Explaining Self-Supervised Image Representations with Visual Probing

Dominika Basaj[1,2*] , **Witold Oleszkiewicz**[1*] , **Igor Sieradzki**[3] , **Michał Górszczak**[3] ,
Barbara Rychalska[1,4] , **Tomasz Trzciński**[1,2,3] and **Bartosz Zieliński**[3,5]
[1]Warsaw University of Technology
[2]Tooploox
[3]Faculty of Mathematics and Computer Science, Jagiellonian University
[4]Synerise
[5]Ardigen
dominika.basaj@tooploox.com, witold.oleszkiewicz@pw.edu.pl, igor.sieradzki@uj.edu.pl,
michal.gorszczak@student.uj.edu.pl, b.rychalska@mini.pw.edu.pl, tomasz.trzcinski@pw.edu.pl,
bartosz.zielinski@uj.edu.pl

Abstract

Recently introduced self-supervised methods for image representation learning provide on par or superior results to their fully supervised competitors, yet the corresponding efforts to explain the self-supervised approaches lag behind. Motivated by this observation, we introduce a novel visual probing framework for explaining the self-supervised models by leveraging probing tasks employed previously in natural language processing. The probing tasks require knowledge about semantic relationships between image parts. Hence, we propose a systematic approach to obtain analogs of natural language in vision, such as visual words, context, and taxonomy. We show the effectiveness and applicability of those analogs in the context of explaining self-supervised representations. Our key findings emphasize that relations between language and vision can serve as an effective yet intuitive tool for discovering how machine learning models work, independently of data modality. Our work opens a plethora of research pathways towards more explainable and transparent AI.

1 Introduction

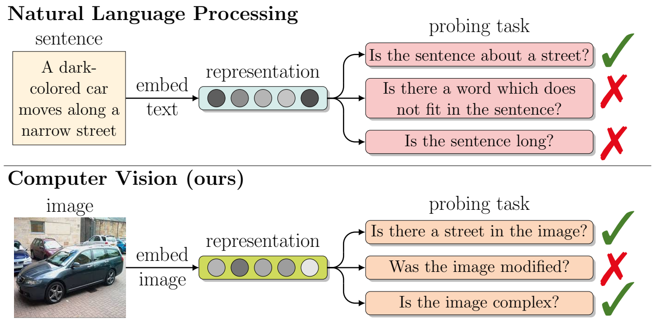

Natural Language Processing
sentence
A dark-colored car moves along a narrow street
embed text
representation
probing task
Is the sentence about a street? ✓
Is there a word which does not fit in the sentence? ✗
Is the sentence long? ✗
Computer Vision (ours)
image
embed image
representation
probing task
Is there a street in the image? ✓
Was the image modified? ✗
Is the image complex? ✓

Figure 1: *Probing tasks*, widely used in natural language processing, validate if a *representation* implicitly encodes a given property, *e.g.*, sentence topic or length. We introduce a visual taxonomy along with the corresponding probing framework that allow to build analogous *visual probing tasks* and explain the self-supervised image representations. As a result, we *e.g.* discover that even though all analysed models build similar semantic knowledge, some of them focus more on texture and therefore achieve better accuracy on target tasks.

Visual representations are cornerstones of a multitude of contemporary computer vision and machine learning applications, ranging from visual search [Sivic and Zisserman, 2006] to image classification [Krizhevsky *et al.*, 2012] and visual question answering (VQA) [Antol *et al.*, 2015]. However, learning representations from data typically requires tedious annotation. Therefore, recently introduced self-supervised representation learning methods concentrate on decreasing the need for data labeling without reducing their performance [Chen *et al.*, 2020b; Grill *et al.*, 2020; Caron *et al.*, 2020]. Because of the fundamental role representations play in real-life applications, a lot of research focuses on explaining these embeddings [Vulić *et al.*, 2020; Eichler *et al.*, 2019; Huang and Li, 2020]. Nevertheless, most of them concentrate on fully supervised embeddings [Zhang and Zhu, 2018] and not on their self-supervised counterparts. Moreover, the majority of the proposed approaches rely on pixel-wise image analysis [Simonyan *et al.*, 2014; Adebayo *et al.*, 2018], while general semantic concepts present in the images are often ignored.

