Comprehensive Information Bottleneck for Unveiling Universal Attribution to Interpret Vision Transformers

Jung-Ho Hong Ho-Joong Kim Kyu-Sung Jeon Seong-Whan Lee*
Dept. of Artificial Intelligence, Korea University, Seoul, Korea
{jungho-hong, hojoong_kim, ksjeon, sw.lee}@korea.ac.kr

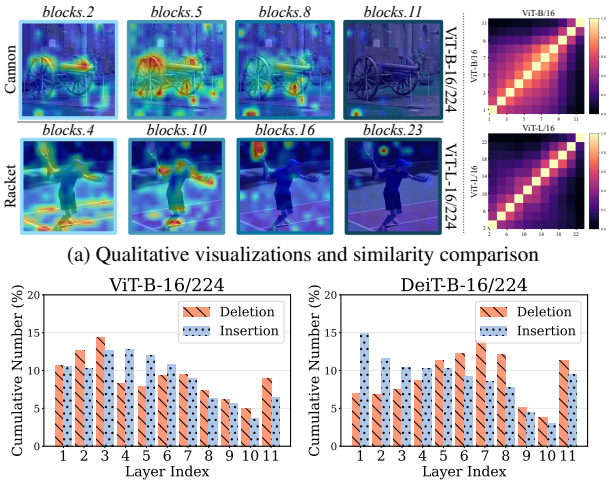
Abstract

The feature attribution method reveals the contribution of input variables to the decision-making process to provide an attribution map for explanation. Existing methods grounded on the information bottleneck principle compute information in a specific layer to obtain attributions, compressing the features by injecting noise via a parametric damping ratio. However, the attribution obtained in a specific layer neglects evidence of the decision-making process distributed across layers. In this paper, we introduce a comprehensive information bottleneck (CoIBA), which discovers the relevant information in each targeted layer to explain the decision-making process. Our core idea is applying information bottleneck in multiple targeted layers to estimate the comprehensive information by sharing a parametric damping ratio across the layers. Leveraging this shared ratio complements the over-compressed information to discover the omitted clues of the decision by sharing the relevant information across the targeted layers. We suggest the variational approach to fairly reflect the relevant information of each layer by upper bounding layer-wise information. Therefore, CoIBA guarantees that the discarded activation is unnecessary in every targeted layer to make a decision. The extensive experimental results demonstrate the enhancement in faithfulness of the feature attributions provided by CoIBA.

1. Introduction

Vision transformer (ViT) achieves remarkable performance in diverse fundamental computer vision tasks, such as multi-modal [24, 36] and self-supervised learning [5, 12]. Despite this achievement, the black box nature stemming from the complex structure of ViT restricts the testability of the end-to-end system, impeding the failure diagnosis [10]. This limitation constrains the application of ViT in safety-critical areas where transparency and interpretability are considered important as well [22, 26]. Feature attribution methods [4, 28, 30] are introduced to reveal the evidence supporting the decision-making process. Many of the existing attribution methods designed to interpret convolutional neural networks show limited explainability to interpret ViT [7]. To address this limitation, feature attribution methods [3, 6–8, 35] tailored for interpreting ViT have been proposed, leveraging either attention weights computed in ViT or intermediate representations, such as the class token, to generate attribution maps.

Although feature attribution methods have advanced in interpreting ViT, they lack theoretical guarantees that low-scored attributions are not necessary to the decision-making process [27]. The methods [27, 37] grounded on the information bottleneck principle [2, 31] address this limitation by guaranteeing the importance of attributions with a



(a) Qualitative visualizations and similarity comparison

(b) Cumulative number of each layer yielding optimal result per sample

Figure 1. **The quantitative and qualitative comparisons of attribution maps generated from different layers.** We illustrate the visualized attribution maps from the first to fourth columns in Fig. 1(a) utilizing IBA [27]. The text on the top of each figure indicates the index of the layer. We select the layer index of the large model with interval 2. The fifth column in Fig. 1(a) shows the similarity comparison between attributions of different layers. Fig. 1(b) illustrates the cumulative number of each layer that provides the best explanation per sample. The uniformly distributed data indicates the absence of the layer optimal for the explanation.

*Corresponding author

variational approximation. Information bottleneck for attribution (IBA) [27] proposes the model-agnostic method to compress the information unnecessary for the prediction by positioning the bottleneck layer into the target layer. InputIBA [37] succeeds in providing a high-resolution attribution map by directly measuring the information in the input domain, assuming the prior distribution with a generative model. However, existing information bottleneck-based methods face two main challenges. First, the resulting attribution map reflects the partial of the decision-making process as the restriction of the information flow is conducted in a specific targeted layer. This constraint provides inconsistent evidence for a single decision-making process, leading to unobvious interpretations. Lastly, no layer dominantly provides the most appropriate relevant information for explaining the decision-making process.

Fig. 1 shows the aforementioned issues with quantitative and qualitative examples. In the first row in Fig. 1(a), the visualization derived from the second layer (*blocks.2*) highlights the *barrel* whereas the fifth (*blocks.5*) and eighth layers (*blocks.8*) highlight the *top* and *bottom of the wheels*, respectively. In the second row in Fig. 1(a), the visualization in the earlier layer highlights the *person holding the racket*, while the deeper layer progressively concentrates on the *racket*. To confirm the layer-wise discrepancy in attributions quantitatively, we measure SSIM [34] in the right row of Fig. 1(a). As the layer providing attributions distance farther, the dissimilarity in those attributions increases. Thus, computing the relevant information from isolated layer-specific information bottlenecks highlights different attributions for the same decision. Among the layers producing the attribution maps, there is no dominant layer that provides the most appropriate relevant information to explain the decision-making process. Fig. 1(b) compares the cumulative number of each layer yielding attributions with the optimal insertion and deletion scores per input sample. The results show a lack of the layer that produces the most appropriate relevant information for explanation. Thus, identifying the most faithful attribution from the layers requires a heuristic search due to the lack of criteria to distinguish the optimal layer.

