Discriminability-enforcing loss to improve representation learning

Florinel-Alin Croitoru Diana-Nicoleta Grigore Radu Tudor Ionescu

University of Bucharest, Romania

Abstract

During the training process, deep neural networks implicitly learn to represent the input data samples through a hierarchy of features, where the size of the hierarchy is determined by the number of layers. In this paper, we focus on enforcing the discriminative power of the high-level representations, that are typically learned by the deeper layers (closer to the output). To this end, we introduce a new loss term inspired by the Gini impurity, which is aimed at minimizing the entropy (increasing the discriminative power) of individual high-level features with respect to the class labels. Although our Gini loss induces highly-discriminative features, it does not ensure that the distribution of the high-level features matches the distribution of the classes. As such, we introduce another loss term to minimize the Kullback–Leibler divergence between the two distributions. We conduct experiments on two image classification data sets (CIFAR-100 and Caltech 101), considering multiple neural architectures ranging from convolutional networks (ResNet-17, ResNet-18, ResNet-50) to transformers (CvT). Our empirical results show that integrating our novel loss terms into the training objective consistently outperforms the models trained with cross-entropy alone, without increasing the inference time at all.

1. Introduction

Learning good data representations is a crucial point towards improving the performance of deep neural networks. Therefore, discovering new methods [2, 4, 11, 15, 16, 20, 23, 27, 28] to improve this capability is an important goal for the research community. *Representation learning* [1, 9] is a vast domain studying this direction, and it covers multiple topics ranging from the design of neural architectures [10, 13, 14, 17] to the development of learning paradigms [2, 4, 19, 26, 27].

Our study fits in this domain, as our focus is on learning better high-level representations by enforcing their discriminability towards the target classes, in the context of supervised learning.

During the training process, deep neural networks implicitly learn to represent the input data samples through a hierarchy of features. As shown in [30], the features closer to the input tend to encode low-level information, e.g. edges, corners, stains, while those closer to the output tend to encode high-level information, e.g. object parts or even entire objects. While the low-level features are generic to all object classes, the high-level features should be specialized in discriminating between object classes. To further boost the discriminative power of the high-level representations, we propose to add two new loss terms to the objective function. The first loss term relies on the Gini impurity to minimize the entropy of individual high-level features with respect to the class labels. The entropy of the features (activation maps) gets minimized as they become more discriminative towards certain classes. Although our Gini loss induces highly-discriminative features, it does not ensure that the distribution of the high-level features matches the distribution of the classes. In other words, it might lead to a disproportionate number of features being specialized on a certain class. As such, we introduce another loss term to minimize the Kullback–Leibler (KL) divergence between the feature and the class distributions.

We conduct experiments on two image classification data sets (CIFAR-100 and Caltech 101), considering multiple neural architectures ranging from convolutional networks (ResNet-17, ResNet-18, ResNet-50) [12] to transformers (CvT) [29]. Our empirical results show that integrating our novel loss terms into the training objective consistently outperforms the models trained with cross-entropy alone. The accuracy gains come at no computational cost during inference, making the use of deeper and more computationally intensive models unnecessary.

Contribution. In summary, our contribution is threefold:

- We introduce a new loss based on the Gini impurity, which boosts the discriminative power of high-level features.
- We introduce a new loss based on the KL divergence, which ensures that the distribution of the high-level representation matches the class distribution.
- We present empirical evidence showing the benefits of our approach across neural architectures and data sets.

2. Related work

Representation learning is a vast research topic [1, 9] and, through deep learning methods, it has attained significant progress in various domains, such as computer vision [12, 29] and natural language processing [6, 22]. Aside from these achievements, learning good representations is important because they unlock solutions for other scenarios via transfer learning [11, 20, 23], domain adaptation [3, 8] or modality learning [24, 25]. What constitutes a good representation is described in [1, 9] through a set of properties.