Here, we attempt to overcome these shortcomings and draw inspiration from a simple yet often overlooked observation that humans use language as a natural tool to explain what they learn about the world through their eyes [Kumar and Talukdar, 2020]. Therefore, considering that the very same machine learning algorithms can be successfully applied to solve both vision and natural language processing (NLP) tasks [Dosovitskiy *et al.*, 2020; Carion *et al.*, 2020], we postulate that the methods used to analyze text representation can also be employed to investigate visual inputs.

Very popular tools for explaining textual embeddings are *probing tasks* [Conneau *et al.*, 2018]. As shown in the up-

[*] Equal contribution
The code is at: github.com/BioNN-InfoTech/visual-probes

per part of Fig. 1, a probing task in NLP is a simple classifier that asks if a given textual representation encodes a particular property, such as a sentence length or its semantic consistency, even though this property was not a direct training objective. For instance, we can create a textual probing task by substituting a word in a sentence and checking if a simple classifier that takes the representation of the original and altered sentence can detect this change. By analyzing the accuracy of a probing task, one can verify if the investigated representation contains certain information and understand the rationale behind embedding creation. However, while probing tasks are straightforward, intuitive, and widely used tools in NLP, their computer vision application is limited [Alain and Bengio, 2017], mainly due to the lack of appropriate analogs between textual and visual modalities.

In this paper, we address this limitation by introducing a mapping between vision and language that enables applying the NLP probing tools in the computer vision (CV) domain. For this purpose, in Sec. 3, we propose a taxonomy of visual units that includes *visual sentences*, *words*, and *characters*. We then employ these units as building blocks for a more general visual probing framework that contains a variety of NLP-inspired probing tasks, such as *word content*, *sentence length*, *character bin*, and *Semantic Odd Man Out* [Conneau *et al.*, 2018; Eichler *et al.*, 2019]. The results we obtain provide us with unprecedented insights into semantic knowledge, complexity, and consistency of self-supervised image representations, *e.g.* we discover that semantics of the image only partially contribute to target task accuracy. Our framework also allows us to compare the existing self-supervised representations from a novel perspective, as we show in Sec. 5.

Our contributions can be therefore summarized as follows:

- We propose a mapping between visual and textual modalities that constructs a visual taxonomy.
- We introduce novel visual probing tasks for comparing self-supervised image representations inspired by similar methods used in NLP.
- We show that leveraging the relationship between language and vision serves as an effective yet intuitive tool for discovering how self-supervised models work.

2 Related Works

The visual probing framework aims to explain image representations obtained from self-supervised methods. It is inspired by probing tasks used in NLP. Therefore, we consider related works from three research areas: self-supervised computer vision models, probing tasks in natural language processing, and explainability methods in computer vision.

Self-supervised computer vision models. Recently published self-supervised methods provide state-of-the-art results across computer vision tasks. They usually base on contrastive loss [Hadsell *et al.*, 2006] that measures the similarities of patches in representation space and aims to discriminate between positive and negative pairs. The positive pair contains modified versions of the same image, while the negative pairs correspond to two images in the same dataset. One of the methods, MoCo v1 [He *et al.*, 2019] trains a slowly progressing visual representation encoder, driven by a momentum update. This encoder plays a role of a memory bank of past representations and delivers negative examples. SimCLR v2 [Chen *et al.*, 2020b], unlike MoCo v1, proposes a different way of generating negative pairs. Instead of a memory bank, they propose to use a large batch size of up to 4096 examples. Other improvements proposed by SimCLR v2 are a projection head and carefully tuned data augmentation. The projection head maps representations to space where contrastive loss is applied, which is important due to the loss of information. BYOL [Grill *et al.*, 2020] also uses the projection head, but unlike MoCo v1 and SimCLR v2, it achieves a state-of-the-art performance without the explicitly defined contrastive loss function, so it does not need negative examples. On the other hand, SwAV [Caron *et al.*, 2020] takes advantage of contrastive methods without pairwise comparisons. Instead, it learns the representations by clustering them and predicting the labels of their clusters. Our paper provides a framework to analyze the representation generated by those methods in terms of the semantic knowledge they encode.