In this paper, we propose a novel comprehensive information bottleneck for attribution (CoIBA), which reveals the information relevant for each of the target layers to provide the attributions. Our core idea is to reveal comprehensive relevant information by eliminating unnecessary information in every targeted layer of the prediction. We estimate the comprehensive information by sharing a parameterized universal damping ratio across layers, thereby removing the requirements of additional heuristics. This sharing strategy compensates for the over-compressed information of individual layers by sharing information in each layer necessary for the decision. Since a channel element placed in the same location but different layers captures different features, we uniformly perturb all the channel elements in a single token representation, to handle the information of different layers with a shared parametric ratio. We suggest a variational upper bound to restrict information flowing in all the targeted layers. Our approach eliminates the heuristic search for balancing inconsistent information along the layers by suggesting a variational upper bound to fairly reflect the layer-wise relevant information.

We conduct experiments to show the correctness performance of CoIBA utilizing various assessments. As the evaluations are solely approximations, *i.e.* no ground truth exists, we include FunnyBirds [15], which provide the ground truth and assess the comprehensive quality of attribution maps. To assess the feature importance assessment, we leverage insertion/deletion [23] and remove-and-debias (ROAD) [25], which evaluate the faithfulness of attribution maps. In addition, we scrutinize the faithfulness of CoIBA by analyzing the confident-aware assessment. We discuss the effectiveness of our method with a rigorous discussion.

2. Related Works

Explanation Methods for Vision Transformer Existing explanatory methods designed to interpret ViT provide the attribution map utilizing the latent representation and the corresponding gradients. However, these methods show limited explainability while being adapted to ViT [7]. To address this limitation, Rollout [1] linearly combines the attention weights across the layers to compute the attribution map. Trans-attr [7] provides the class-discriminative attribution map by constructing the layer-wise relevance propagation rule for ViT. Generic [6] produces the attribution map with a generalized procedure to interpret diversified transformer models, leveraging the gradient of the attention map to produce the class-discriminative attribution maps. IIA [3] introduces the iterative integration across the input image, leveraging the internal representations processed by the model and their gradients. ViT-CX [35] utilizes the patch embeddings and measures causal impacts to provide attributions. Beyond [8] unfolds attention blocks with the chain rule between final prediction, CLS, and tokens to obtain token contribution. Existing propagation-based approaches require a specific implementation to weigh the contributions of input variables. In contrast to this constraint, CoIBA does not require implementing additional procedures as same as IBA. Furthermore, aligning with the information bottleneck principle, CoIBA guarantees the highlighted attributions are important in all targeted layers for the decision.

Information Bottleneck Approach The information bottleneck principle is broadly utilized to obtain information in the activation necessary for the specific objective. To this end, the information bottleneck-based methods compress the information unnecessary for the objective while

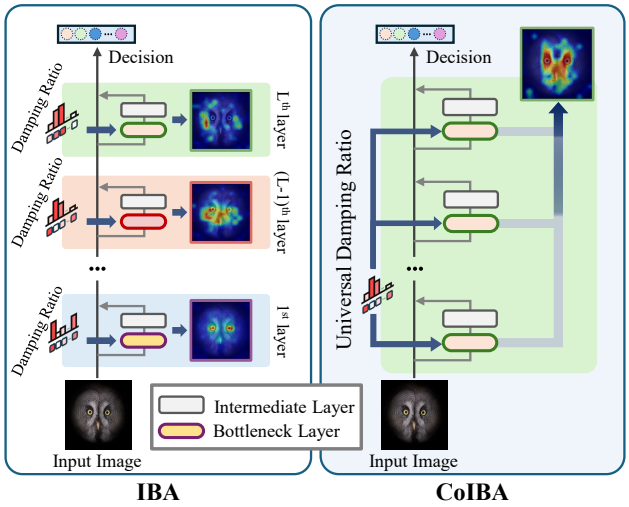


Figure 2. **Main difference between IBA and CoIBA.** IBA obtains the relevant information in the specific bottleneck-inserted layer. Thus, obtaining layer-wise attribution maps necessarily requires iterative running over the layers. In contrast, CoIBA seeks the layer-wise relevant information from each bottleneck-inserted layer with a universal damping ratio.

maintaining the relevant information. Given the challenges of directly estimating the amount of information, a variational approach [18] is used to approximate the intractable posteriors and marginals. The Deep Variational Information Bottleneck [2] method employs the information bottleneck principle in deep networks, using a variational lower bound alongside the reparameterization trick. Referring to this strategy, IBA [27] adapts the information bottleneck principles to restrict the information flow in an intermediate layer and provides the relevant information as an attribution map. InputIBA [37] provides high-resolution attributions by directly estimating relevant information in the input domain. However, existing methods compute relevant information solely considering a specific layer, overlooking the information necessary in the different layers. In contrast to this process, CoIBA obtains comprehensive relevant information from the multiple layers, highlighting the evidence important in every targeted layer to make a decision.

3. Method

The main goal of CoIBA is to produce an attribution map by reflecting the information relevant to the targeted layers, which is opposed to the IBA as shown in Fig. 2. To show this procedure, we first introduce IBA in Sec. 3.1 and discuss the limitations of the existing information bottleneck-based approach in Sec. 3.2. Then, in Sec. 3.3, we describe the overall procedure of CoIBA. Finally, we introduce a variational upper bound, which enhances the reflection of layer-wise relevant information in Sec. 3.4.

