Which Visual Features Impact the Performance of Target Task in Self-supervised Learning?

Witold Oleszkiewicz[1][0000−0002−7234−393X], Dominika Basaj[2][0000−0002−9377−3517], Tomasz Trzciński[1,2,3][0000−0002−1486−8906], and Bartosz Zieliński[3,4][0000−0002−3063−3621]

[1] Warsaw University of Technology, plac Politechniki 1, Warszawa, Poland
witold.oleszkiewicz@pw.edu.pl
[2] Tooploox, Teczowa 7, Wrocław, Poland
[3] Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, Kraków, Poland
[4] Ardigen, Podole 76, Kraków, Poland

Abstract. Self-supervised methods gain popularity by achieving results on par with supervised methods using fewer labels. However, their explaining techniques ignore the general semantic concepts present in the picture, limiting to local features at a pixel level. An exception is the visual probing framework that analyzes the vision concepts of an image using probing tasks. However, it does not explain if analyzed concepts are critical for target task performance. This work fills this gap by introducing amnesic visual probing that removes information about particular visual concepts from image representations and measures how it affects the target task accuracy. Moreover, it applies Marr's computational theory of vision to examine the biases in visual representations. As a result of experiments and user studies conducted for multiple self-supervised methods, we conclude, among others, that removing information about 3D forms from the representation decrease classification accuracy much more significantly than removing textures.

Keywords: Explainability · Self-supervision · Probing tasks

1 Introduction

Visual representations are critical in many computer vision and machine learning applications. The spectrum of these applications is broad, starting with visual search [21] to image classification [16] and visual question answering [3]. However, supervised representation learning requires a large amount of labeled data, usually time-consuming and expensive. Hence, self-supervised methods gain popularity, achieving results on par with supervised methods using fewer labels [6, 8, 13].

Along with the increasing proliferation of self-supervised methods for representation learning, there is a growing interest in developing methods that allow the interpretation of the resulting representation space and draw conclusions regarding the information it conveys. However, most of them focus on supervised

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

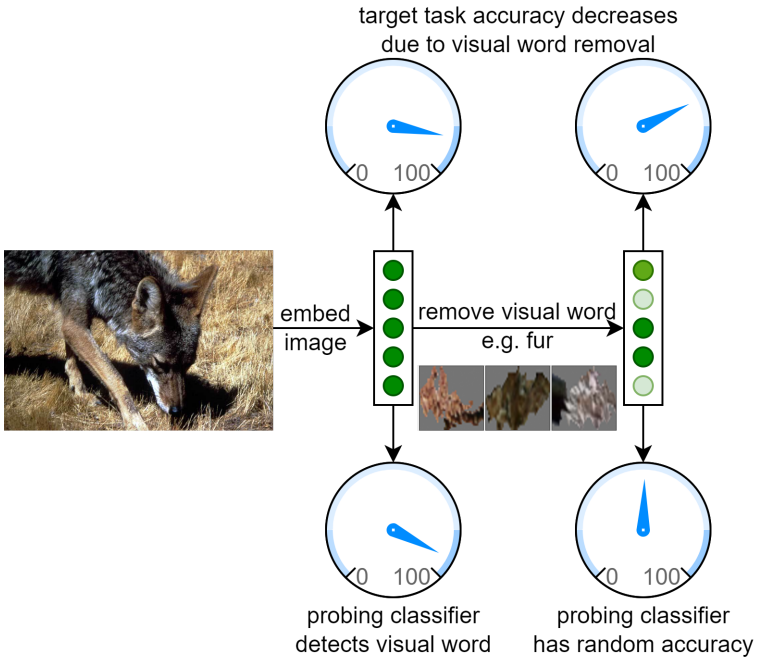


Fig. 1: Amnesic visual probing removes a specific visual concept (here corresponding to fur) from the self-supervised representation of an image (here corresponding to a wolf). As a result, the probing classifier cannot detect the presence of fur in the representation, and the target task accuracy decreases. The level of decrease represents the importance of the considered concept.

approaches and study local features at a pixel level [2, 20]. At the same time, the general semantic concepts present in the image are often overlooked, and their influence on model decisions is unknown. From this perspective, an exception is visual probing [4] that analyzes the vision concepts of an image using probing tasks. The probing tasks provide information about the presence of visual concepts in the representations but do not explain if they are critical for target task performance.

