B-Cos Aligned Transformers Learn Human-Interpretable Features

Manuel Tran[1,2,3], Amal Lahiani[1], Yashin Dicente Cid[1], Melanie Boxberg[3,5], Peter Lienemann[2,4], Christian Matek[2,6], Sophia J. Wagner[2,3], Fabian J. Theis[2,3], Eldad Klaiman[1], and Tingying Peng[2(✉)]

[1] Roche Diagnostics, Penzberg, Germany
[2] Helmholtz AI, Helmholtz Munich, Neuherberg, Germany
tying84ster@gmail.com
[3] Technical University of Munich, Munich, Germany
[4] Ludwig Maximilian University of Munich, Munich, Germany
[5] Pathology Munich-North, Munich, Germany
[6] University Hospital Erlangen, Erlangen, Germany

Abstract. Vision Transformers (ViTs) and Swin Transformers (Swin) are currently state-of-the-art in computational pathology. However, domain experts are still reluctant to use these models due to their lack of interpretability. This is not surprising, as critical decisions need to be transparent and understandable. The most common approach to understanding transformers is to visualize their attention. However, attention maps of ViTs are often fragmented, leading to unsatisfactory explanations. Here, we introduce a novel architecture called the B-cos Vision Transformer (BvT) that is designed to be more interpretable. It replaces all linear transformations with the B-cos transform to promote weight-input alignment. In a blinded study, medical experts clearly ranked BvTs above ViTs, suggesting that our network is better at capturing biomedically relevant structures. This is also true for the B-cos Swin Transformer (Bwin). Compared to the Swin Transformer, it even improves the F1-score by up to 4.7% on two public datasets.

Keywords: transformer · self-attention · explainability · interpretability

1 Introduction

Making artificial neural networks more interpretable, transparent, and trustworthy remains one of the biggest challenges in deep learning. They are often still considered black boxes, limiting their application in safety-critical domains such

E. Klaiman—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43993-3_50.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14227, pp. 514–524, 2023.
https://doi.org/10.1007/978-3-031-43993-3_50

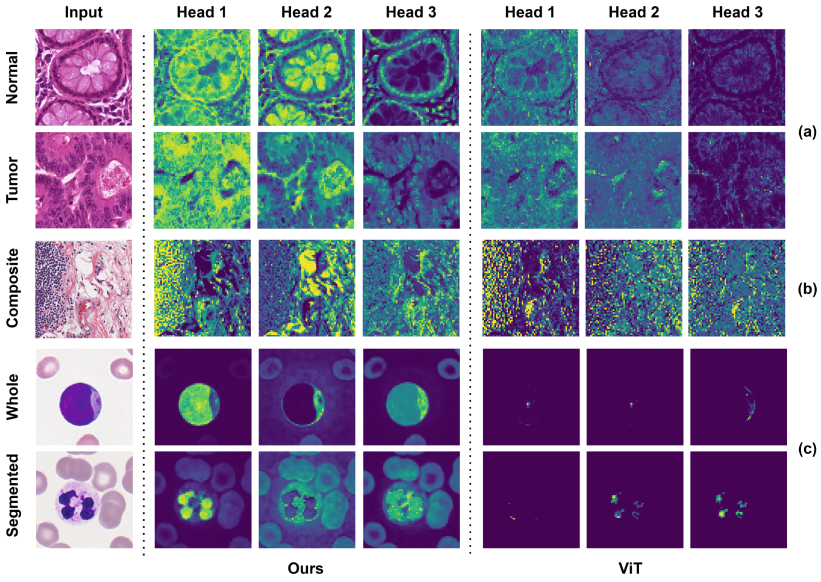


Fig. 1. Attention maps of ViT and BvT (ours) on the test set of (a) NCT-CRC-HE-100K, (b) TCGA-COAD-20X, and (c) Munich-AML-Morphology. BvT attends to various diagnostically relevant features such as cancer tissue, cells, and nuclei.

as healthcare. Histopathology is a prime example of this. For years, the number of pathologists has been decreasing while their workload has been increasing [23]. Consequently, the need for explainable computer-aided diagnostic tools has become more urgent.