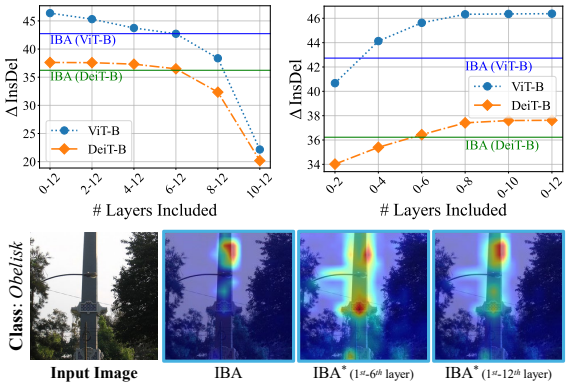


Figure 3. **Quantitative and qualitative results of IBA*.** In the first row, the horizontal line indicates the performance of layer-specific relevant information, *i.e.*, IBA. We report the difference in insertion and deletion scores (ΔInsDel). The higher score is better for this metric. The visualized attribution maps are illustrated in the second row obtained from VIT-B-16/224 [11].

3.1. Background – IBA

IBA adopts an information bottleneck principle to produce relevant information to provide an attribution map. To this end, in l-th layer, IBA computes the bottleneck representation Z_l as follows:

$$Z_l = \lambda_l R_l + (1 - \lambda_l)\epsilon_l \,. \tag{1}$$

Here, the damping ratio λ_l is a learnable parameter assigned at the bottleneck layer, manipulating the degree of damping signal between the activation R_l and the independent noise ϵ_l. The independent noise ϵ_l is sampled from the Gaussian distribution. To maintain the statistic of internal representation, the Gaussian distribution shares the mean μ_{R_l} and variance σ_{R_l} with the representation R_l.

IBA minimizes the shared information between the activation R_l and bottleneck variable Z_l while maximizing the shared information between the bottleneck variable Z_l and the label Y as follows:

$$\max_{\lambda_l} I[Z_l; Y] - \beta I[R_l, Z_l] \,. \tag{2}$$

Here, β is a hyperparameter manipulating the trade-off between compression and relevancy in l-th layer.

3.2. Motivation

IBA selects a specific layer to obtain relevant information to provide an attribution map. However, the relevant information of a specific single layer does not reflect the overall evidence required in the sequence of layers to make a decision. To demonstrate our insistence, we illustrate the quantitative comparison of the attribution maps obtained in a specific layer and multiple layers in Fig. 3, leveraging insertion/deletion [23]. We iterate IBA multiple times along the layers to obtain the set of individual relevant information

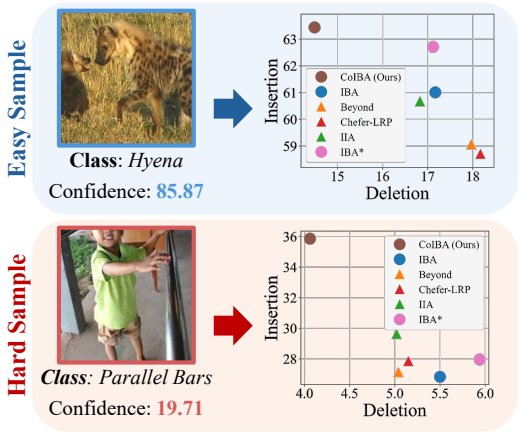


Figure 4. **The quantitative comparisons of explanation methods in high-confident (easy) and low-confident (hard) input samples.** The confidence score is the *softmax* probability of the target class predicted by the model. We compare the information-bottleneck-based methods including IBA [27], IBA*, and CoIBA (marked with a circle), and the propagation-based explanation baselines including Beyond [8], IIA [3], and Chefer-LRP [7] methods (marked with a triangle). We utilize DeiT-B-16/224 [32] for the evaluation. For insertion and deletion, higher and lower scores are better, respectively.

for each layer. After that, we compute a linear combination over the set of individual relevant information. We denote this procedure as IBA*. As shown in the results, the attribution maps produced by IBA* archive better scores compared to the IBA. To scrutinize the quantitative results between IBA* and IBA, we split the insertion/deletion scores by the confidence scores of the model in Fig. 4. As shown in the results, the increase in performance is only observed in high-confident samples, unlike in low-confident samples. Therefore, even iterating IBA, which requires high computational cost, shows limited performance enhancement according to the difficulty of each sample.

3.3. Comprehensive Information Bottleneck

Our comprehensive information bottleneck (CoIBA) inserts the bottleneck layer into the sequence of targeted layers to restrict the information flowing in the intermediate layer. Within the bottleneck layer, CoIBA restricts the information with the learnable universal damping ratio shared across the bottleneck-inserted layers. Our goal is to optimize the universal damping ratio to obtain the layer-wise relevant information while fulfilling the information bottleneck objective. The universal damping ratio enables the attributions to identify distributed relevant information omitted by IBA, which restricts information in a layer-specific manner.