Probing tasks in NLP. One of the classic examples of the NLP probing task aims to probe sentence embeddings for interesting linguistic features such as the depth of the parse tree or whether the sentence contains a specific word [Conneau *et al.*, 2018]. Others propose to focus on lexical knowledge concerning the qualities of individual words more than the whole sentences [Vulić *et al.*, 2020; Eichler *et al.*, 2019]. We consider both these objectives in our approach, i.e. we study probing tasks on specific concepts and their compositions. Moreover, while most works on probing tasks focus on one selected language, the others [Eichler *et al.*, 2019] are designed with multilingual settings in mind. This paper reflects the latter because it can be applied to various image domains.

Explainability methods in CV representation learning. The existing methods for explaining image representations either verify the relevance of hidden layers of supervised classification networks [Alain and Bengio, 2017] or highlight individual pixels that are essential for the model [Simonyan *et al.*, 2014; Adebayo *et al.*, 2018]. Moreover, they usually generate the important regions as pixel clouds, which are not understood as concrete semantic concepts. In contrast, approaches such as [Huang and Li, 2020; Ghorbani *et al.*, 2019] aim to detect important image segments but are often difficult to understand in practice, even though they are crucial for the model objective. In this work, we extend the existing methods by analyzing the semantic information stored in the self-supervised representation.

3 Visual Probing

This section introduces a novel visual probing framework that analyzes the information stored in self-supervised image representations. For this purpose, in Sec. 3.1, we propose a mapping between visual and textual modalities that constructs a visual taxonomy. As a result, the image becomes a "visual sentence" and can be analyzed with visual probing tasks inspired by similar methods used in NLP (see Sec. 3.2).

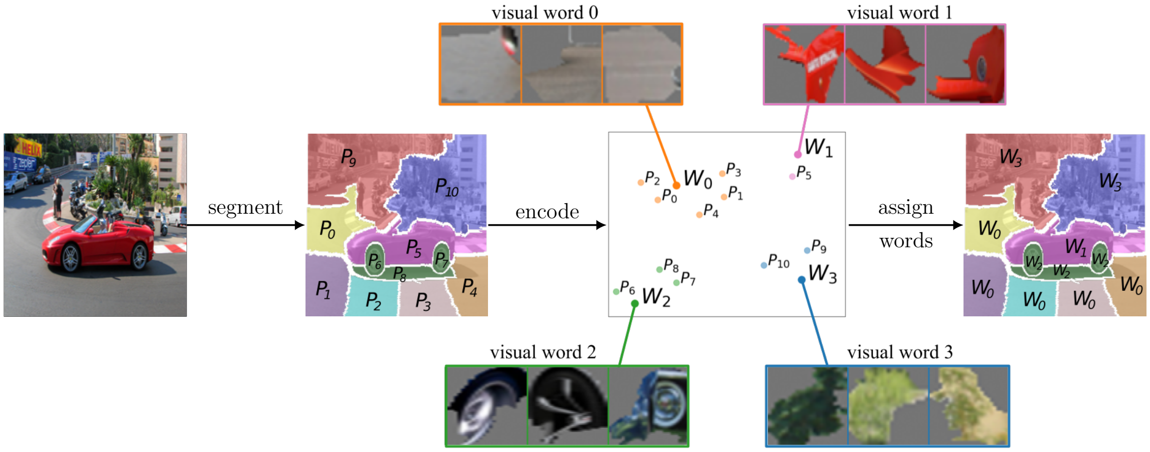


Figure 2: The process of dividing an image into visual words. First, an image is segmented into multiple superpixels: $P_0, P_1, \ldots, P_{10}$. Then, each superpixel is embedded in the latent space previously used to generate the dictionary of visual words: W_0, W_1, W_2, W_3. Finally, each superpixel is assigned to the closest word in the visual word dictionary. This results in mapping between vision and language and enables using the visual probing framework that includes a variety of NLP-inspired probing tasks.

3.1 Mapping Between Vision and NLP

While an image can be considered a sentence equivalent in a probing task, the question remains, what is the equivalent to words and characters? There are multiple possible answers to this question. One of the intuitive ones is to divide an image into non-overlapping superpixels that group pixels into perceptually meaningful atomic regions [Achanta *et al.*, 2012]. As a result, we obtain an image built from superpixels as an analogy of a sentence built from the words. The superpixels, similarly like words, have order and meaning. Moreover, each superpixel contains a specific number of pixels, like the number of word's characters. As a consequence, we obtain an intuitive mapping between visual and textual domains.