A considerable effort in learning better representations has been made in unsupervised settings, where architectures such as auto-encoders [14], variational auto-encoders (VAE) [13, 17], and generative adversarial networks (GANs) [5, 10, 21] have been proposed. Deep convolutional GAN (DCGAN) [21], InfoGAN [5] and $\beta-$VAE [13] are able to learn expressive representations with multiple explanatory factors [1], such as emotion and hairstyle in images of human faces, or digit type and rotation in MNIST images. However, despite these improvements, in computer vision, the state-of-the-art solutions in image recognition [12, 29] still employ supervised transfer learning to learn better features. Semi-supervised learning [19, 26] utilizes both labeled and unlabeled data for learning representations. Other approaches, such as pre-trained language models [6, 22] and contrastive learning frameworks [2, 4, 27], fall under the self-supervised paradigm. Interestingly, the method presented in [2] is very close in terms of image classification performance to the supervised counterpart. In natural language processing, a common technique is to pre-train language models on an extensive unsupervised corpus to learn language representations, and then fine-tune them on supervised tasks. Popular examples such as BERT [6] and GPT [22] brought notable improvements on multiple language understanding tasks.

Closer to our work, there are several methods that use other objective functions besides cross-entropy to impose certain properties on the learned features [15, 16, 28]. In [16], the contrastive loss from the self-supervised case [2] is adapted to supervised scenarios, the inflicted properties being smoothness and space coherence [1, 9]. The same can be said about other works [15, 28], where the additional center loss penalizes a large variation within each class. Our loss function is distinct from related methods by not being directly connected to the distances or similarities in the representation space. Instead, it simply enforces the high-level features to be discriminative across classes. To the best of our knowledge, we are the first to employ a discriminability-enforcing loss to learn better high-level features.

3. Method

The cross-entropy loss is broadly employed as an objective function for training deep classifiers, being able to push any model to learn separable representations across classes. However, the training objective is not necessarily the most suitable for generalizing to unseen data, because the implicitly induced separability can be fragile. We address this shortage by optimizing an auxiliary loss function that can compel any model to learn more discriminative representations, meaning more class-oriented representations. Further in this section, we describe our approach comprising two novel components added to the loss function, while also explaining their roles.

The Gini impurity is often used to estimate the discriminative power of features in random forests, when a feature needs to be selected for a certain node while building a decision tree. Inspired by this choice, we propose to introduce a loss based on the Gini impurity to enforce the discriminative power of a set of convolutional activation maps $A \in \mathbb{R}^{m \times c \times h \times w}$, where m is the number of training data samples, c is the number of channels, and h and w are the height and width of an activation map. Our Gini loss is defined as follows:

$$\mathcal{L}_{\text{Gini}} = \frac{1}{c} \sum_{i=1}^{c} \left(1 - \sum_{j=1}^{n} \bar{s}_{ij}^2 \right), \tag{1}$$

where n represents the number of classes, and $\bar{s}_{ij}$ are elements of the matrix $\bar{S}$ having $c \times n$ components, whose components are probabilities computed as follows:

$$\bar{s}_{ij} = \frac{s_{ij}}{\sum_{k=1}^{n} s_{ik}}, \forall i \in \{1, 2, ..., c\}, j \in \{1, 2, ..., n\}. \tag{2}$$

Here, s_{ij} represents the average activation on channel i for class j. The average activations s_{ij} are elements of the matrix S of size $c \times n$, which is computed as follows:

$$S = M^{\top} \cdot \bar{Y}, \tag{3}$$

where M is a matrix of $m \times c$ components resulted after applying global average pooling over the set of activation maps A, and $\bar{Y}$ is a matrix of $m \times n$ components denoted as $\bar{y}_{ij}$, which are computed as follows:

$$\bar{y}_{ij} = \frac{y_{ij}}{\sum_{k=1}^{m} y_{kj}}, \forall i \in \{1, 2, ..., m\}, j \in \{1, 2, ..., n\}, \tag{4}$$

where y_{ij} are elements of the matrix Y of size $m \times n$ containing the target labels (as n-dimensional one-hot encodings) for the m training samples.