Formulating Information Restriction The model is trained to make a decision $\tilde{Y} = f(X)$ to predict the ground truth Y by taking input X. To make a decision, the input sample X is contextualized by passing through the sequence of intermediate layers. Here, we insert the bottleneck layer into the targeted layer. The role of the bottleneck layer is to restrict the information by dampening the signal of the intermediate representation passing through the L bottleneck-inserted layers. We perturb the intermediate representations in the l-th layer to dampen the signal. To dampen the signal, we inject the independent noise ϵ_l into the intermediate bottleneck representation R'_l, which is computed from the bottleneck variables Z_{l-1} of the preceding layer. Note that the first element of intermediate bottleneck representations is the non-perturbed intermediate representation, *i.e.*, $R'_1 = R_1$. Thus, we obtain the bottleneck representation Z_l as follows:

$$Z_l = \lambda R'_l + (1 - \lambda)\epsilon_l \,, \tag{3}$$

Here, a universal damping ratio $\lambda \in \mathbb{R}^{P\times 1}$ manipulates the degree of perturbation for each token, denoting P as the number of patches. Thus, λ is ranged from 0 to 1, allowing the propagation of the signal when $\lambda = 1$ otherwise blocking it when $\lambda = 0$. As the universal damping ratio λ is consistently adapted to the layers, we omit the layer index in the notation. We compute $\lambda = \text{sigmoid}(\alpha)$ by passing the trainable parameter α, which is initialized with 5, to the sigmoid function. The universal damping ratio uniformly perturbs the signal along the channel dimension. This setting avoids handling the neurons that capture different features with the same coefficient and induces concentration on the token's importance. We empirically show this uniform perturbation across the channels enhances the faithfulness of resulting relevant information. The independent noise ϵ_l shares the dimension and is sampled from the Gaussian distribution with the mean μ_{R_l} and variance $\sigma^2_{R_l}$, such that $\epsilon_l \sim \mathcal{N}(\mu_{R_l}, \sigma^2_{R_l})$. To minimize statistical differences between the bottleneck variable and non-perturbed activations, we leverage the mean μ_{R_l} and variance $\sigma^2_{R_l}$ from those of the non-perturbed activations.

Information Bottleneck Objective The information bottleneck principle desires to obtain information necessary for the objective while compressing irrelevant information from the signal. Compressing irrelevant information is performed by minimizing the shared information between the activation of a bottleneck-inserted network and the bottleneck representation. Building upon this, we compress the amount of information by minimizing the mutual information between internal sequences of bottleneck representations while maximizing the mutual information between the ground truth and bottleneck representations. Combining two aspects with the ground truth Y, we maximize the mutual information between the bottleneck variable and the ground truth while minimizing the mutual information between bottleneck representations as follows:

$$\max_{\lambda} I[Z_L; Y] - \frac{1}{L}\left(\sum\nolimits_{l=1}^{L} \beta_l I[Z_{l-1}; Z_l]\right) . \tag{4}$$

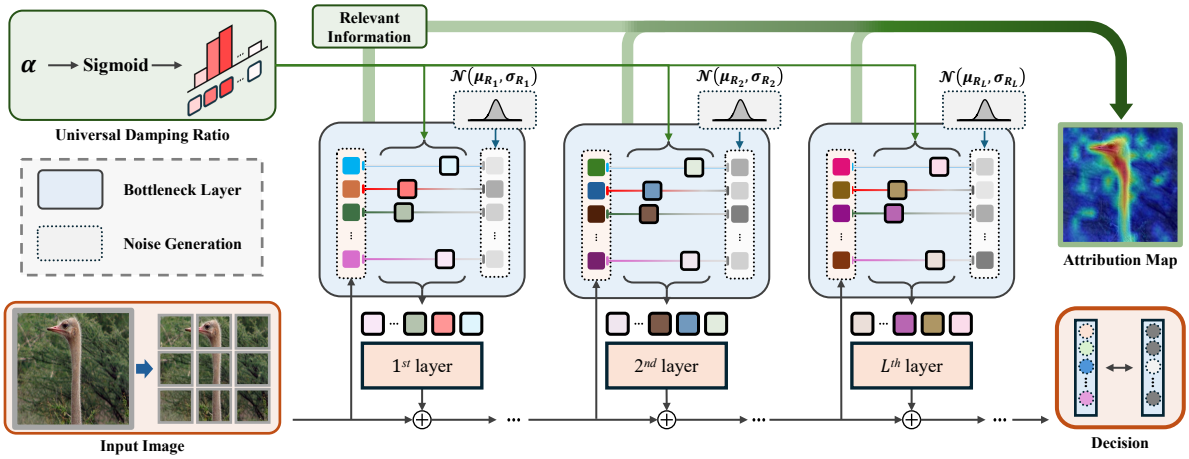


Figure 5. **An overview of CoIBA.** CoIBA restricts the information flow using a universal damping ratio while upper-bounding the information of each layer. After that, CoIBA aggregates the relevant information from each layer to produce an attribution map.

Here, the trade-off parameter β_l governs the degree of compression in each layer. Since the bottleneck layers are inserted into the sequence of layers, the compression term includes the accumulated sequence of mutual information of bottleneck variables $I[Z_{l-1}; Z_l]$. Here, the Z_0 indicates the original intermediate representation as the first element, such that $Z_0 = R_1$. We maximize the relevancy term $I[Z_L; Y]$ while compression to maintain the shared information between bottleneck variables and the ground truth. We maximize this term by minimizing the cross-entropy loss $\mathcal{L}_{ce}$.