In this work, we overcome this limitation, providing a method that investigates the importance of visual features in the context of target task performance, referring to the amnesic probing [10] used in natural language processing (NLP). We remove information about particular visual concepts from image representations using the Iterative Nullspace Projection [19] and measure how it affects the target task accuracy. In addition, we conduct user studies to describe the visual concepts using Marr's computational theory of vision [17]. As a consequence, we can examine the biases in image representations.

Our contributions can be summarized as follows:

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

- We propose amnesic visual probing, a method for analyzing which visual features impact the performance of a target task.
- We apply Marr's computational theory of vision to examine the biases in visual representations.
- We conduct a complete user study and assign automatically generated visual concepts to one of six visual features from Marr's computational theory of vision.

2 Related Works

Our work corresponds to two research areas: self-supervised learning and probing tasks. We briefly cover the latest achievements in these two topics in the following paragraphs.

Self-supervised image representations Image representations obtained in a self-supervised manner are increasingly popular due to the competitive performance compared to supervised approaches. It is because they leverage the power of datasets without label annotations. One of the methods, called MoCo v1 [14], is based on a dictionary treated as a queue of data samples. It contains two encoders for query and keys, which are matched by contrastive loss. This queue enables to use of a large dictionary of examples previously limited to the batch size. SimCLR v2 [8] is another powerful method, which builds upon its predecessor, SimCLR [7] that maximizes the agreement between two views of the same sample by contrastive loss. In [8], the authors use a deeper and thinner backbone (ResNet-152 3x), deepen the projection head, which is not removed after contrastive training, and adapt memory mechanism from MoCo to increase the pool of negative examples. SwAV [6] takes advantage of contrastive methods. However, it compares clusters of data instead of single examples. The consistency between clusters, which can be seen as views of the same data sample, is enforced by learning to predict one view from another. In contrast to the above methods, BYOL [13] does not use the explicitly defined contrastive loss function, so it does not need negative samples. Instead, it uses two neural networks, referred to as online and target networks, that interact and learn the representation of the same image from each other.

Probing tasks The probing tasks originally come from Natural Language Processing (NLP). Their objective is to discover the characteristics interpretable by humans, which are encoded in the representation obtained by neural networks [5]. Probing is usually a simple classifier applied to trained representations like word embeddings. The probing classifier predicts whether the linguistic phenomenon that we want to verify exists or not. The probing classifiers in the NLP research community are popular tools for inspecting the internals of representations. However, some recent work extends the usability of probing tasks by introducing the concept of amnesic probing [10] to measure the influence of the phenomenons on the target task performance.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

Although probing tasks are popular in NLP, they only recently have been adapted to the Computer Vision (CV) domain in [4] based on the mapping defined between NLP and CV domains. These visual probing tasks allow one to gain intuition about the knowledge conveyed in the representation by the various self-supervised methods. However, there is no clear consensus on their impact on the target task performance.

3 Methods

This section introduces amnesic visual probing (AVP), a tool for explaining visual representations. It analyzes how important are particular visual concepts for a target task. Therefore, to define AVP, we first provide visual concepts (here called Visual Words, VW) and then obtain their meaning. Finally, we remove information about VW from the representation and analyze how it influences a target task.

Generating visual words To generate visual words, we use the established ACE algorithm [12]. It starts by dividing the image into superpixels using the SLIC algorithm [1]. Because different superpixel sizes are preferred, we run the algorithm three times with different parameters and obtain three sets with 15, 50, and 80 superpixels for each image. Then, we pass all the superpixels through the network trained on ImageNet to obtain their representations. These representations are clustered separately for each class using the k-means algorithm with $k = 25$ (infrequent and unpopular clusters are removed as described in [12]). Clusters obtained this way could be directly used as visual words. However, so many visual words would be impractical due to the similarity between concepts of ImageNet classes. Therefore, to obtain a credible dictionary with visual words shared between different classes, we filter out concepts with the smallest TCAV score [15] and cluster the remaining 6,000 ones using the k-means algorithm into $N = 50$ new clusters. These N clusters are visual words that form our visual language (see Fig. 2).