However, superpixels treated as visual words would significantly differ from their linguistic counterparts because they do not repeat between images, while in text, the words often repeat between sentences. Therefore, we propose to define visual words as the clusters of superpixels in representation space and assign each superpixel to the closest centroid. For this purpose, we could use the original definition of visual words from [Leung and Malik, 2001]. However, it does not take into account the importance of those words for a model's prediction. Therefore, instead of that, we use TCAV methodology [Kim *et al.*, 2018; Ghorbani *et al.*, 2019] that generates high-level concepts, which are important for prediction and easily understandable by humans. Such an approach requires a supervisory training network but generates visual words independent of any compared self-supervised techniques, which is crucial for a fair comparison. Therefore, the process of dividing an image into visual words consists of three steps: segmentation into superpixels, their encoding, and assignment to visual words (see Fig. 2).

3.2 Visual Probing Tasks

After dividing an image into visual words, it can be analyzed by the visual probing framework, which can adapt almost any NLP probing task. Here, we describe the four that are well known by the NLP community [Conneau *et al.*, 2018; Eichler *et al.*, 2019]. Moreover, except for defining visual probing tasks, we provide their original NLP definitions to make the paper self-contained.

Word Content (WC). The word content probing task aims to identify which visual words are present in the image. The *input* of this probing task is a self-supervised representation of the image. The *target labels* represent the presence of a particular visual word. As we describe in Sec. 4, we select 100 representative visual words. Hence, there are 100 binary *target labels*. Fig. 2 illustrates the process of determining which visual words are present in the image. The NLP inspiration of the task probes for surface information, the type of information that does not require any linguistic knowledge [Conneau *et al.*, 2018]. In contrast, its adaptation requires ***semantic knowledge*** to understand which concept is represented by a superpixel.

Sentence Length (SL). The aim of the sentence length probing task is to distinguish between simple and complex images, as presented in Fig. 3. The *input* of this probing task is a self-supervised representation of the image. The *target label* is the number of unique visual words in the image, which can be determined based on the WC labels. The original NLP probing task predicts the number of words (or tokens) and retains only surface information [Conneau *et al.*, 2018]. At CV, it serves as a proxy for ***semantic complexity***, requiring the semantic understanding of the image.

Character Bin (CB). The aim of the character bin probing task is to check whether the representation stores information about the complexity of the image. The *input* of this probing task is a self-supervised representation of the image's superpixel. The *target label* is the size of the superpixel defined as the number of non-grey pixels, as presented in Fig. 4. The original NLP probing task is defined as a classifier of the number of characters in a single word [Eichler *et al.*, 2019]. From this perspective, the character bin retains only surface

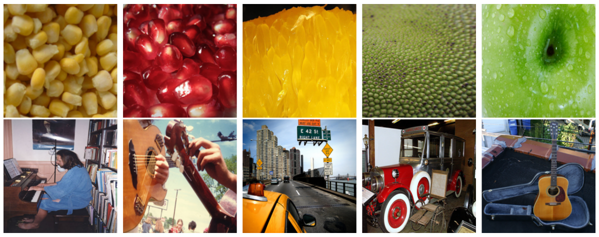


Figure 3: The SL probing task measures how well the representation encodes the information about the number of unique visual words in the image. *Top row:* a low number of unique visual words (<13). *Bottom row*: a high number of unique visual words (>42).

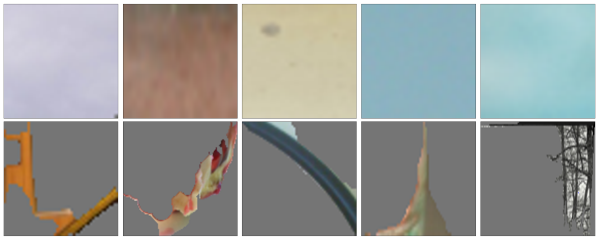


Figure 4: The CB probing task measures if the representation retains information about the superpixel's size. *Top row:* examples of large superpixels. *Bottom row:* example of small superpixels.

information in both domains.