Utilizing KL divergence, the mutual information between internal representations $I[Z_{l-1}; Z_l]$ is formulated by:

$$I[Z_{l-1}; Z_l] = \mathbb{E}_{Z_{l-1}}[D_{KL}[P(Z_l|Z_{l-1})||P(Z_l)]] \quad (5)$$

However, the direct estimation of the prior distribution $P(Z_l)$ is intractable as it requires the integration of bottleneck representations $P(Z_l) = \int_{Z_{l-1}} P(Z_l|Z_{l-1})P(Z_{l-1})dZ_{l-1}$ of the corresponding l-th layer. Thus, we assume the prior distribution $Q(Z_l)$ as Gaussian distribution $\mathcal{N}(\mu_{R_l}, \sigma^2_{R_l})$, reconstituting the mutual information computation as follows:

$$I[Z_l; Z_{l-1}] \leq \mathbb{E}_{Z_{l-1}}[D_{KL}[P(Z_l|Z_{l-1})||Q(Z_l)]]. \quad (6)$$

Since assuming the prior distribution with Gaussian distribution only overestimates the mutual information, minimizing this term suppresses the mutual information of the compression term.

3.4. Variational Upper Bound

As discussed in Sec. 4.6, the relevant information obtained by suppressing layer-wise mutual information suffers from reflecting layer-wise relevant information due to the inconsistent amount of information along the layers. Thus, obtaining the sequence of trade-off parameters $\{\beta_l\}_{l=1}^L$ requires an extensive heuristic search to balance the degree of compression for each layer. To overcome this issue, we suggest the variational upper bound for CoIBA, referring to the two following points. First, since there is no additional information propagated during the forwarding pass, the mutual information of intermediate layers would not be greater than the mutual information between the model input and output. Second, the iterated noise-injection procedure diminishes the mutual information among the internal and bottleneck representations, *i.e.*, $I[Z_l; Z_{l+1}] \leq I[Z_{l-1}; Z_l]$. Building upon these bases, we establish an upper bound that encompasses the combined mutual information from all subsequent layers as follows:

$$I[R_1; Z_1] \geq \frac{1}{L}\left(\sum\nolimits_{l=1}^{L} I[Z_{l-1}; Z_l]\right) . \quad (7)$$

Accordingly, we reconstitute the objective to be a simplified formula, leaving only a single hyperparameter β as follows:

$$\max_{\lambda} I[Z_L; Y] - \beta I[R_1; Z_1] . \quad (8)$$

To compress the subsequent layers, the simplified objective necessitates calculating only the mutual information of the first layer $I[R_1; Z_1]$. This term solely utilizes the non-perturbed intermediate and bottleneck representations of the first layer, simplifying the objective calculation. In terms of relevant information computation, referring to the data processing inequality, the inequality $I[Y; Z_l] \leq I[Y; Z_{l-1}]$ holds. Thereby, in contrast to the layer-specific information bottleneck, which overestimates mutual information in the earlier layers during compression [37], CoIBA relieves the overestimation. This is because, as discussed in Sec. 4.6, the relevant information is compensated as the sequence of bottleneck variables joins the objective computation.

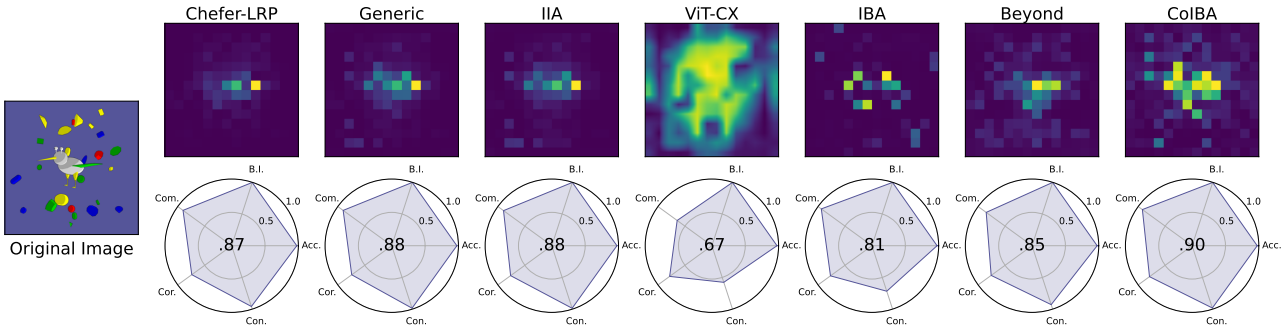


Figure 6. **Quantitative comprehensive assessment on FunnyBirds experiment:** The abbreviations Com. Cor. and Con. indicate *Completeness*, *correctness*, and *contrastivity*. The center score indicates the comprehensive result of enumerated aspects. Acc. and B.I. indicate accuracy and background independence, respectively. We provide detailed numeric in the supplementary material.

Variant	Model	Chefer-LRP [7]	Generic [6]	IIA [3]	ViT-CX [35]	IBA [27]	Beyond [8]	CoIBA
ViT	ViT*-S-16/224	-	14.56/56.08	13.69/57.31	14.98/56.65	13.56/58.64	13.50/58.47	**11.08/63.35**
	ViT*-T-16/224	-	9.00/48.95	8.50/49.75	12.08/45.09	8.02/49.22	11.26/44.07	**6.97/54.23**
	ViT-B-16/224	17.67/56.69	17.92/57.03	16.56/58.7	16.94/58.17	17.23/59.06	15.84/60.37	**13.01/62.58**
	ViT-L-16/224	20.97/53.81	20.17/55.54	20.14/55.22	19.76/57.38	20.88/54.85	18.99/56.55	**15.40/61.14**
DeiT	DeiT-S-16/224	12.76/47.51	12.10/48.64	10.89/49.76	17.57/44.43	11.43/47.87	11.13/48.88	**9.55/53.38**
	DeiT-B-16/224	14.63/48.79	15.10/49.41	13.69/50.51	18.21/46.97	13.95/50.18	14.45/48.88	**11.79/53.96**
	DeiT3-B-16/224	-	15.70/52.42	15.38/52.70	18.88/50.40	15.07/53.32	14.71/54.07	**12.97/56.54**
	DeiT3-L-16/224	-	22.10/62.73	17.75/55.93	26.11/59.60	21.01/65.60	19.34/64.27	**17.64/67.20**
Swin	Swin-B	-	33.76/42.94	-	-	18.03/52.81	23.73/50.39	**17.09/54.80**
	Swin2-B	-	34.71/47.16	-	-	20.94/53.85	31.32/49.44	**18.88/55.77**

Table 1. **Quantitative feature importance assessment on insertion ↑ / deletion ↓.** $(^*)$ denotes the ViT trained with strong regularization techniques [29]. We underline the state-of-the-art performance among the baselines.