Cognitive vision systematic To obtain the meaning of the generated visual words, we use cognitive visual systematic [18] based on Marr's computational theory of vision [17]. According to Marr's theory, three levels of visual representations play an essential role in perception and discovering essential features of visible objects. These are the primal sketch, the 2.5D sketch, and the 3D model representation. The primal sketch is a two-dimensional image representation that uses light intensity changes, edges, colors, and textures. The 2.5D sketch represents mostly two-dimensional shapes, and the 3D model representation allows an observer to imagine the spatial object features based on its two-dimensional image. We will analyze six visual features from Marr's theory: brightness, color, texture, and lines (all primal sketch), shape (2.5D sketch), and form (3D model representation). We conduct user studies to establish the relationship between these features and individual visual words (see Fig. 3).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

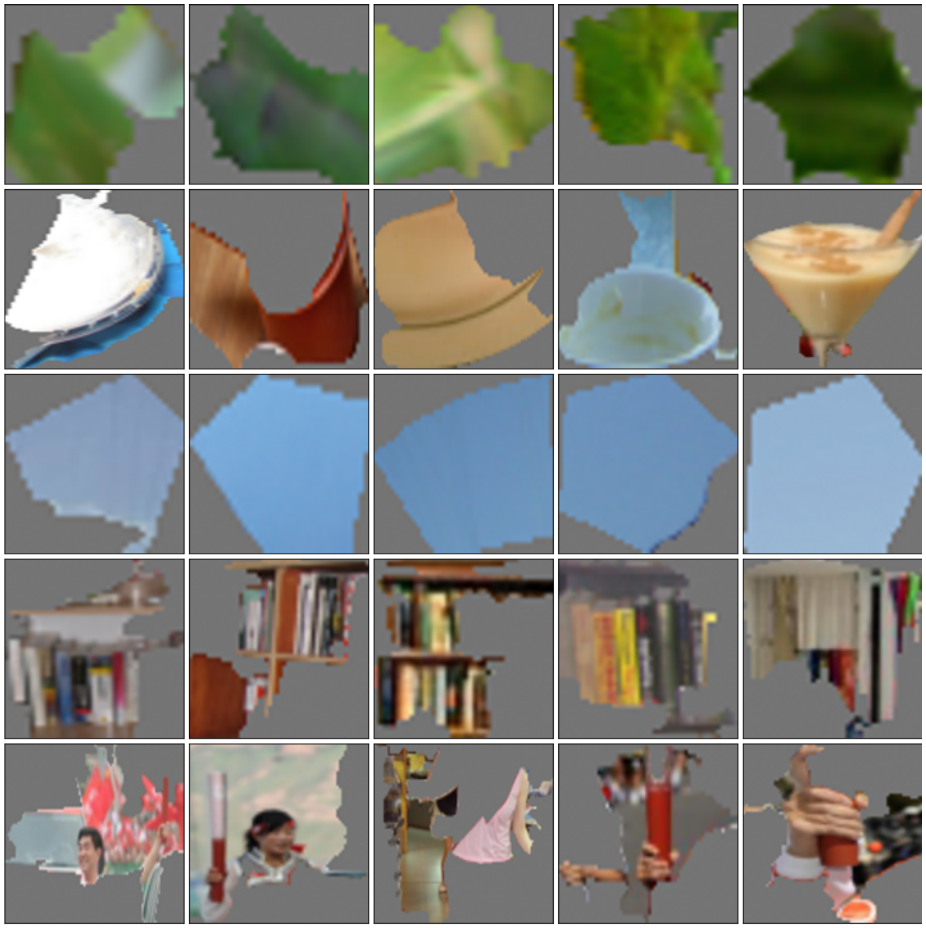


Fig. 2: Sample visual words, each represented by a row of 5 superpixels.

Amnesic visual probing We want to remove the information about a visual word from the representations and analyze how they differ from the original ones. For this purpose, we divide an image into superpixels, pass them through the network to obtain their representations, and assign them to the closest visual word. Then, we *define Word Content labels* $z_i \in \{0,1\}^N$ for representations $x_i \in \mathbb{R}^d$, where $z_i[j] = 1$ means that at least one superpixel of i-th image is assigned to j-th visual word.