As a result, research in explainable artificial intelligence is thriving [20]. Much of it focuses on convolutional neural networks (CNNs) [13]. However, with the rise of transformers [31] in computational pathology, and their increasing application to cancer classification, segmentation, survival prediction, and mutation detection tasks [26,32,33], the old tools need to be reconsidered. Visualizing filter maps does not work for transformers, and Grad-CAM [30] has known limitations for both CNNs and transformers.

The usual way to interpret transformer-based models is to plot their multi-head self-attention scores [8]. But these often lead to fragmented and unsatisfactory explanations [10]. In addition, there is an ongoing controversy about their trustworthiness [5]. To address these issues, we propose a novel family of transformer architectures based on the B-cos transform originally developed for CNNs [7]. By aligning the inputs and weights during training, the models are implicitly forced to learn more biomedically relevant and meaningful features (Fig. 1). Overall, our contributions are as follows:

- We propose the B-cos Vision Transformer (BvT) as a more explainable alternative to the Vision Transformer (ViT) [12].
- We extensively evaluate both models on three public datasets: NCT-CRC-HE-100K [18], TCGA-COAD-20X [19], Munich-AML-Morphology [25].

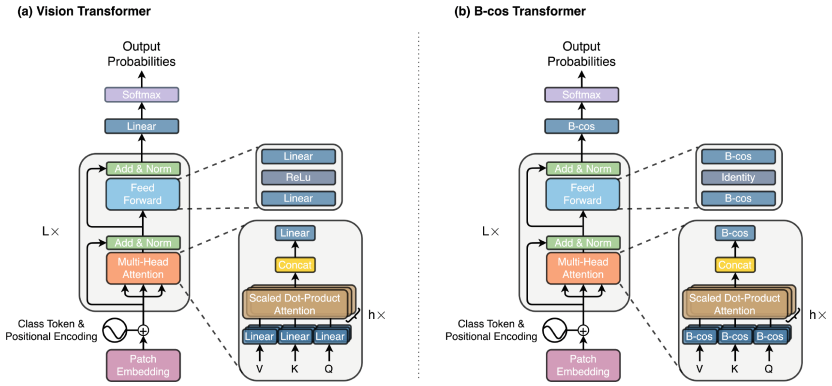


Fig. 2. The model architecture of ViT and BvT (ours). We replace all linear transformations in ViT with the B-cos transform and remove all ReLU activation functions.

- We apply various post-hoc visualization techniques and conduct a blind study with domain experts to assess model interpretability.
- We derive the B-cos Swin Transformer (Bwin) based on the Swin Transformer [21] (Swin) in a generalization study.

2 Related Work

Explainability, interpretability, and relevancy are terms used to describe the ability of machine learning models to provide insight into their decision-making process. Although these terms have subtle differences, they are often used interchangeably in the literature [15].

Recent research on understanding vision models has mostly focused on attribution methods [13,20], which aim to identify important parts of an image and highlight them in a saliency map. Gradient-based approaches like Grad-CAM [30] or attribution propagation strategies such as Deep Taylor Decomposition [27] and LRP [6] are commonly used methods. Perturbation-based techniques, such as SHAP [22], are another way to extract salient features from images. Besides saliency maps, one can also visualize the activations of the model using Activation Maximization [14].

However, it is still controversial whether the above methods can correctly reflect the behavior of the model and accurately explain the learned function (model-faithfulness [17]). For example, it has been shown that some saliency maps are independent of both the data on which the model was trained and the model parameters [2]. In addition, they are often considered unreliable for medical applications [4]. As a result, inherently interpretable models have been proposed as a more reliable and transparent solution. The most recent contribution are B-cos CNNs [7], which use a novel nonlinear transformation (the B-cos transformation) instead of the traditional linear transformation.