Semantic Odd Man Out (SOMO). The objective of the SOMO probing task is to predict whether the image was modified by replacing a random superpixel in the image with a similarly shaped superpixel from another image, and corresponding to a different visual word, as presented in Fig. 5. The *input* of this probing task is a self-supervised representation of the image. The *target label* is binary, i.e. the image was modified or not. The original NLP task predicts if the sentence was altered by replacing a random noun or verb [Conneau *et al.*, 2018]. In both domains, it requires the ability to detect alterations in ***semantic consistency***.

4 Experimental Setup

In this section, we describe the procedure of generating visual words and training probing tasks.

Generating visual words. We use the original settings of the ACE algorithm described in [Ghorbani *et al.*, 2019] that first divides images into superpixels using SLIC algorithm [Achanta *et al.*, 2012] with three resolutions of 15, 50, and 80 segments for each image. It then computes representations of these superpixels as an output of a *mixed4c* layer of GoogLeNet [Szegedy *et al.*, 2015] trained on the ImageNet dataset. Finally, representations are clustered using the k-means algorithm, resulting in clusters that correspond to the visual words (see Fig. 6). As there are over a dozen visual words generated for each of the classes, the dictionary's size grows significantly with the size of the analyzed dataset. Therefore, in this paper, we decided to analyze its subset containing 55 classes grouped into 5 categories: animals, vehicles, musical instruments, buildings, fruits. Moreover, to further limit the dictionary size, we only keep 100 of the most relevant visual words (according to TCAV score [Kim *et al.*, 2018]), while ensuring that each class is represented by at least one of them. These 100 visual words form our visual language.

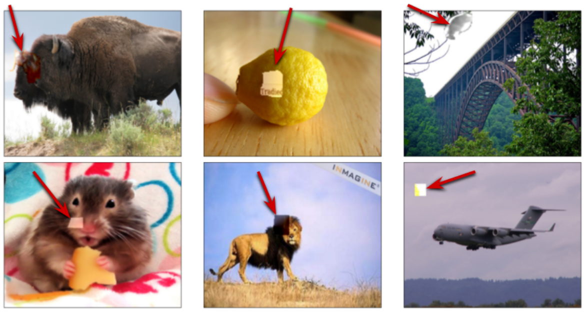


Figure 5: The SOMO probing task predicts if the image was altered by replacing a random superpixel. *Top row:* examples of images for which SimCLR v2 correctly recognizes the modification, while SwAV fails. *Bottom row:* examples of images where both SimCLR v2 and SwAV do not recognize superpixel modification.

Generating a self-supervised representation. We examine four self-supervised methods: MoCo v1 [He *et al.*, 2019], SimCLR v2 [Chen *et al.*, 2020b], BYOL [Grill *et al.*, 2020], and SwAV [Caron *et al.*, 2020]. For all of them, we use publicly available models trained with ImageNet. Although they all use the penultimate layer of ResNet-50 to generate representations, their training hyperparameters differ, which is described in Supplementary Materials (SM in the following)[1].

Assigning visual words. To assign a superpixel to a visual word, we pass it through the GoogLeNet to generate a representation from the *mixed4c* layer (similarly to generating visual words). We can then determine the visual word closest to a superpixel, as both are embedded in the same space.

Training probing tasks. We use a logistic regression classifier with a maximum of 1000 iterations and the LBFGS solver to train all diagnostic classifiers. As an input, we use representations generated by the self-supervised methods. The output depends on the probing task. In the case of the WC, we train 100 classifiers corresponding to 100 visual words. We expect an image to be assigned to a particular visual word if at least one of its superpixels is assigned to it. Finally, we report the average AUC scores over 100 classifiers (see Tab. 1). To obtain classification setup in the sentence length probing task, we group the possible output into 5 equally-wide bins, resulting in one-vs-rest OVR AUC, which is resistant to class imbalance. A similar procedure is applied to the character bin probing task, except that we use 6 bins in this case. SOMO is formulated as a binary classification task, in which we predict whether the image was modified or not. The training and validation datasets are balanced. We conduct all of our experiments on the ImageNet dataset [Deng *et al.*, 2009], keeping its standard train/validation split. More-