Then, we *remove information about j-th visual word from a representation* x_i. For this purpose, we adapt an algorithm called Iterative Nullspace Projection (INLP) [19]. The probing classifier for $z_i[j]$ is parameterized by the matrix W_0. We first construct a projection matrix P_0 such that $W_0(P_0 x_i) = 0$ for all representations x_i (using method from [19]). Then, we iteratively train additional classifiers W_1 and perform the same procedure until no linear information re-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

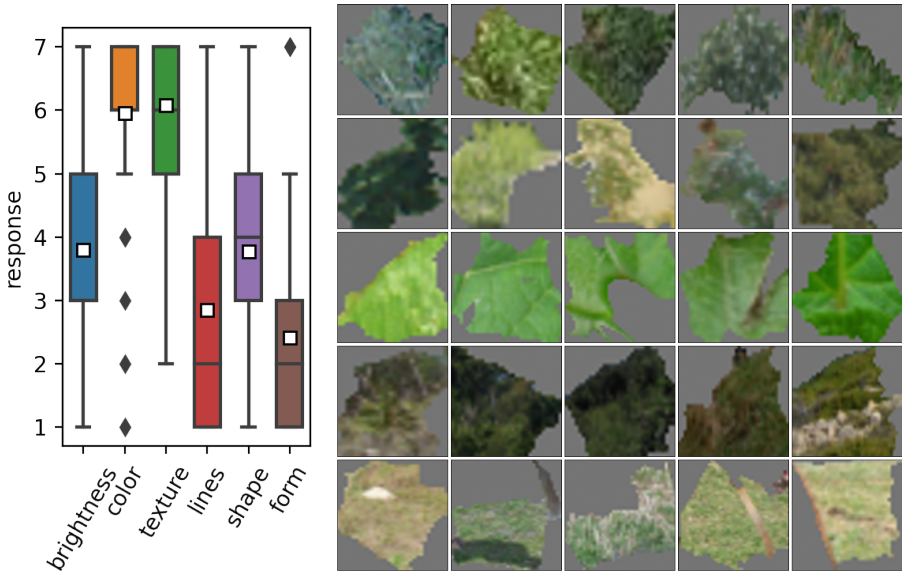


Fig. 3: Sample visual word (corresponding to grass) and its distributions of Likert scores obtained from user studies. One can observe that users mostly decided to assign this word to color and texture from the Marr's computational theory of vision.

garding $z_i[j]$ remains in x_i, i.e., until the chance of predicting the presence of a j-th visual word by the linear model is random. As a result we obtain a matrix $P_n \cdot P_{n-1} \cdot \ldots \cdot P_0$ which, when applied to representation, removes information about visual word j.

Finally, one can analyze changes in target task performance after removing information about a particular visual word. In this case, a target task is defined as multi-class classification with labels $y_i \in \{1, \ldots, k\}$, where $k = 1000$ is the number of ImageNet's classes. It is trained and tested for two types of representations, original and with removed visual word information.

4 User studies

To understand the meaning of visual words, we conduct user studies with 97 volunteers (64 males, 32 females, and 2 others aged 25 ± 7 years), including 71.1% students or graduates of computer science and related fields. Users completed an online survey with the number of questions corresponding to the number of visual words. We presented 12 typical (randomly chosen) superpixels for each visual word, and we asked to what extent a particular visual feature was essential for its creation. In reference to Marr's computational theory of vision [17] (see Section 3), six features were taken into consideration: brightness, color, texture,

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

Algorithm 1 Amnesic visual probing (AVP)

Require: X – set of image representations, Y – set of target labels, Z – set of visual words labels, C – codebook of visual words,
getNullSpaceProj(X, Z) – returns projection matrix that removes information about a visual word from representations,
trainValProb(X, Z) – trains model on probing task and returns validation accuracy,
trainValTarget(X, Y) – trains model on target task and returns validation accuracy
for each: $c \in$C
  $X_{proj} \leftarrow X$
  repeat
    $P \leftarrow$getNullSpaceProj(X_{proj}, Z)
    $X_{proj} \leftarrow PX_{proj}$
    $acc_{prob} \leftarrow$trainValProb(X_{proj}, Z)
  until $acc_{prob} \geq \frac{1}{2}$
  $acc_{target} \leftarrow$trainValTarget(X, Y)
  $acc_{target}^{\neg c} \leftarrow$trainValTarget$(X_{proj}, Y)$
  $influence^c = acc_{target}^{\neg c} - acc_{target}$

lines (all primal sketch), shape (2.5D sketch) and form (3D model representation). We use the Likert scale with seven numerical responses from 1 to 7, corresponding to insignificant and key features, respectively.