Compared to CNNs, there is limited research on understanding transformers beyond attention visualization [10]. Post-hoc methods such as Grad-CAM [30]

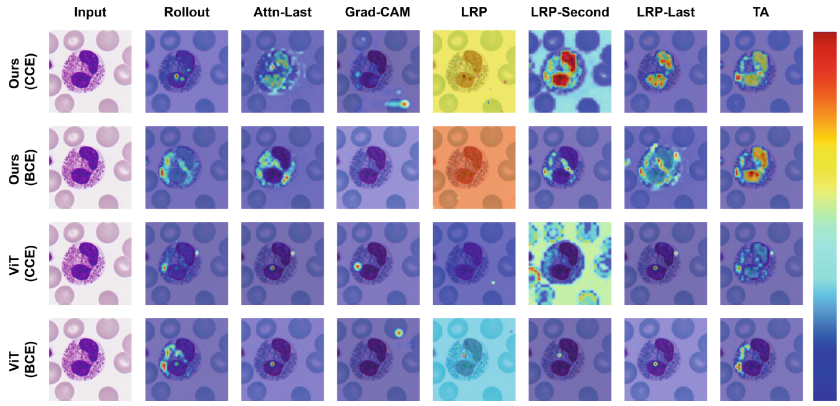


Fig. 3. Rollout, Attention-Last (Attn-Last), Grad-CAM, LRP, LRP of the second layer (LRP-Second), LRP of the last layer (LRP-Last), and Transformer Attribution (TA) applied on the test set of Munich-AML-Morphology. The image shows an eosinophil, which is characterized by its split, but connected nucleus, large specific granules (pink structures in the cytoplasm), and dense chromatin (dark spots inside the nuclei) [29]. Across all visualization techniques, BvT focuses on these exact features unlike ViT.

and Activation Maximization [14] used for CNNs can also be applied to transformers. But in practice, the focus is on visualizing the raw attention values (see Attention-Last [16], Integrated Attention Maps [12], Rollout [1], or Attention Flow [1]). More recent approaches such as Generic Attention [9], Transformer Attribution [10], and Conservative Propagation [3] go a step further and introduce novel visualization techniques that better integrate the attention modules with contributions from different parts of the network. Note that these methods are all post-hoc methods applied after training to visualize the model's reasoning.

On the other hand, the ConceptTransformer [28], achieves better explainability by cross-attending user-defined concept tokens in the classifier head during training. More recently, HIPT [11] combines multi-scale images and DINO [8] pre-training to learn hierarchical visual concepts in a self-supervised fashion. Unlike all of these methods, interpretability is already an integral part of our architecture. Therefore, these methods can be easily applied to our models. In Fig. 3 and Fig. 6, we show that the B-Cos Transformer produces superior feature maps over various post-hoc approaches – suggesting that our architecture does indeed learn human-plausible features that are independent of the specific visualization technique used.

3 Methods

We focus on the original Vision Transformer [12]: The input image is divided into non-overlapping patches, flattened, and projected into a latent space of dimension d. Class tokens $[cls]$ are then prepended to these patch embeddings. In addition, positional encodings $[pos]$ are added to preserve topological information. In the scaled dot-product attention [31], the model learns different features

Table 1. F1-score, top-1, and top-3 accuracy from the test set of NCT-CRC-HE-100K, Munich-AML-Morphology, and TCGA-COAD-20X. We compare ViT and BvT (ours) trained with categorical cross-entropy (CCE) and binary cross-entropy (BCE) loss using two model configurations: T/8 and S/8 (see Sect. 4).

Models	NCT			Munich			TCGA		
	F1	Top-1	Top-3	F1	Top-1	Top-3	F1	Top-1	Top-3
ViT-T/8_{CCE}	**90.9**	92.7	99.1	**57.3**	90.1	98.9	57.1	78.8	94.4
ViT-S/8_{CCE}	89.2	91.1	99.5	56.3	93.1	99.0	56.3	78.6	92.9
BvT-T/8_{CCE}	88.8	91.1	99.3	54.0	87.1	98.6	**61.0**	77.4	93.1
BvT-S/8_{CCE}	88.4	90.1	99.4	52.9	89.8	98.6	60.2	76.3	92.9
ViT-T/8_{BCE}	90.0	91.4	98.4	54.8	90.0	99.0	53.6	79.6	93.9
ViT-S/8_{BCE}	**90.2**	92.2	99.3	**55.4**	92.8	99.0	54.1	77.0	88.9
BvT-T/8_{BCE}	86.7	90.1	98.5	51.1	83.5	97.9	57.7	79.8	93.4
BvT-S/8_{BCE}	87.5	90.4	99.4	52.4	85.0	98.3	**59.0**	74.5	88.9

(query Q, key K, and value V) from the input vectors through a linear transformation. Both query and key are then correlated with a scaled dot-product and normalized with a softmax. These self-attention scores are then used to weight the value by importance:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{1}$$

To extract more information, this process is repeated h times in parallel (multi-headed self-attention). Each self-attention layer is followed by a fully-connected layer consisting of two linear transformations and a ReLU activation.