4. Experiments

4.1. Setup

Architectures We select variants of ViT for our experiment, including original ViT [11] and DeiT models [32, 33], trained with ImageNet-21k (IN-21k) and ImageNet-1k (IN-1k), respectively. We include the variants in depth interpolated from ViT-T, ViT-S, ViT-B, ViT-L, and ViT-H models. For the ViT-T and ViT-S models, we utilize the model pre-trained with massive regularization techniques [29]. The ViT-H model is pre-trained with CLIP [24]. We present the settings of the ViT-B model with patch size 16 and image resolution 224 as ViT-B-16/224. We include Swin transformers [20, 21], demonstrating the generalized ability of CoIBA in multi-scale features processing. The details of settings and further results about convolutional neural networks are included in the supplementary material.

Feature Attribution Methods We compare the various feature attribution methods including Chefer-LRP [7], Generic [6], IIA [3], ViT-CX [35], Beyond [8], and IBA [27] as baselines. We select $\beta = 10$ for IBA and 6-th layer to insert bottleneck. For CoIBA, we select trade-off hyper-parameter β as 1. We insert the bottleneck layer from s-th departure to e-th arrival layers, which are chosen, 4, and 12, respectively, for ViT-B. For the optimization, we set the learning rate, batch size, and optimizer as 1, 10, and Adam [17]. The additional hyperparameter settings are included in the supplementary material. For Swin transformers, we only report the explanation methods that do not require methodology modification, including Generic, ViT-CX, IBA, Beyond, and CoIBA. We utilize RTX A6000 GPU for all the experiments.

Datasets We utilize ImageNet-1k [9] (IN-1k), ImageNet-A (IN-A) [14] and ImageNet-R (IN-R) [13] validation datasets for the experiments. We leverage IN-A and IN-R to conduct the difficult-aware analysis.

4.2. FunnyBirds Assessment

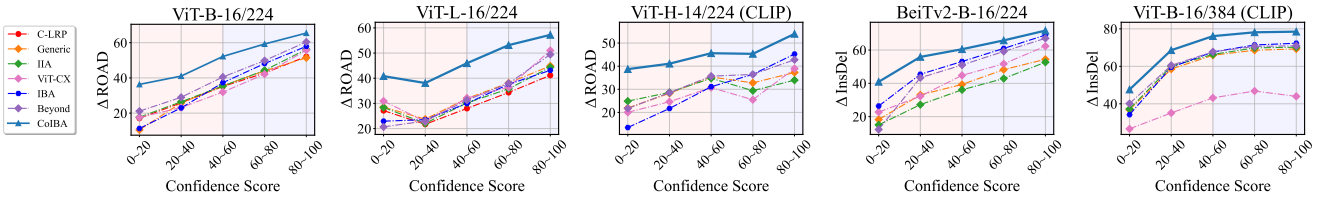
The FunnyBirds [15] experiment provides the ground truth for the evaluation. With the help of ground truth, the Funnybird framework assesses the quality of the attribution map from three perspectives: *completeness*, *correctness*, and *contrastivity*. Fig. 6 illustrates the qualitative and quantitative results of the Funnybirds experiment. As shown in the results, CoIBA outperforms all the baselines, including the propagation or gradient-based approaches. Furthermore, in contrast to IBA, CoIBA obtains an attribution map with significantly enhanced *contrastivity*.

4.3. Insertion/Deletion

Insertion/deletion [23] measures the correctness of feature importance highlighted by an attribution map. The insertion/deletion method gradually inserts or deletes informative pixels in ascending order of attribution to compute the area under the curve scores. In insertion/deletion, methods

Model	Chefer-LRP [7]	Generic [6]	IIA [3]	ViT-CX [35]	IBA [27]	Beyond [8]	CoIBA
ViT-B-16/224	20.87/64.37	22.15/65.31	19.58/66.19	19.04/65.36	20.99/68.80	17.93/68.44	**16.63/73.68**
DeiT-B-16/224	15.31/59.55	16.76/59.78	14.74/61.21	18.37/57.44	15.62/61.53	15.37/60.06	**11.59/67.12**
ViT-L-16/224	29.92/64.87	28.31/66.51	28.39/65.92	24.30/67.27	30.55/66.97	25.47/66.68	**22.03/74.88**
DeiT3-L-16/224	-	20.28/69.87	19.72/70.02	24.43/66.60	19.56/71.23	19.41/70.52	**15.02/76.13**
ViT†-H-16/224	-	30.06/61.34	32.27/60.98	30.34/57.70	28.30/62.29	28.11/62.96	**23.38/66.95**

Table 2. **Quantitative feature importance evaluation of ROAD (MoRF ↓ / LeRF ↑).** The higher and lower scores indicate better attribution map quality for MoRF and LeRF, respectively. The underlined scores indicate the highest performance among the baselines. (†) denotes the ViT trained with CLIP [24].



