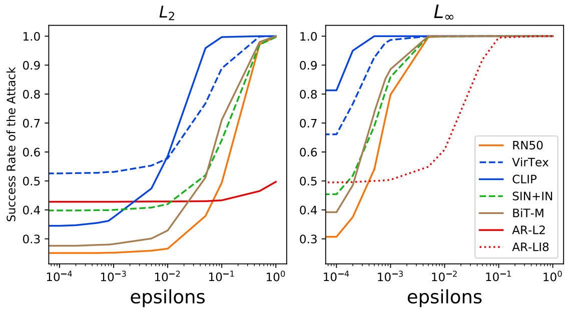

Figure 2: Robustness of some of the models to untargeted random projected gradient descent (RPGD) attacks for varying epsilons, with L_2 (left) or L_∞ norm (right). AR models are robust by design. Multimodal networks (CLIP, VirTex) are less robust than vision-only models (RN50, SIN+IN, BiT-M).

2020). We use the same datasets as in the other tasks, each time training a linear probe using the Adam optimizer.

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., 2017). As the results are qualitatively similar, we only report here the untargeted attacks. We use L_2 and L_∞ norms to distinguish any norm-specific effects. Figure 2 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).

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

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 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. We then compute the representational dissimilarity matrix $RDM = [d(\mathcal{F}_i, \mathcal{F}_j)]_{i,j}$ where

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

is the norm of the unequal variance t-test (Welch, 1947): this allows us to normalize the distances between class centroids with respect to their variances.

In the case of language models, we encode 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. As there is only one sentence per class, there is no variance associated with the feature vector of each class and 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 (Kriegeskorte et al., 2008).

Results Figure 3 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 2). 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.*

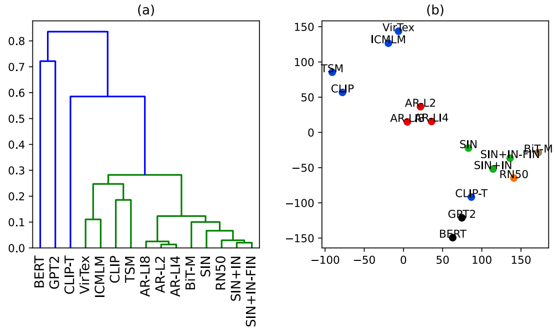

Figure 3: (a) Dendrogram of a hierarchical clustering of the RDMs. (b) t-SNE of the RDMs.

5 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.

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 3). 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 Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *arXiv preprint arXiv:2006.16228*.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model.

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 *International Conference on Learning Representations*.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2007. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

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.

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, Ruslan Salakhutdinov, et al. 2012. Multimodal learning with deep boltzmann machines. In *NIPS*, volume 1, page 2. Citeseer.

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.