Before completing the survey, users got familiarized with the examples of visual words with particular features selected by a trained cognitivist. They also completed two training trials to become familiar with the main task. Moreover, completing the task was not limited in time. Finally, due to the high number of visual words, assessing all 50 visual words would be tedious for the users. Therefore, we have prepared four questionnaire versions (one with twenty visual words and three with ten visual words) and assigned them to users randomly.

Based on the user studies results, we ranked the most representative visual words for each of the six features: brightness, color, texture, lines, shape, and form. We used those rankings to obtain detailed results of the amnesic visual probing.

5 Experimental Setup

Models We examine four self-supervised methods (MoCo v1 [14], SimCLR v2 [8], BYOL [13], and SwAV [6]), with a publicly available implementation based on the ResNet-50 (1x) architecture, trained on the entire ImageNet dataset[5]. We use the penultimate layer of ResNet-50 to generate representations with a length of 2048.

[5] We use the following implementations of the self-supervised methods: https://github.com/{google-research/simclr, yaox12/BYOL-PyTorch, facebookresearch/swav, facebookresearch/moco}.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

Data and target task We consider ImageNet [9] classification as the target task, but our approach could also be applied to other tasks. In order to get the classification model, we freeze the self-supervised trained model and fine-tune an ultimate fully-connected layer for 100 epochs. We conduct our experiments with a standard train/validation split.

Removing visual words Interventions that remove visual words are parametrized by 2048×2048 matrices applied to self-supervised representations. We obtain these matrices with our adaptation of the INLP algorithm, where we iterate until the probing classifier (detecting a visual word) achieves random accuracy.

Metric We consider the difference in top-5 classification accuracy before and after the intervention. For each self-supervised method, we carry out a series of interventions, removing the information about successive visual words from the ranking obtained based on the user studies (see Section 4). For each of the six features, we start with visual words considered as crucial for a given feature.

6 Results

As shown in Table 1, *removing visual words from self-supervised representations reduces the top-5 accuracy* of the target task. It is expected because, as presented in [4], image representation contains semantic knowledge. However, depending on a self-supervised model and a type of visual word, the level of degradation significantly differs. In the case of SimCLR v2, visual words related to the shape and form have the most significant influence on the classifier decisions. For BYOL, brightness and form have the greatest influence. Results for SimCLR and BYOL are also similar because they are least sensitive to texture removal from the representations. In contrast, MoCo and SwAV are the least sensitive to removing shape. In the case of MoCo, we also observe the most significant decrease in classification accuracy when removing forms, while the performance of SwAV is the most sensitive to color removal.

In Fig. 4, we present the most important visual words (according to our user studies) for each of the six visual concepts from Marr's computational theory of vision. These are visual words that we first remove from the representation.

In general, except for MoCo v1, representations are the least sensitive to removing textures from representations, which is inconsistent with what is found in [11]. Also, the two-dimensional shape is the most influential feature only for the classifier using the SimCLR v2 model. On the other hand, on average, removing visual words corresponding to the three-dimensional form and color from the self-supervised representation causes the most significant drop in the classification accuracy.

In Fig. 5, we present the change of target task accuracy when removing the successive most important visual words of the considered Marr's visual features (obtained with user studies). In general, the classification accuracy decreases as we remove the successive visual words. There are only a few exceptions to this,

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

Table 1: Removing visual words from the self-supervised representations influences the top-5 accuracy. The results are presented for six visual concepts from Marr's computational theory of vision. For each visual feature we remove five visual words according to the ranking obtained based on the user studies. The colors denote higher (dark blue) or lower (light blue) accuracy drop (in percentage points). These results demonstrate the biases in the self-supervised representations.