We propose to replace all linear transforms in the original ViT (Fig. 2)

$$\text{Linear}(x, w) = w^T x = \|w\|\|x\|c(x, w), \tag{2}$$

$$c(x, w) = \cos(\angle(x, w)), \quad \angle\text{...angle between vectors} \tag{3}$$

with the B-cos* transform [7]

$$\text{B-cos*}(x; w) = \underbrace{\|\hat{w}\|}_{=1}\|x\||c(x, \hat{w})|^B \times \text{sgn}(c(x, \hat{w})), \tag{4}$$

where $B \in \mathbb{N}$. Similar to [7], an additional nonlinearity is applied after each B-cos* transform. Specifically, each input is processed by two B-cos* transforms, and the subsequent MaxOut activation passes only the larger output. This ensures that only weight vectors with higher cosine similarity to the inputs are selected, which further increases the alignment pressure during optimization. Thus, the final B-cos transform is given by

$$\text{B-cos}(x; w) = \max_{i \in \{1,2\}} \text{B-cos*}(x; w_i). \tag{5}$$

To see the significance of these changes, we look at Eq. 4 and derive

$$\|\hat{w}\| = 1 \Rightarrow \text{B-cos*}(x; w) \leq \|x\|. \tag{6}$$

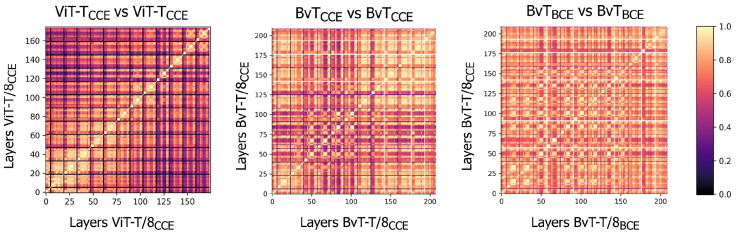
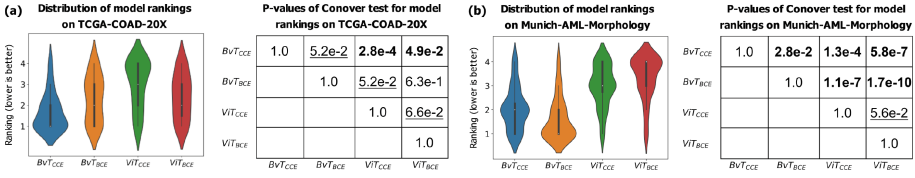


Fig. 4. We compute the central kernel alignment (CKA), which measures the representation similarity between each hidden layer. Since the B-cos transform aligns the weights with the inputs, BvT (ours) achieves a more uniform representation structure compared to ViT (values closer to 1). When trained with the binary cross-entropy loss (BCE) instead of the categorical cross-entropy loss (CCE), the alignment is higher.

Since $|c(x, \hat{w})| \leq 1$, equality is only achieved if x and w are collinear, i.e., if they are aligned. Intuitively, this forces the weight vector to be more similar to the input. Query, key, and value thus capture more patterns in an image – which the attention mechanism can then attend to. This can be shown visually by plotting the centered kernel alignment (CKA). It measures the similarity between layers by comparing their internal representation structure. Compared to ViTs, BvTs achieve a highly uniform representation across all layers (Fig. 4).