[1] Supplementary materials: http://www.ii.uj.edu.pl/~zielinsb/papers/visual_probing_ijcai_supplement.pdf

		Probing tasks (ours)			
	Target	**WC**	**SL**	**CB**	**SOMO**
MoCo v1	0.606	0.790	0.868	0.937	0.559
SimCLR v2	0.717	**0.800**	**0.877**	**0.964**	**0.625**
BYOL	0.723	0.795	0.876	0.961	0.615
SwAV	**0.753**	0.761	0.838	0.956	0.530

Table 1: AUC score for our probing tasks and accuracy on the linear evaluation (Target). Like the linear evaluation, our probing tasks are also trained on top of the frozen base network, and test accuracy is used as a proxy for representation quality. Hence, they provide complementary knowledge about the representation.

over, we apply the random over-sampling if needed to deal with the imbalanced classes. The details on the experimental setup are presented in SM.

5 Results and Discussion

Tab. 1 summarizes the results obtained in our experiments. It presents the performance of our probing tasks and the target task accuracy for reference. The reported target task performance is the classification accuracy calculated for the whole ImageNet validation set. The first conclusion is that self-supervised representations retain information about semantic knowledge and semantic complexity, but they do not code much information about image consistency. Secondly, the performance on probing tasks do not correlate with accuracy on the target task. Finally, SimCLR v2 overpasses other methods in all probing tasks. In the following, we analyze those aspects in greater detail.

Self-supervised representations contain strong semantic knowledge. As outlined in 3.2, we treat the results of the word content probing as an approximation of semantic knowledge present in a representation. The AUC scores for this probing task reported in Tab. 1 vary from 0.76 for SwAV to 0.8 for SimCLR v2. This shows ability to predict which visual words are present in the image. Based on this we can say that semantic knowledge is encoded in the examined self-supervised representations.

The level of semantic knowledge does not correlate with target task accuracy. It is surprising that although examined self-supervised methods have diverse target task accuracy, they all have a similar level of semantic knowledge. E.g. MoCo v1 obtains the worst target task accuracy (61%), but the results of the WC probing task is on par with stronger self-supervised methods. Even more surprising is that SwAV, despite its highest accuracy on the target task, is below the scores of other tested methods in terms of semantic knowledge measured by the WC probing task. This finding supports the view that semantic knowledge only partially contributes to the target task accuracy [Geirhos *et al.*, 2020].

Certain types of semantic knowledge are represented better than others. The probing task's ability to predict which visual words are present in the image varies, as some words are better predicted than others. We conducted a user study to understand the difference between best and worst predicted visual words presented in Fig. 6. According to the results, the five best recognizable visual words are perceived to have distinct, non-uniform textures. In contrast, the five worst recognizable visual words have more uniform textures. This may indicate that self-supervised representations are pattern-biased. This sheds new light on this problem, as previous results [Geirhos *et al.*, 2020] suggest the opposite. See SM for details on the user study.

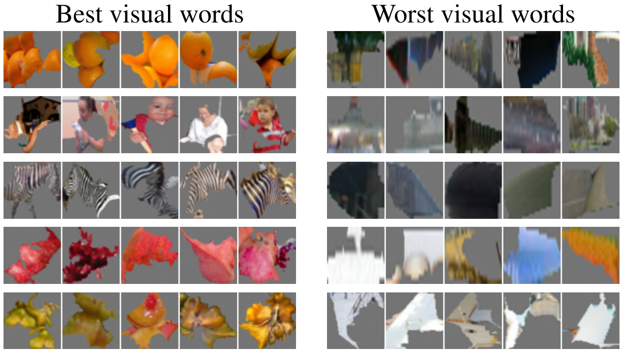
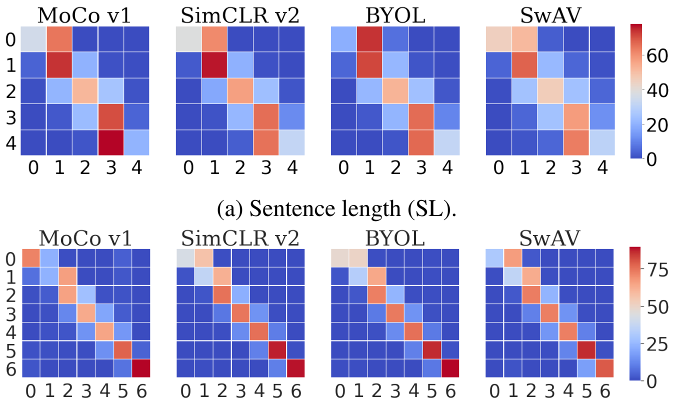