(a) IN-1k (b) IN-A (c) IN-R

Figure 7. **Quantitative difficulty-aware correctness assessment on insertion/deletion and ROAD.** We measure differences in insertion/deletion (ΔInsDel↑) and ROAD (ΔROAD↑) scores. We fill the regions including low-confident samples and high-confident samples with red and blue, respectively, based on the prediction made by the model. The first to third columns illustrate the quantitative results of the IN-1k. The fourth and fifth columns include the results of IN-A and IN-R datasets, respectively.

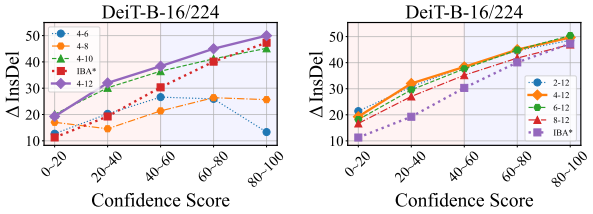
providing better correctness performance achieve higher scores for insertion and lower scores for deletion. We gradually deleted or inserted around every 3.5% pixels of the input image and randomly sampled 6,000 samples from the IN-1k validation dataset for the assessment. As shown in Table 1, the attributions provided by CoIBA yield predominant correctness performance compared to baselines. Especially, the consistent increase in correctness from the Swin transformer demonstrates the generalizability of CoIBA.

4.4. Remove and Debias

The remove-and-debias (ROAD) [25] experiments measure the correctness of the attribution maps, addressing the class information leakage problem (*i.e.*, the shape of the mask) observed in the remove-and-retrain (ROAR) [16] experiment. We average the quantitative scores over the 6,000 randomly selected images when 20, 40, 60, and 80% of pixels are imputed in two perspectives: most relevant first (MoRF) and least relevant first (LeRF). As shown in Table 2, CoIBA outperforms existing methods in MoRF and LeRF experiments with a sizable gap. These results demonstrate that regardless of the imputation type, CoIBA outperforms all the baselines regardless of the model setting (patch size and depth) and pre-trained dataset.

4.5. Difficulty-aware Correctness Assessment

We compare the quantitative quality of attribution maps yielded by CoIBA and baseline methods in terms of difficulty-aware assessment, in addition to correctness assessments of overall samples. Leveraging the confidence scores, and the prediction probability of a target class, we divide the confidence scores with 20 intervals as in Sec. 3.2. As shown in Fig. 7, CoIBA consistently outperforms all the baselines regardless of the confidence scores predicted with each sample, predicted by the model. Along with IN-1k, we include the quantitative results of IN-A and IN-R. As both IN-A and IN-R datasets include the low-confident (difficult) samples compared to IN-1k, these quantitative results support the demonstration that CoIBA outperforms the correct performance with low-confident samples.

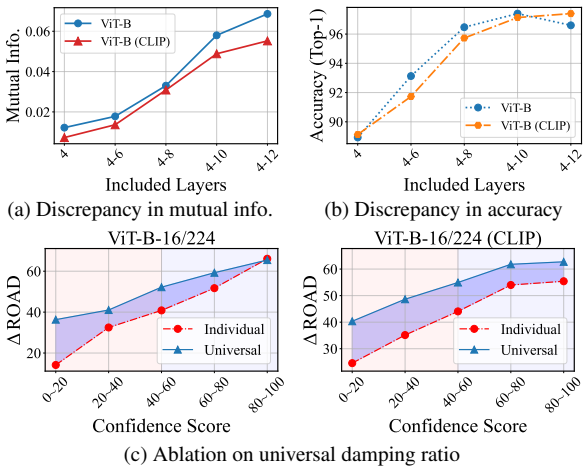


(a) Ablation on departure layer (b) Ablation on arrival layer

Figure 8. **Comparisons of different departure s and arrival e layer settings.** We plot the hyper-parameters selected for CoIBA as a solid line. We compare the quantitative results of the discrepancy between AUC scores of insertion/deletion (ΔInsDel) with different intervals of confidence scores outputted by the model. The better attribution method achieves a higher score. We provide further results in the supplementary material.

4.6. Discussion

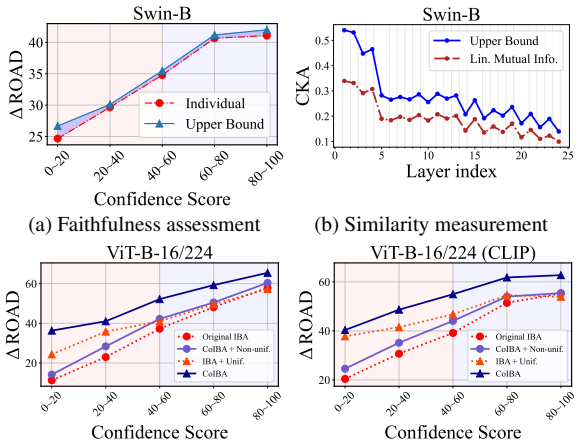
Ablation on Bottleneck-inserted Layers In this section, we confirm whether including multiple layers enhances the correctness of the resulting attribution map. Fig. 9 shows the comparison of the quantitative results by interpolating departure (s) and arrival (e) layers from earlier to deeper layers. The results demonstrate that increasing the number of layers to compute relevant information consistently enhances the correctness quality of an attribution map. Furthermore, CoIBA consistently outperforms IBA* with large margins, requiring only a single iteration. Thus, in CoIBA, various layers gladly reveal the omitted relevant information obtained from a specific layer.