4 Implementation and Evaluation Details

Task-Based Evaluation: Cancer classification and segmentation is an important first step for many downstream tasks such as grading or staging. Therefore, we choose this problem as our target. We classify image patches from the public colorectal cancer dataset NCT-CRC-HE-100K [18]. We then apply our method to TCGA-COAD-20X [19], which consists of 38 annotated slides from the TCGA colorectal cancer cohort, to evaluate the effectiveness of transfer learning. This dataset is highly unbalanced and not color normalized compared



	BvT$_{CCE}$	BvT$_{BCE}$	ViT$_{CCE}$	ViT$_{BCE}$
BvT$_{CCE}$	1.0	5.2e-2	**2.8e-4**	**4.9e-2**
BvT$_{BCE}$		1.0	5.2e-2	6.3e-1
ViT$_{CCE}$			1.0	6.6e-2
ViT$_{BCE}$				1.0

	BvT$_{CCE}$	BvT$_{BCE}$	ViT$_{CCE}$	ViT$_{BCE}$
BvT$_{CCE}$	1.0	**2.8e-2**	**1.3e-4**	**5.8e-7**
BvT$_{BCE}$		1.0	**1.1e-7**	**1.7e-10**
ViT$_{CCE}$			1.0	5.6e-2
ViT$_{BCE}$				1.0

Fig. 5. In a blinded study, domain experts ranked models (lower is better) based on whether the models focus on biomedically relevant features that are known in the literature to be important for diagnosis. We then performed the Conover post-hoc test after Friedman with adjusted p-values according to the two-stage Benjamini-Hochberg procedure. BvT ranks above ViT with $p < 0.1$ (underlined) and $p < 0.05$ (bold).

Table 2. Results of Swin and Bwin (ours) experiments on the test set of NCT-CRC-HE-100K and Munich-AML-Morphology. We report F1-score, top-1, and top-3 accuracy.

Models	NCT			AML		
	F1	Top-1	Top-3	F1	Top-1	Top-3
Swin-T_{CCE}	89.1	92.1	99.0	48.2	94.1	98.8
Swin-T_{CCE} (modified)	89.8	92.0	99.6	49.1	94.2	98.9
Bwin-T_{CCE}	91.5	93.5	99.5	53.0	93.9	98.6
Bwin-T_{CCE} (modified)	**92.5**	94.3	99.6	**53.3**	93.8	98.7

to the first dataset. Additionally, we demonstrate that the B-cos Vision Transformer is adaptable to domains beyond histopathology by training the model on the single white blood cell dataset Munich-AML-Morphology [25], which is also highly unbalanced and also publicly available.

Domain-Expert Evaluation: Our primary objective is to develop an extension of the Vision Transformer that is more transparent and trusted by medical professionals. To assess this, we propose a blinded study with four steps: (i) randomly selecting images from the test set of TCGA-COAD-20X (32 samples) and Munich-AML-Morphology (56 samples), (ii) plotting the last-layer attention and transformer attributions for each image, (iii) anonymizing and randomly shuffling the outputs, (iv) submitting them to two domain experts in histology and cytology for evaluation. Most importantly, we show them all the available saliency maps without pre-selecting them to get their unbiased opinion.

Implementation Details: In our experiments, we compare different variants of the B-cos Vision Transformer and the Vision Transformer. Specifically, we implement two versions of ViT: ViT-T/8 and ViT-S/8. They only differ in parameter size (5M for T models and 22M for S models) and use the same patch size of 8. All BvT models (BvT-T/8 and BvT-S/8) are derivatives of the corresponding ViT models. The B-cos transform used in the BvT models has an exponent of $B = 2$. We use AdamW with a cosine learning rate scheduler for optimization and a separate validation set for hyperparameter selection. Following the findings of [7], we add $[1-r, 1-g, 1-b]$ to the RGB channels $[r, g, b]$ of BvT. This allows us to encode each pixel with the direction of the color channel vector, forcing the model to capture more color information. Furthermore, we train models with two different loss functions: the standard categorical cross-entropy loss (CCE) and the binary cross-entropy loss (BCE) with one-hot encoded entries. It was suggested in [7] that BCE is a more appropriate loss for B-cos CNNs. We explore whether this is also true for transformers in our experiments. Additional details on training, optimization, and datasets can be found in the Appendix.