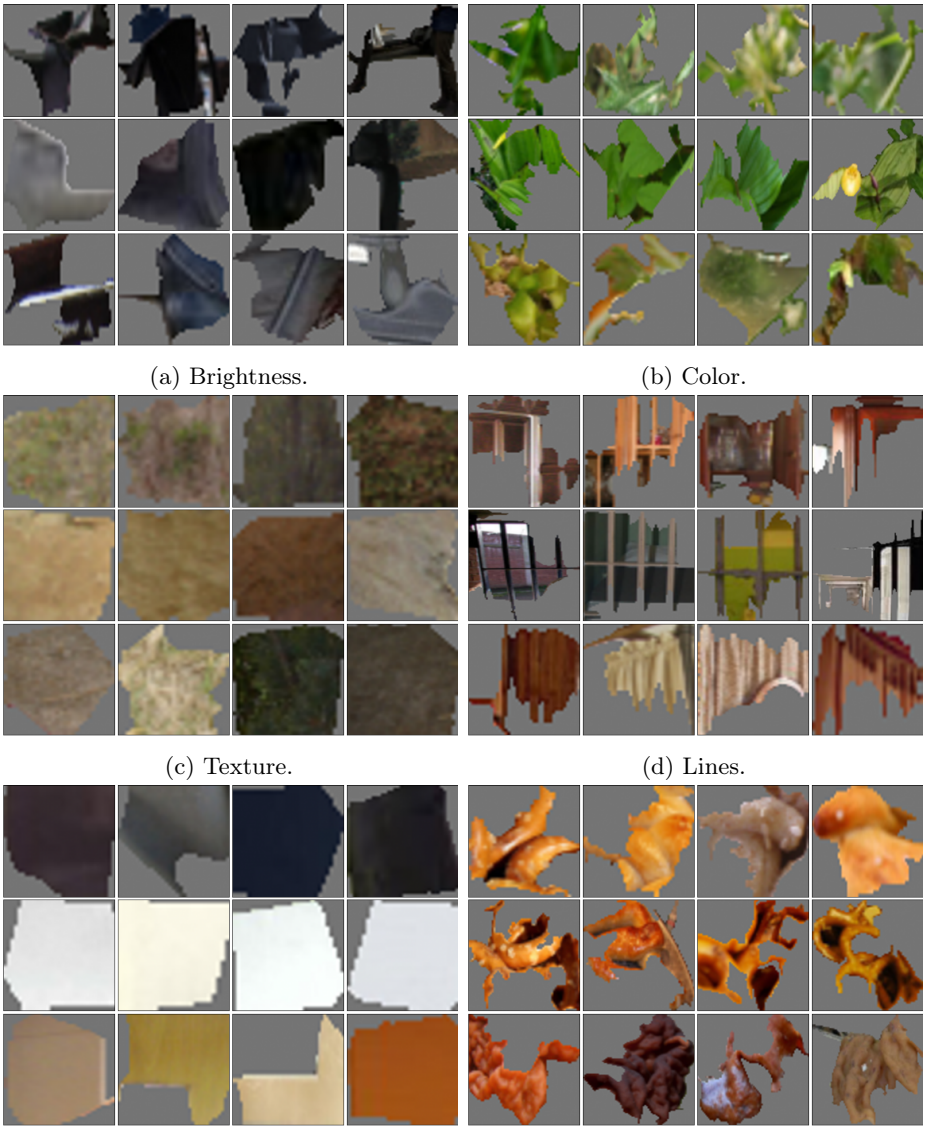
	top-5 acc.	**decrease in top-5 acc.**					
	no interv.	**remove visual words**					
		bright.	**color**	**texture**	**lines**	**shape**	**form**
MoCo v1	82.5	−3.09	−4.27	−3.84	−4.04	−2.98	−4.73
SimCLR v2	86.0	−2.00	−2.44	−1.60	−1.68	−2.51	−2.51
BYOL	86.5	−3.99	−3.37	−2.35	−2.75	−2.36	−3.49
SwAV	92.4	−2.20	−2.94	−1.56	−1.85	−1.00	−2.09

most notable in the case of SimCLR v2. We notice that in some cases, after removing two or three visual words from a given category, deleting the next ones causes only a slight further decrease in accuracy. It happens, for example, when removing visual words related to shape from SwAV representations or texture from SimCLR v2 representation. We also notice that in the case of three models (except MoCo v1), initially, when removing a small number of visual words, the most significant loss of accuracy occurs when removing the simplest visual features such as brightness (BYOL and SwAV) and color (SwAV and SimCLR v2). However, as we remove more visual words, the impact of removing more complex visual words corresponding to three-dimensional forms increases. This result may be because three-dimensional forms are more diverse and heterogeneous than colors and brightness.