(a) Discrepancy in mutual info. (b) Discrepancy in accuracy

(c) Ablation on universal damping ratio

Figure 9. **Ablation study on layers to analyze universal damping ratio.** Fig. 9(a) illustrates the change in mutual information for compression terms $I[Z_1; R_1]$ in Eq. (8) and the Fig. 9(b) shows the change in accuracy. We utilize the ViT-B-16/224 model for these experiments. The mutual info. denotes mutual information.



(a) Faithfulness assessment (b) Similarity measurement

(c) Ablation on channel uniform perturbation

Figure 10. **Quantitative analysis on CoIBA.** We report the results divided by the confidence score computed by the model. For the comparison, we include the ViT-B model pre-trained with CLIP in Fig. 10(c). We filled the discrepancy in performances with blue color in Fig. 10(c). The solid line indicates the results of CoIBA.

Effectiveness of Universal Damping Ratio We investigate whether leveraging the universal damping ratio compensates for over-compression in Fig. 9(a) while amplifying the relevant information in Fig. 9(b). We compare the different numbers of targeted layers to compare the quantitative results. The results illustrated in Fig. 9(a) show that leveraging the universal damping ratio complements the over-compressed information that occurred in earlier layers, due to the delivered relevant information from deeper layers. Concurrently, as shown in Fig. 9(c), leveraging the universal damping ratio amplifies the relevant information term in Eq (8). Aligning with these results, compared by assigning individual damping ratios to each layer, our universal damping ratio enhances the correctness performance of CoIBA. Therefore, leveraging the universal damping ratio to targeted multiple layers significantly enhances the correctness performance of CoIBA by compensating for the over-compression and amplifying the relevant information.

Effectiveness of Variational Upper Bound We demonstrate the effectiveness of our variational upper bound by measuring correctness performance and whether the relevancy of different layers is fairly reflected in attributions. We compare quantitative results obtained by leveraging a variational upper bound (Eq. (8)) versus linearly combining layer-wise mutual information (Eq. (4)) in Fig. 10(a). As shown in the results, leveraging our variational upper bound is better at correctly identifying feature importance, regardless of confidence scores. In addition to this result, we confirm whether the relevant information of each layer is fairly reflected in the attribution map by comparing similarities. Concretely, we compare the similarity of both settings (Eq. (4) and (8)) with layer-specific relevant information computed by iterating IBA with channel uniform perturbation, denoted as IBA*. We employ CKA [19] as a similarity metric. As shown in Fig. 10(b), leveraging our upper bound yields the comprehensive relevant information of all the layers. This result demonstrates that our variational upper bound encourages the resulting comprehensive relevant information to fairly reflect the relevancy of each layer.

Effectiveness of Uniform Channel Perturbation We demonstrate the effectiveness of uniform channel perturbation by comparing the correctness performance of uniform (unif.) and non-uniform (non-unif.) channel perturbation strategies. As shown in Fig. 10(c), the uniform channel perturbation consistently enhances the correctness performance of IBA for low-confident samples. In addition to this result, uniform channel perturbation significantly enlarges the correctness of resulting attribution maps, regardless of the confidence scores. These enhanced correctness performances demonstrate the effectiveness of our uniform channel perturbation, regardless of the method.

5. Conclusion

In this paper, we introduce CoIBA, which reveals the comprehensive relevant information to produce an attribution map, revealing the omitted relevancy in IBA with a theoretical guarantee. CoIBA shares the universal damping ratio to compensate for the over-compressed information, delivering relevancy among the bottleneck-inserted layers. CoIBA clearly judges the importance of each token by leveraging uniform channel perturbation. We leverage variational approximation to upper bound the information to ensure the eliminated activations are not necessary for the bottleneck-inserted layers to make a decision. CoIBA demonstrates a substantial improvement over existing methods in numerous experiments and discussions.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant, funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), IITP-2025-RS-2024-00436857, ITRC (Information Technology Research Center), and No. RS-2022-II220984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation).

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020. 2

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. 1, 3

[3] Oren Barkan, Yuval Asher, Amit Eshel, Noam Koenigstein, et al. Visual explanations via iterated integrated attributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2073–2084, 2023. 1, 2, 4, 6, 7

[4] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 63–71. Springer, 2016. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 1, 2, 6, 7

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 1, 2, 4, 6, 7

[8] Jiamin Chen, Xuhong Li, Lei Yu, Dejing Dou, and Haoyi Xiong. Beyond intuition: Rethinking token attributions inside transformers. *Transactions on Machine Learning Research*, 2022. 1, 2, 4, 6, 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 6

[10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 1

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3, 6

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 6

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6

[15] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3981–3991, 2023. 2, 6

[16] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[19] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 8

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6

[21] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 6

[22] Clemens Otte. Safe and interpretable machine learning: a methodological review. *Computational Intelligence in Intelligent Data Analysis*, pages 111–122, 2013. 1

[23] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *British Machine Vision Conference*, 2018. 2, 3, 6

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 6, 7

[25] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, pages 18770–18795. PMLR, 2022. 2, 7

[26] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22, 2019. 1

[27] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4, 6, 7

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 618–626, 2017. 1

[29] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2024. 6

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 1

[31] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 1

[32] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 4, 6

[33] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *Proceedings of the European Conference on Computer Vision*, pages 516–533. Springer, 2022. 6

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2

[35] Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L Zhang. Vit-cx: causal explanation of vision transformers. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1569–1577, 2023. 1, 2, 6, 7

[36] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1

[37] Yang Zhang, Ashkan Khakzar, Yawei Li, Azade Farshad, Seong Tae Kim, and Nassir Navab. Fine-grained neural network explanation by identifying input features with predictive information. *Advances in Neural Information Processing Systems*, 34:20040–20051, 2021. 1, 2, 3, 5