Explaining the Transfer Learning Ability of a Deep Neural Networks by Means of Representations

German Magai[1], Artem Soroka[2]

[1] HSE University, Moscow, Russia
[2] National Research Nuclear University MEPhI, Moscow, Russia

Abstract

The basis of transfer learning methods is the ability of deep neural networks to use knowledge from one domain to learn in another domain. However, another important task is the analysis and explanation of the internal representations of deep neural networks models in the process of transfer learning. Some deep models are known to be better at transferring knowledge than others. In this re-search, we apply the Centered Kernel Alignment (CKA) method to analyze the in-ternal representations of deep neural networks and propose a method to evaluate the ability of a neural network architecture to transfer knowledge based on the quantitative change in representations during the learning process. We introduce the Transfer Ability Score (TAs) measure to assess the ability of an architecture to effectively transfer learning. We test our approach using Vision Transformer (ViT-B/16) and CNN (ResNet, DenseNet) architectures in computer vision tasks in several datasets, including medical images. Our work is a contribution to the field of explainable AI and an attempt to explain the learning transfer process.

Keywords

Transfer learning, knowledge representation

1. Introduction

Excellent results of deep learning models are mainly achieved by fine-tuning models that are pre-trained on Big Data. Knowledge transfer is one of key approaches to achieving high performance. Models learn to transfer and generalize knowledge in one data field (target domain) using information obtained in another one (source domain). In our work, we consider the process of knowledge transfer from the point of view of the similarity of feature representations. The contribution of this research can be divided into several points. First, we propose a method for evaluating the ability of a particular deep neural network (DNN) architecture to transfer knowledge to a new domain. We also compare different DNN architectures (ViT and CNNs) on different tasks in terms of the ability to transfer knowledge and explore the dynamics of the similarities of internal features representations in the process of fine tuning.

2. Related work

Various methods are used to evaluate the similarity of neural representations: Linear-Reg [1], SVCCA[2], PWCCA[3], HSIC[4], but the most common is the Central Kernel Alignment (CKA) method. The CKA analysis in [5] shows the block structure of CNN. The paper [6] notes the fundamental differences between ViT and CNN in terms of the similarity of representation. There are many works that explore the problem of knowledge transfer [7–11]. In [12–14] it is argued

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

gera128a@gmail.com (G. Magai); copoka11@gmail.com (*A. Soroka*)

https://www.hse.ru/org/persons/364631586 (G. Magai);

that ViT has better transfer learning performance than CNN in the medical imaging task. LEEP, NCE, LogMe, OTCE [15–19] have been proposed to assess the transfer knowledge ability of a DNN.

3. First level heading

The deep neural network $DNN_\theta(x_i) = y_i$ is mapping from the example x_i space to the class labels y_i space. $DNN=f_L \circ ... \circ f_1$. where functions f_i, $1 \le i \le L$, are called layer functions, θ is a set of parameters. The design paradigms of modern DNN model architectures are divided into architectures based on the convolution (CNN) [20] and self-attention (ViT) [21]. Due to the large number of existing DNN architectures, the question arises as to whether each one is suitable for efficient transfer learning. Let $X, Y \in R^{n \times d}$ denote the 2 sets of neural activations of layer i and j of the DNN model with $d = d_1$ and d_2 neurons respectively in response to a batch of n examples. The measure CKA ∈ [0,1] shows how sets X and Y are similar to each other. The CKA is based on the principle of the Hilbert-Schmidt Independence Criterion (HSIC) [22, 23]:

$$HSIC(X,Y) = \frac{1}{(n-1)^2} tr(XX^T YY^T) = |cov(X^T Y^T)|_F^2, \quad (1)$$

where tr is the trace matrix, cov is the covariance matrix, F is the Frobenius norm, n – the number of examples in a batch. Linear CKA can be calculated as follows:

$$CKA(X,Y) = \frac{HSIC(X,Y)}{\sqrt{HSIC(X,X)HSIC(Y,Y)}}, \quad (2)$$

We propose a Transferability score (TAs) – a measure of the ability of a DNN to transfer knowledge to a new domain. Consider the problem of transferring knowledge by the model with architecture A_k from source domain D_S to target domain D_t. The adaptation to the D_t can be interpreted via evolving of the feature space on different layers. A slight change in feature representations on different layers during finetuning on domain D_t indicates that the DNN has a high ability to transfer knowledge to a new domain. In contrast significantly change shows that the information extracted from D_s is not enough to generalize knowledge to a new domain D_t, or the domains are very different and a substantial change in the learned features representation is required. A low TAs value is an indication of less parameter change during DNN training.