While constraining the activation maps to be more discriminative towards certain classes, $\mathcal{L}_{\text{Gini}}$ does not influence the distribution of the activation maps over classes. Hence, we might end up having many discriminative activation maps for some classes and none for other classes. To ensure that the distribution of discriminative activation maps matches the training class distribution, we propose to introduce a loss based on the Kullback–Leibler divergence, defined as:

$$\mathcal{L}_{\text{KL}} = \text{KL}\left(H^S, H^Y\right), \tag{5}$$

where H^S and H^Y are histograms of n bins representing the discrete distributions of the activations maps and the class labels, respectively. Each component $h_i^S \in H^S$ is computed by summing up and normalizing the average activations in $\bar{S}$, as follows:

$$h_i^S = \frac{\sum_{j=1}^{c} \bar{s}_{ji}}{\sum_{k=1}^{n}\sum_{j=1}^{c} \bar{s}_{jk}}, \forall i \in \{1, 2, ..., n\}. \quad (6)$$

Similarly, each component $h_i^Y \in H^Y$ is computed by summing up and normalizing the one-hot labels in Y, as follows:

$$h_i^Y = \frac{\sum_{j=1}^{m} y_{ji}}{m}, \forall i \in \{1, 2, ..., n\}. \quad (7)$$

In Eq. (7), we use the fact that the sum of all components in Y is equal to m (the number of training samples).

When integrating our approach into a neural model having its own loss function $\mathcal{L}_*$, our losses can simply be added to the respective loss, resulting in a new loss function that comprises all terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_* + \lambda_1 \cdot \mathcal{L}_{\text{Gini}} + \lambda_2 \cdot \mathcal{L}_{\text{KL}}, \quad (8)$$

where $\lambda_1, \lambda_2 \in \mathbb{R}^+$ are hyperparameters that control the importance of each loss term with respect to the original loss $\mathcal{L}_*$. For simplicity, we set $\lambda_1 = \lambda_2$ in our experiments.

4. Experiments

4.1. Data Sets

CIFAR-100. The CIFAR-100 data set [18] consists of 60,000 color images of 32×32 pixels. The images are grouped into 20 superclasses that are further divided into 100 mutually exclusive classes, each containing 500 training images and 100 test images. We kept a subset of 4,000 images (40 images per category) from the training set for validation.

Caltech 101. The Caltech 101 data set [7] contains a total of 9,146 pictures of objects from 101 different categories and a special background class. Each category includes between 40 and 800 images, with the more common or important classes being better represented. The images have varying sizes of around 300×200 pixels.

4.2. Evaluation Setup

In-domain and cross-domain evaluation. We conduct both in-domain and cross-domain experiments to exhibit the generalization capacity of our approach across different scenarios. The in-domain experiments are conducted on CIFAR-100. To perform cross-database assessments of the neural models, we create a custom subset of Caltech 101, which contains all the categories from the intersection with CIFAR-100. Due to the imperfect automatic mapping between category names, we performed manual label matching. Finally, the resulting Caltech 101 subset consists of 12 object categories. For the cross-domain experiments, the models are trained on CIFAR-100 and evaluated on the Caltech 101 subset.

Evaluation measure. We quantify the performance of neural models in terms of the classification accuracy. We train and evaluate each neural network in 5 trials, reporting its average accuracy and standard deviation.

Baselines. We study the impact of our method on two types of deep architectures, namely residual neural networks (ResNets) [12] and Convolutional vision Transformers (CvT) [29]. ResNets [12] form a class of convolutional neural networks approaching the gradient propagation problem, which appears in very deep models, with residual connections implemented as identity or projection mappings. In this work, we employ three such models that vary in depth: ResNet-17, ResNet-18 and ResNet-50. We derive ResNet-17 from ResNet-18 by removing the fully connected layer at the end and adjusting the last convolutional layer to have the number of filters equal to the number of class labels, thus obtaining a lighter architecture.