Amnesic visual probing vs. Word Content probing task The correlation between the results of amnesic visual probing and the Word Content (WC) probing task is relatively weak, as presented in Fig. 6. The Pearson correlation coefficient ranges from 0.14 for SimCLR v2 to 0.52 for MoCo v1. In Fig. 6 we can see that although the WC probing task shows that there is a similar level of information about the visual words corresponding to lines and forms in SimCLR's representation, removing forms from this representation causes a much more significant decrease of target task accuracy than removing lines. The same relationship regarding lines and forms is also valid for BYOL, in which case the correlation between target task accuracy and WC results is the largest among the examined methods, even though it is still weak.

In general, this weak correlation supports the thesis that the WC probing task focuses on what visual words are encoded in the representation, but it does not assess how this information is used. Therefore, we conclude that the *Word Content probing task cannot be directly used to evaluate target task accuracy,*

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24



(a) Brightness. (b) Color.

(c) Texture. (d) Lines.

(e) Shape (2D). (f) Form (3D).

Fig. 4: The most important visual words (according to our user studies) for each of the six visual concepts from Marr's computational theory of vision.

which justifies the introduction of amnesic visual probing. Nevertheless, WC is still needed for amnesic visual probing to analyze the representation and should be considered as a complementary tool.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

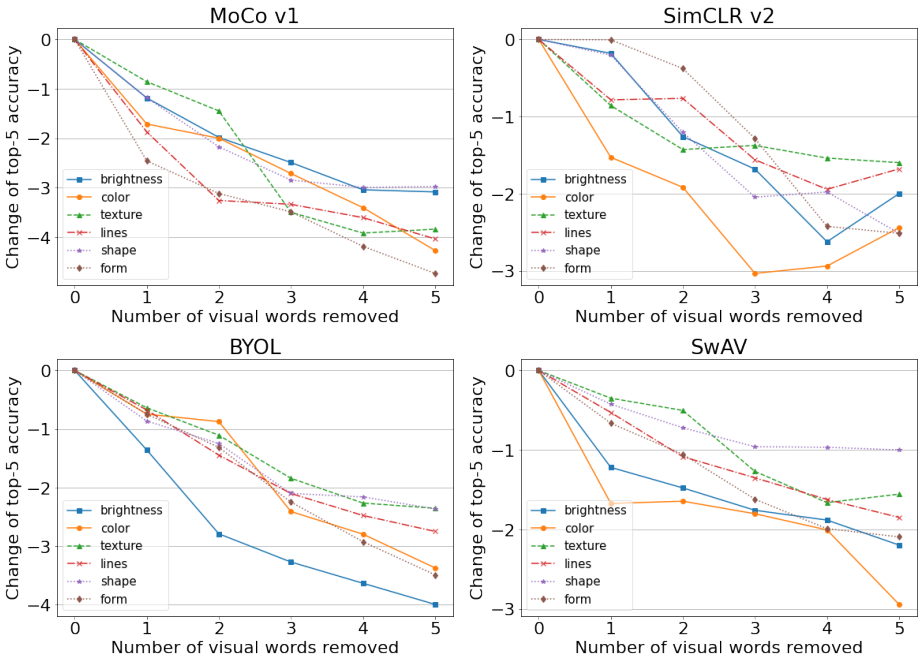


Fig. 5: Decrease in top-5 accuracy (in percentage points) when removing the information about successive visual words according to the ranking obtained based on the user studies, presented for six visual concepts from Marr's computational theory of vision.

7 Conclusions

The visual probing framework provides interesting insight into the self-supervised representations. However, this insight does not correspond to the performance of the target task. Hence, we propose Amnesic Visual Probing (AVP) to analyze the visual concepts that influence the target task. Thanks to preserving the semantic taxonomy of visual words from the visual probing framework, we can use AVP to examine and compare the biases of individual self-supervised methods. Finally, the user studies allow us to describe those biases using six visual features from Marr's computational theory of vision.

Acknowledgments

This research was funded by Foundation for Polish Science (grant no POIR.04.04.00-00-14DE/18-00 carried out within the Team-Net program co-financed by the European Union under the European Regional Development Fund), National Science Centre, Poland (grant no 2020/39/B/ST6/01511). The authors have applied a CC BY license to any Author Accepted Manuscript (AAM) version aris-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