Table 1
Empirical comparison of DNN architectures on CIFAR-10 D_t.

Architecture	Test accuracy	Number of layers	TA score
ResNet-50	86.7	151	0.1737
ViT-B/16	**95.2**	**140**	**0.1528**
DenseNet-121	84.3	433	0.2574

Let $\{X_m\}_{m=1}^n = \{X_1, X_2 .. X_n\}$ is a set of representations for model DNN_X with n_1 layers trained on D_s and $\{Y_m\}_{m=1}^n = \{Y_1, Y_2 .. Y_n\}$ is a set of representations for model DNN_Y with n_2 layers fine-tuned on D_t. Let's define CKA matrix M_1, where m_{1ij} is the value of the $CKA(X_i, X_j)$ between the X representations on layers i and j. And CKA matrix M_2, where m_{2ij} is the value of the $CKA(X_i, Y_j)$ between the X and Y representations on layers i and j, respectively. Let's denote $M' = M_1 - M_2$, M' shows how much the representations on different layers have changed after fine-tuned on the target domain. m'_{ij} – i,j-th element of matrix M'. We estimate the ability of a model with A_k architecture to transfer knowledge (Transferability score – TAs) from the D_s domain to the D_t domain via a quantitative change in the feature space after fine-tuning and define it as $TAs = \sum_{i,j=1}^n |m'_{ij}| / n^2$. The m'_{ij} values show the absolute change in the similarities of representations. The lower the Transferability score, the greater the DNN model's ability to transfer knowledge. In addition, the M' matrix provides a visual understanding of how much the similarity of representations on different layers of the DNN has changed after fine tuning on data in the D_t.

4. Experiments

We test ResNet-50 [24], ResNet-101, DenseNet-121[25] and ViT-B/16 architecture models pre-trained on ImageNet-1k [26]. We analyze the ability of various DNN models to transfer knowledge to a new target domain on several datasets: Eurosat (ESAT) [27], PatchCamelyon (PCAM) [28], The Cars dataset [29], DTD [30], CIFAR-10 [31]. For DNN training we used Adam [32] stochastic optimizer, lr = $5 \cdot 10^{-5}$, batch size = 32.

The success of transfer learning depends on the similarity between the D_s and D_t: the more similar the data, the more effective the transfer of knowledge [8,33]. Difference between the CKA matrices showing the difference between the source and fine-tuned models for different D_t (Figure 1). ImageNet's D_s partially includes information contained in DTD, CIFAR-10, and Stanford cars, so the representations do not change as much as for PCAM and ESAT, which are very different from ImageNet. To adapt to the PCAM and ESAT domains, the DNN model needs to learn new feature representations, which is strongly reflected in the M' matrices. It can also be seen that the ViT-B/16 architecture changes representations less significantly than ResNet-50, which indicates that ViT-B/16 are able to extract more information from D_s and it is easier for ViT to adapt to D_t. This is consistent with the greater accuracy of ViT models in knowledge transfer than CNN models (Table 1).

The dynamics of TAs during fine-tuning to a new dataset shows that when the accuracy of the test stabilizes, the values of the TA score also stabilize (Figure 2). In ViT, we observe a slight change in representations, because when trained on D_s, the ViT model extracts more complete information from a large dataset and generalizes better, and when adapted to D_t, the adaptation of the feature space is not so significant [34], which is consistent with the lower value of TAs.