CvT [29] is a transformer-based architecture that organizes the transformer blocks into stages, incorporating the convolution operation via two methods, called convolutional token embedding and convolutional projection. In our experiments, we employ the lighter CvT-7 architecture, which contains three stages. The first stage has one transformer block, the second has two blocks and the third has four blocks. In each stage, the convolutional token embedding layers have 32, 128 and 258 filters, respectively. The number of heads of the Multi-Head Self-Attention module also varies between stages: the blocks in the first stage have one head, those in the second stage have two heads, and those in the third stage have six heads.

Hyperparameter tuning. We trained both ResNet-17 and ResNet-18 for 200 epochs on mini-batches of 500 examples. We started with a learning rate of 10^{-3} and periodically reduced it by a factor of 0.5 after 10 epochs of no improvement, with a threshold of 10^{-2}. We used Adam to optimize both models. In a similar fashion, we trained ResNet-50 for 250 epochs on mini-batches of size 200. We optimized the model using stochastic gradient descent with momentum, setting the momentum rate to 0.9. We started with a learning rate of 10^{-1} and a weight decay of $5 \cdot 10^{-4}$. At epochs 60, 120, 160, and 200, we reduced the learning rate by a factor of 0.2. For CvT, we set the number of epochs to 200 and the mini-batch size to 200. We optimized CvT using AdaMax, with an initial learning rate of $2 \cdot 10^{-3}$. All baselines are trained using the categorical cross-entropy as the loss function $\mathcal{L}_*$.

Aside from setting the common hyperparameters mentioned above, our method requires setting additional values, such as the two weights that control the importance of our new loss terms. In a preliminary set of validation experiments, we observed that setting both λ_1 and λ_2 to 0.5

Figure 1. Examples from Caltech 101 that are wrongly classified by a ResNet-50 trained on CIFAR-100 using cross-entropy (baseline). The model is able to correctly label the samples upon introducing our novel loss. Best viewed in color.

Table 1. Accuracy scores (in %) on CIFAR-100 and Caltech 101. The baselines trained with cross-entropy alone are compared with our models trained with the loss defined in Eq. (8). Top scores for each architecture are highlighted in bold.

Model	CIFAR-100	Caltech 101
ResNet-17 (baseline)	71.88 ± 0.36	63.22 ± 0.68
ResNet-17 (ours)	$\mathbf{73.00} \pm 0.20$	$\mathbf{63.90} \pm 1.56$
ResNet-18 (baseline)	71.47 ± 0.36	63.04 ± 1.35
ResNet-18 (ours)	$\mathbf{72.34} \pm 0.13$	$\mathbf{64.33} \pm 0.84$
ResNet-50 (baseline)	76.45 ± 0.51	70.14 ± 1.26
ResNet-50 (ours)	$\mathbf{76.92} \pm 0.61$	$\mathbf{70.68} \pm 0.74$
CvT-7 (baseline)	63.34 ± 0.56	53.35 ± 1.26
CvT-7 (ours)	$\mathbf{63.76} \pm 0.25$	$\mathbf{53.79} \pm 0.89$

works reasonably well across different neural architectures. Another necessary configuration is the epoch index from which we begin to introduce $\mathcal{L}_{\text{Gini}}$ and $\mathcal{L}_{\text{KL}}$. We set the starting epoch to 10 in all our experiments. Our loss terms are applied on the last convolutional layer of each residual architecture. Analogously, for CvT, we employ them on the convolutional token embedding layer from the last stage.

4.3. Results

We present the results on CIFAR-100 and Caltech 101 in Table 1. First, we observe that all the models trained with our new loss perform better than their counterparts based solely on cross-entropy. In addition, our models yield better accuracy scores in the cross-domain setup, demonstrating their capability of learning more general representations than the baselines. Another remark is that the standard deviation is generally lower for our approach, suggesting that the results are more stable across multiple runs. We underline that the accuracy gains come at no additional computational cost during inference. Moreover, we observe that the lighter ResNet-17 architecture usually attains better performance than ResNet-18, indicating that more efficient models can also be more effective.