5 Results and Discussion

Task-Based Evaluation: When trained from scratch, all BvT models underperform their ViT counterparts by about 2% on NCT-CRC-HE-100K and

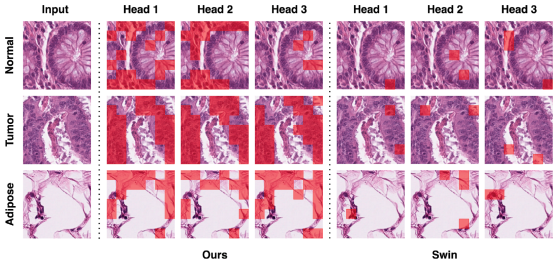


Fig. 6. Attention maps of the last layer of the modified Swin and Bwin (ours). Bwin focuses on cells and nuclei, while Swin mostly focuses on a few spots.

3% on Munich AML-Morphology (Table 1). However, when we use the pre-trained weights from NCT-CRC-HE-100K and transfer them to TCGA-COAD-20X for fine-tuning, BvT outperforms ViT by up to 5% (Table 1). We believe this is due to the simultaneous optimization of two objectives: classification loss and weight-input alignment. With a pre-trained model, BvT is likely to focus more on the former. In addition, we observe that models trained with BCE tend to perform worse than those trained with CCE. However, their saliency maps seem to be more interpretable (see Fig. 3).

Domain-Expert Evaluation: The results show that BvTs are significantly more trustworthy than ViTs ($p < 0.05$). This indicates that BvT consistently attends to biomedically relevant features such as cancer cells, nuclei, cytoplasm, or membrane [24] (Fig. 5). In many visualization techniques, we see that BvT, unlike ViT, focuses exclusively on these structures (Fig. 3). In contrast, ViT attributes high attention to seemingly irrelevant features, such as the edges of the cells. A third expert points out that ViT might overfit certain patterns in this dataset, which could aid the model in improving its performance.

6 Generalization to Other Architectures

We aim to explore whether the B-cos transform can enhance the interpretability of other transformer-based architectures. The Swin Transformer (Swin) [21] is a popular alternative to ViT (e.g., it is currently the SOTA feature extractor for histopathological images [33]). Swin utilizes window attention and feed-forward layers. In this study, we replace all its linear transforms with the B-cos transform, resulting in the B-cos Swin Transformer (Bwin). However, unlike BvT and ViT, it is not obvious how to visualize the window attention. Therefore, we introduce a modified variant here that has a regular ViT/BvT block in the last layer.

In our experiments (Table 2), we observe that Bwin outperforms Swin by up to 2.7% and 4.8% in F1-score on NCT-CRC-HE-100K and Munich-AML-Morphology, respectively. This is consistent with the observations made in Sect. 5: When BvT is trained from scratch, the model faces a trade-off between

learning the weight and input alignment and finding the appropriate inductive bias to solve the classification task. By reintroducing many of the inductive biases of CNNs through the window attention in the case of Swin or transfer learning in the case of BvT, the model likely overcomes this initial problem.

Moreover, we would like to emphasize that the modified models have no negative impact on the model's performance. In fact, all metrics remain similar or even improve. The accumulated attention heads (we keep 50% of the mass) demonstrate that Bwin solely focuses on nuclei and other cellular features (Fig. 6). Conversely, Swin has very sparse attention heads, pointing to a few spots. Consistent with the BvT vs ViT blind study, our pathologists also agree that Bwin is more plausible than Swin ($p < 0.05$).

7 Conclusion

We have introduced the B-cos Vision Transformer (BvT) and the B-cos Swin Transformer (Bwin) as two alternatives to the Vision Transformer (ViT) and the Swin Transformer (Swin) that are more interpretable and explainable. These models use the B-cos transform to enforce similarity between weights and inputs. In a blinded study, domain experts clearly preferred both BvT and Bwin over ViT and Swin. We have also shown that BvT is competitive with ViT in terms of quantitative performance. Moreover, using Bwin or transfer learning for BvT, we can even outperform the original models.

Acknowledgements. M.T. and S.J.W. are supported by the Helmholtz Association under the joint research school "Munich School for Data Science".