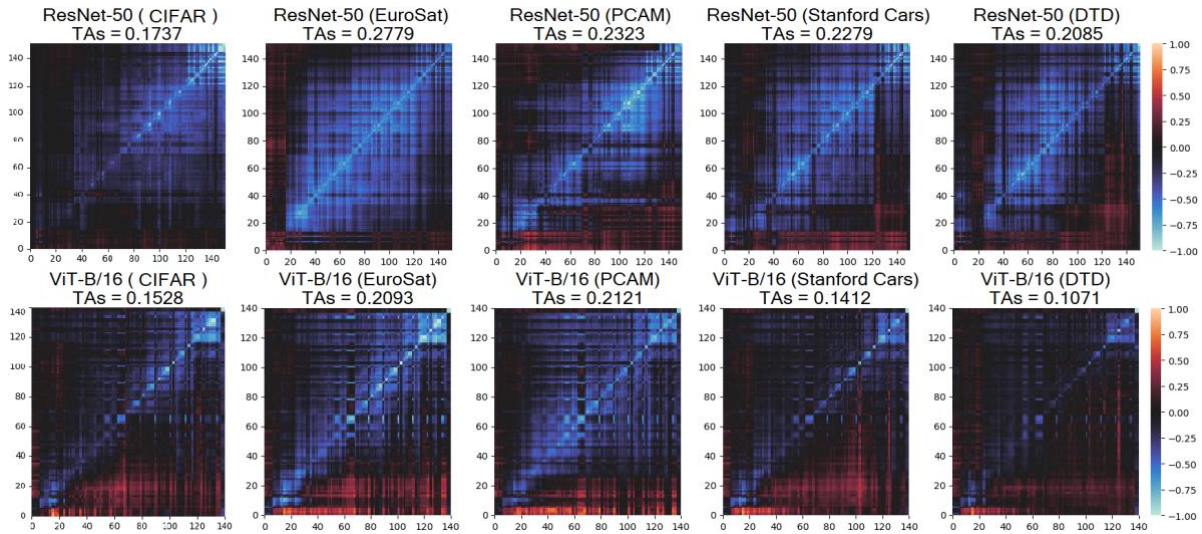

Figure 1: Differences in the representations of the M' matrices for different target domains.

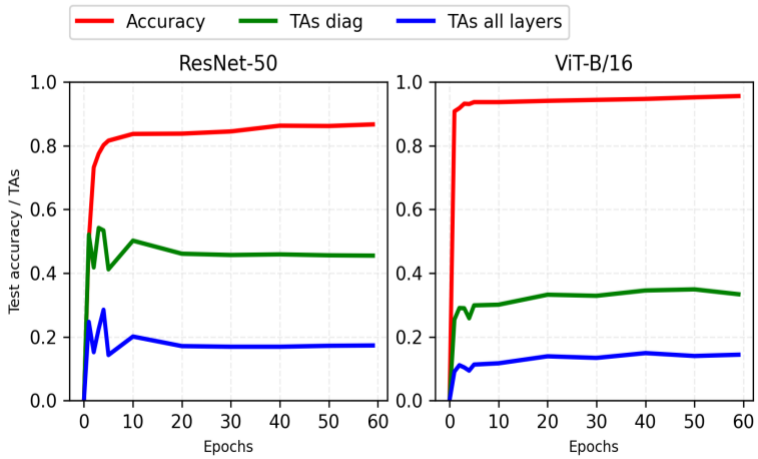

Figure 2: Dynamics of TAs in the fine-tuning process, ResNet-50 and ViT-B/16.

5. Discussion

In this paper, we touch upon the issue of interpreting the change in the similarity of internal features representations in the transfer learning process. We have proposed a method to evaluate the ability of a DNN architecture to transfer knowledge from the source domain to target domain based on similarity of feature representations. Experiments were performed for several architectures on different datasets. Based on TAs we can conclude the ViT architecture has a better ability to transfer knowledge than CNN models, which is consistent with previous research [7-9].

Improving our approach may be useful for choosing the optimal architecture. For future research, we propose to pay attention to the transfer of knowledge not only within the modality of images, but also cross-modality, for example, the use of features extracted from an image for an audio or text classification task.

Acknowledgements

The work of G. Magai was supported by the HSE University Basic Research Program. The work of A. Soroka was performed in the Tensor Processors laboratory of the Mephius Full-cycle Microelectronics Design Center (NRNU MEPhI) and IVA Technologies (HiTech).

References

[1] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2014.

[2] M. Raghu, J. Gilmer, J. Yosinski, J. Sohl-Dickstein, Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. Advances in neural in-formation processing systems, 2017.

[3] A. Morcos, M. Raghu, S. Bengio, Insights on representational similarity in neural networks with canonical correlation. Advances in Neural Information Processing Systems, 2018.