In Table 2, we present ablation results that show the effect of each loss term on the final accuracy. The ablation study indicates that using our loss terms independently leads to performance drops. The results demonstrate the importance of jointly using our novel loss terms, confirming the necessity of using both terms to achieve the desired accuracy gains.

Table 2. Ablation results for ResNet-18 on CIFAR-100, obtained by excluding various loss terms from Eq. (8). $\mathcal{L}_*$ represents the categorical cross-entropy.

$\mathcal{L}_*$	$\mathcal{L}_{\text{Gini}}$	$\mathcal{L}_{\text{KL}}$	Accuracy
✓			71.47 ± 0.36
✓	✓		69.87 ± 0.50
✓		✓	69.72 ± 0.38
✓	✓	✓	72.34 ± 0.13

We illustrate a few qualitative results in Figure 1. The most noteworthy examples are the second and third images (counting from left to right), which show that the baseline confuses two similar classes, lobster and crab, while our model is able to correctly tag the respective examples.

5. Conclusion

In this paper, we proposed a novel approach to boost the discriminative power of high-level representations in convolutional and transformer architectures. Our approach is based on jointly integrating two novel loss terms, one that enforces the discriminability of individual features and another that ensures the alignment between the high-level feature distribution and the class distribution. We presented empirical evidence indicating that our approach brings accuracy gains for multiple neural architectures across different evaluation scenarios.

In future work, we aim to evaluate our approach on further benchmarks and neural architectures. We also aim to extend our approach to regression tasks, which could require modifying our loss terms.

Acknowledgments

The research leading to these results has received funding from the NO Grants 2014-2021, under project ELO-Hyp contract no. 24/2020. This article has also benefited from the support of the Romanian Young Academy, which is funded by Stiftung Mercator and the Alexander von Humboldt Foundation for the period 2020-2022.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1, 2

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of NeurIPS*, volume 33, pages 9912–9924, 2020. 1, 2

[3] Minmin Chen, Zhixiang Xu, Kilian Q. Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of ICML*, pages 1627–1634, 2012. 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, volume 119, pages 1597–1607, 2020. 1, 2

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Proceedings of NIPS*, volume 29, pages 2180–2188, 2016. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, 2019. 2

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *Proceedings of CVPRW*, 2004. 3

[8] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of ICML*, pages 513–520, 2011. 2

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 1, 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of NIPS*, volume 27, pages 2672–2680, 2014. 1, 2

[11] Ian J. Goodfellow, Aaron Courville, and Yoshua Bengio. Spike-and-slab sparse coding for unsupervised feature discovery. In *Proceedings of NIPS Workshop on Challenges in Learning Hierarchical Models*, 2011. 1, 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*, pages 770–778, 2016. 1, 2, 3

[13] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *Proceedings of ICLR*, 2017. 1, 2

[14] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 1, 2

[15] Xiuyu Huang, Nan Zhou, and Kup-Sze Choi. A Generalizable and Discriminative Learning Method for Deep EEG-Based Motor Imagery Classification. *Frontiers in Neuroscience*, 15, 2021. 1, 2

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of NeurIPS*, volume 33, pages 18661–18673, 2020. 1, 2

[17] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of ICLR*, 2014. 1, 2

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 3

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proceedings of ICLR*, 2017. 1, 2

[20] Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, Pascal Vincent, Aaron Courville, and James Bergstra. Unsupervised and transfer learning challenge: a deep learning approach. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27, pages 97–110, 2012. 1, 2

[21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of ICLR*, 2016. 2

[22] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 2

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 1, 2

[24] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learning Research*, 15(84):2949–2980, 2014. 2

[25] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of EMNLP*, pages 5100–5111, 2019. 2

[26] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of NIPS*, volume 30, pages 1195–1204, 2017. 1, 2

[27] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2

[28] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of ECCV*, pages 499–515, 2016. 1, 2

[29] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of ICCV*, pages 22–31, 2021. 1, 2, 3

[30] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of ECCV*, pages 818–833, 2014. 1