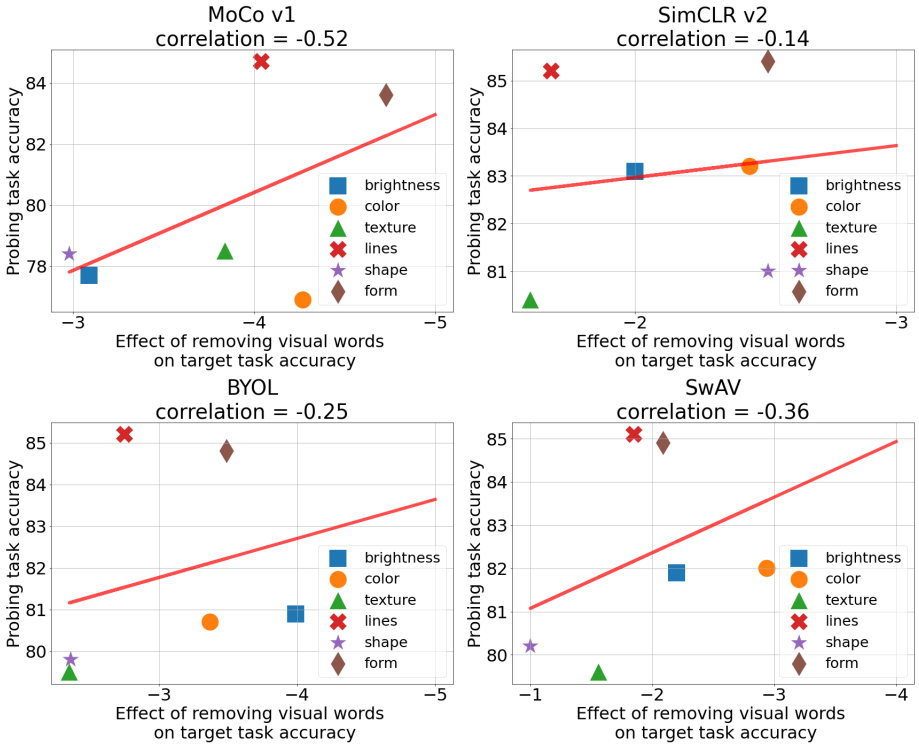


Fig. 6: There is a weak correlation between the results of amnesic visual probing (in percentage points) and the Word Content (WC) probing task (in percents). It means that WC cannot be directly used to evaluate target task accuracy. Hence introducing the amnesic visual probing is justified.

ing from this submission, in accordance with the grants' open access conditions. Dominika Basaj was financially supported by grant no 2018/31/N/ST6/02273 funded by National Science Centre, Poland.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2274–2282 (2012)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. arXiv preprint arXiv:1810.03292 (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
4. Basaj, D., Oleszkiewicz, W., Sieradzki, I., Górszczak, M., Rychalska, B., Trzcinski, T., Zieliński, B.: Explaining self-supervised image representations with visual probing. In: IJCAI-21. pp. 592–598 (8 2021). https://doi.org/10.24963/ijcai.2021/82

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

5. Belinkov, Y., Glass, J.: Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics **7**, 49–72 (2019)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020)
8. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 22243–22255. Curran Associates, Inc. (2020)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Elazar, Y., Ravfogel, S., Jacovi, A., Goldberg, Y.: Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. Transactions of the Association for Computational Linguistics **9**, 160–175 (03 2021)
11. Geirhos, R., Narayanappa, K., Mitzkus, B., Bethge, M., Wichmann, F.A., Brendel, W.: On the surprising similarities between supervised and self-supervised models. arXiv preprint arXiv:2010.08377 (2020)
12. Ghorbani, A., Wexler, J., Zou, J., Kim, B.: Towards automatic concept-based explanations. arXiv preprint arXiv:1902.03129 (2019)
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 21271–21284. Curran Associates, Inc. (2020)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9726–9735 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
17. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Co., Inc., New York, NY, USA (1982)
18. Oleszkiewicz, W., Basaj, D., Sieradzki, I., Górszczak, M., Rychalska, B., Lewandowska, K., Trzcinski, T., Zielinski, B.: Visual probing: Cognitive framework for explaining self-supervised image representations. CoRR **abs/2106.11054** (2021), https://arxiv.org/abs/2106.11054
19. Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., Goldberg, Y.: Null it out: Guarding protected attributes by iterative nullspace projection. In: Proceedings

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24

of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7237–7256. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.647
20. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
21. Sivic, J., Zisserman, A.: Video google: Efficient visual search of videos. In: Toward category-level object recognition, pp. 127–144. Springer (2006)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_24