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: 58th ACL, pp. 4190–4197. ACL (2020)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency map. In: 32th NeurIPS, pp. 1–11. Curran Associates, Inc. (2018)
3. Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.R., Wolf, L.: XAI for transformers: better explanations through conservative propagation. In: 39th ICML, pp. 435–451. PMLR (2022)
4. Arun, N., et al.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol. Artif. Intell. **3**(6), 1–12 (2021)
5. Bibal, A., et al.: Is attention explanation? An introduction to the debate. In: 60th ACL, pp. 3889–3900. ACL (2022)
6. Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., Samek, W.: Layer-wise relevance propagation for neural networks with local renormalization layers. In: Villa, A.E.P., Masulli, P., Pons Rivero, A.J. (eds.) ICANN 2016. LNCS, vol. 9887, pp. 63–71. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44781-0_8
7. Böhle, M., Fritz, M., Schiele, B.: B-Cos networks: alignment is all we need for interpretability. In: 2022 IEEE/CVF CVPR, pp. 10329–10338. IEEE (2022)
8. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF ICCV, pp. 9650–9660. IEEE (2021)

9. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: 2021 IEEE/CVF CVPR, pp. 397–406. IEEE (2021)
10. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: 2021 IEEE/CVF CVPR, pp. 782–791. IEEE (2021)
11. Chen, R.J., et al.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: 2022 IEEE/CVF CVPR, pp. 16144–16155. IEEE (2022)
12. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: 9th ICLR, pp. 1–21. ICLR (2021)
13. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Commun. ACM **63**(1), 68–77 (2020)
14. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. Technical reports University of Montreal, vol. 1341, no. (3), p. 1 (2009)
15. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th DSAA, pp. 80–89. IEEE (2018)
16. Hollenstein, N., Beinborn, L.: Relative importance in sentence processing. In: 59th ACL and the 11th IJCNLP, pp. 141–150. ACL (2021)
17. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In: Proceedings of the 58th ACL, pp. 4198–4205. ACL (2020)
18. Kather, J.N., et al.: Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. PLOS Med. **16**(1), 1–22 (2019)
19. Kirk, S., et al.: Radiology data from the Cancer Genome Atlas Colon Adenocarcinoma [TCGA-COAD] collection. The cancer imaging archive. Technical report, University of North Carolina, Brigham & Women's Hospital Boston, Roswell Park Cancer Institute (2016)
20. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 1–45 (2020)
21. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF ICCV, pp. 9992–10002. IEEE (2021)
22. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: 31th NeurIPS, pp. 4765–4774. Curran Associates, Inc. (2017)
23. Märkl, B., Füzesi, L., Huss, R., Bauer, S., Schaller, T.: Number of pathologists in Germany: comparison with European countries, USA, and Canada. Virchows Arch. **478**, 335–341 (2021)
24. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T., Marr, C.: Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. Blood **138**(20), 1917–1927 (2021)
25. Matek, C., Schwarz, S., Spiekermann, K., Marr, C.: Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks. Nat. Mach. Intell. **1**(11), 538–544 (2019)
26. Matsoukas, C., Haslum, J.F., Sorkhei, M., Söderberg, M., Smith, K.: What makes transfer learning work for medical images: feature reuse & other factors. In: 2022 IEEE/CVF CVPR, pp. 9225–9234. IEEE (2022)
27. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn. **65**, 211–222 (2017)

28. Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., Scotton, P.: Attention-based interpretability with concept transformers. In: 10th ICLR, pp. 1–16. ICLR (2022)
29. Rosenberg, H.F., Dyer, K.D., Foster, P.S.: Eosinophils: changing perspectives in health and disease. Nat. Rev. Immunol. **13**(1), 9–22 (2013)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE/CVF CVPR, pp. 618–626. IEEE (2017)
31. Vaswani, A., et al.: Attention is all you need. In: 31st NIPS, pp. 6000–6010. Curran Associates, Inc. (2017)
32. Wagner, S.J., et al.: Fully transformer-based biomarker prediction from colorectal cancer histology: a large-scale multicentric study. arXiv preprint arXiv:2301.09617 (2023)
33. Wang, X., et al.: Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. **81**(102559), 1–21 (2022)