Figure 6: Visualization of the best and the worst predicted visual words, according to the results of the WC probing task (on average by all self-supervised methods). Our user study shows that the best recognizable visual words are perceived to have distinct non-uniform textures contrary to the worst recognizable ones which have more uniform textures. This may indicate that self-supervised representations better encode information about patterns.

There are visible differences in semantic knowledge retained by different self-supervised methods. Our user study shows variability by comparing the semantic content of representations on individual visual words. We take a closer look at the visual words that some self-supervised methods encode better or worse than others. The examples of these visual words are in SM. Looking at the top five visual words that MoCo v1 encodes better than the other representations, we can see that these words have distinct patterns. Moreover, the user study shows that MoCo v1 is better than the others at recognizing non-uniform textures. On the other hand, SimCLR v2, BYOL, and SwAV are above average in recognizing uniform textures.

Self-supervised representations contain information about semantic complexity. We design two probing tasks - sentence length and character bin - which validate the complexity of an image. Based on the results in Tab. 1, we observe that representations reflect the level of semantic complexity to a high degree. Information about the number of unique visual words (SL) is equally well predicted by probing classifiers for all self-supervised representations. These results are consistent with the results for semantic knowledge. For both probing tasks, SwAV's performance is slightly below the scores of other tested methods. This demonstrates the link between semantic complexity and semantic knowledge. AUCs are even higher for predicting the size of a visual word, which indicates that representation encodes the approximation of its shape (although technically, we predict the number of pixels).



(a) Sentence length (SL).

(b) Character bin (CB).

Figure 7: Confusion matrices for SL and CB (results in %). The results indicate that the ability of self-supervised representations to retain information about complexity differs depending on the level of image complexity. Moreover, even though the final AUCs are similar, their confusion matrices vary.

The ability of self-supervised representations to retain information about semantic complexity differs. There are no substantial differences in the ability to encode the complexity of the images between self-supervised methods. However, some preferences can be observed once we do not aggregate predictions into one AUC number. A closer look at the confusion matrices for the SL probing task in Fig. 7 shows that BYOL does worse when it comes to recognizing less complex images, but it performs well in comparison to other self-supervised methods. That is in contrast to SwAV, which overall has the lowest AUC metric, but it stands out when it comes to passing on information about simple images.

Self-supervised representations struggle to retain information about semantic consistency. Contrary to what we observe for semantic knowledge and complexity, self-supervised representations do not encode well the information about semantic consistency. The ability to distinguish altered images differs between methods, with the smallest AUC metric (53%) for representations extracted by SwAV and the highest for SimCLR v2 (63%). Manual inspection of examples from the top and bottom performers classified as true positive and false negative with high (>80%) certainty indicates differences in decision making of probing classifiers. Firstly, we observe that SimCLR v2 does relatively well with examples that people easily recognize as modified. However, it performs worse on more blended alterations (Fig. 5), which do not disturb the huge chunks of textures or colors. At the same time, in most cases, information encoded in SwAV's representation does not reflect well enough even such visible alterations. Fig. 5 shows correct predictions for SimCLR v2 which SwAV predicted as not changed. Analysis of visual words for which we replaced the original ones across true positive and false negative for both SimCLR v2 and SwAV does not indicate any substantial differences between them. Hence, we conclude that the performance of the SOMO does not depend on the visual word we use as a replacement, but rather to what extent the semantic sentence is altered.