[4] W. D. K. Ma, J. P. Lewis, W. B. Kleijn, The HSIC bottleneck: Deep learning without back-propagation. In: Proceedings of the AAAI conference on artificial intelligence, 2020.

[5] S. Kornblith, M. Norouzi, H. Lee, G. Hinton, Similarity of neural network representations revisited. In: International Conference on Machine Learning (pp. 3519-3529). PMLR, 2019.

[6] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision trans-formers see like convolutional neural networks? Advances in Neural Information Processing Systems, 34, 12116-12128, 2021.

[7] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How Transferable Are Features in Deep Neural Networks? In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, pp. 3320–3328. MIT Press, 2014.

[8] I. Redko, E. Morvant, A. Habrard, M. Sebban, Y. Bennani, Advances in domain adaptation theory. Elsevier, 2019.

[9] L. Lin, G. Wang, W. Zuo, X. Feng, L. Zhang, Cross-domain visual matching via generalized similarity measure and feature learning. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1089-1102, 2016.

[10] J. D. Ferreira, F. M. Couto, Multi-domain semantic similarity in biomedical research. BMC bioinformatics, 20(10), 23-31, 2019.

[11] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029v2, 2020.

[12] M. Usman, T. Zia, A. Tariq, Analyzing transfer learning of vision transformers for interpreting chest radiography. Journal of digital imaging, 35(6), 1445-1462, 2022.

[13] J. Yang, Leveraging CNN and Vision Transformer with Transfer Learning to Diagnose Pigmented Skin Lesions. Highlights in Science, Engineering and Technology, 39, 408-412, 2023.

[14] G. Ayana, K. Dese, Y. Dereje, Y. Kebede, H. Barki, D. Amdissa, S. W. Choe, Vision-Transformer-Based Transfer Learning for Mammogram Classification. Diagnostics, 13(2), 178, 2023.

[15] C. Nguyen, T. Hassner, M. Seeger, C. Archambeau, Leep: A new measure to evaluate transferability of learned representations. In: International Conference on Machine Learning (pp. 7294-7305). PMLR, 2020.

[16] A. T. Tran, C. V. Nguyen, T. Hassner, Transferability and hardness of supervised classification tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1395-1405), 2019.

[17] Y. Bao, Y. Li, S. L. Huang, L. Zhang, L. Zheng, A. Zamir, L. Guibas, An information-theoretic approach to transferability in task transfer learning. In: 2019 IEEE international conference on image processing (ICIP) (pp. 2309-2313). IEEE, 2019.

[18] Y. Tan, Y. Li, S. L. Huang, Otce: A transferability metric for cross-domain cross-task representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-tern Recognition (pp. 15779-15788), 2021.

[19] K. You, Y. Liu, J. Wang, M. Long, Logme: Practical assessment of pre-trained models for transfer learning. In: International Conference on Machine Learning (pp. 12133-12143). PMLR, 2021.

[20] J. Gu, Z. Wang, J. Kuen, J., L. Ma, A. Shahroudy, B. Shuai, T. Chen, Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377, 2018.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, I. Polosukhin, Attention is all you need. Advances in neural information processing systems, 2017.
[22] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, A. Smola, A kernel statistical test of independence. Advances in neural information processing systems, 20, 2007.
[23] D. Greenfeld, U. Shalit, Robust learning with the hilbert-schmidt independence criterion. In: International Conference on Machine Learning (pp. 3759-3768). PMLR, 2020.
[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778), 2016.
[25] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708), 2017.
[26] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE, 2009.
[27] P. Helber, B. Bischke, A. Dengel, D. Borth, Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(7), 2217-2226, 2019.
[28] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling, Rotation equivariant CNNs for digital pathology. In: Medical Image Computing and Computer Assisted Intervention–MICCAI, 2018.
[29] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops, 2013.
[30] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi, Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014.
[31] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, 2009.
[32] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[33] E. Otović, M. Njirjak, D. Jozinović, G. Mauša, A. Michelini, I. Štajduhar, Intra-domain and cross-domain transfer learning for time series data — How transferable are the features? Knowledge-Based Systems, 239, 107976, 2022.
[34] J. Kim, K. Shim, J. Kim, B. Shim, Vision Transformer-Based Feature Extraction for Generalized Zero-Shot Learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE, 2023.