Self-supervised representations are resistant to modifications. Even though the replacements of visual words do not disturb the substantial part of the image, this lack of ability to distinguish alterations is interesting in the light of [Hendrycks *et al.*, 2019], which claims that self-supervised methods improve out-of-distribution detection. We do not contradict this conclusion, but our results show that in particular setups, self-supervised representations do not exhibit enough ability to distinguish between corrupted and not corrupted images. Considering that various, even minor and not visible, alterations might lead to a change in the outcome of the prediction, we postulate that the tendency of self-supervised representations not to retain information about consistency might pose a risk. When it comes to the differences in the AUC for the examined representations, they might be partially explained by differences in the architecture. E.g. SimCLR v2 and BYOL are trained with projection head, whereas SwAV and MoCo v1 are not. The projection allows retaining information about the transformation of the input image [Chen *et al.*, 2020a]. Therefore, we hypothesize that this information may cause differences in the AUC score.

6 Conclusions

In this work, we introduce a novel visual probing framework that analyzes the information stored in self-supervised image representations. It is inspired by probing tasks employed in NLP and requires similar taxonomy. Hence, we propose a set of mappings between visual and textual modalities to construct visual sentences, words, and characters. The results of the experiments confirm the effectiveness and applicability of this framework in understanding self-supervised representations. We verify that the representations contain information about semantic knowledge and complexity of the images, although they struggle to retain information about image consistency. Moreover, a detailed analysis of each probing task reveals differences in the representations encoded by various methods. This provides knowledge about representation complementary to the accuracy of linear evaluation.

Finally, we show that the relations between language and vision can serve as an effective yet intuitive tool for explainable AI. Hence, we believe that our work will open new research directions in this domain.

Acknowledgments

This research was supported by: grant no POIR.04.04.00-00-14DE/18-00 carried out within the Team-Net program of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund; grant no 2018/31/N/ST6/02273 funded by National Science Centre, Poland; Priority Research Area Digiworld under the program Excellence Initiative – Research University at the Jagiellonian University in Kraków; and POB Research Centre for Artificial Intelligence and Robotics of Warsaw University of Technology within the Excellence Initiative Program - Research University (ID-UB).

References

[Achanta *et al.*, 2012] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11), 2012.

[Adebayo *et al.*, 2018] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018.

[Alain and Bengio, 2017] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017.

[Antol *et al.*, 2015] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[Carion *et al.*, 2020] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[Caron *et al.*, 2020] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint 2006.09882*, 2020.

[Chen *et al.*, 2020a] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[Chen *et al.*, 2020b] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020.

[Conneau *et al.*, 2018] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*, 2018.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[Dosovitskiy *et al.*, 2020] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint 2010.11929*, 2020.

[Eichler *et al.*, 2019] M. Eichler, G. G. Şahin, and Ir Gurevych. LINSPECTOR WEB: A multilingual probing suite for word representations. In *EMNLP-IJCNLP: System Demonstrations*, 2019.

[Geirhos *et al.*, 2020] R. Geirhos, K. Narayanappa, B. Mitzkus, M. Bethge, F. A. Wichmann, and W. Brendel. On the surprising similarities between supervised and self-supervised models. In *NeurIPS Workshop*, 2020.

[Ghorbani *et al.*, 2019] A. Ghorbani, James Wexler, J. Zou, and Been Kim. Towards automatic concept-based explanations. In *NeurIPS*, 2019.

[Grill *et al.*, 2020] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint 2006.07733*, 2020.

[Hadsell *et al.*, 2006] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[He *et al.*, 2019] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint 1911.05722*, 2019.

[Hendrycks *et al.*, 2019] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019.

[Huang and Li, 2020] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *CVPR*, 2020.

[Kim *et al.*, 2018] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[Kumar and Talukdar, 2020] Sawan Kumar and Partha Talukdar. NILE : Natural language inference with faithful natural language explanations. In *ACL*, 2020.

[Leung and Malik, 2001] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. In *International journal of computer vision*. Springer, 2001.

[Simonyan *et al.*, 2014] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[Sivic and Zisserman, 2006] Josef Sivic and Andrew Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*. Springer, 2006.

[Szegedy *et al.*, 2015] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[Vulić *et al.*, 2020] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen. Probing pretrained language models for lexical semantics. *arXiv preprint arXiv:2010.05731*, 2020.

[Zhang and Zhu, 2018] Quan-shi Zhang and Song-chun Zhu. Visual interpretability for deep learning: A survey. In *Frontiers of Information Technology & Electronic Engineering*, 2018.