Explanation Bottleneck Models

Shin'ya Yamaguchi[1,2*] **and Kosuke Nishida**[1]

[1]NTT
[2]Kyoto University

Abstract

Recent concept-based interpretable models have succeeded in providing meaningful explanations by pre-defined concept sets. However, the dependency on the pre-defined concepts restricts the application because of the limited number of concepts for explanations. This paper proposes a novel interpretable deep neural network called *explanation bottleneck models* (XBMs). XBMs generate a text explanation from the input without pre-defined concepts and then predict a final task prediction based on the generated explanation by leveraging pre-trained vision-language encoder-decoder models. To achieve both the target task performance and the explanation quality, we train XBMs through the target task loss with the regularization penalizing the explanation decoder via the distillation from the frozen pre-trained decoder. Our experiments, including a comparison to state-of-the-art concept bottleneck models, confirm that XBMs provide accurate and fluent natural language explanations without pre-defined concept sets.

Code — https://github.com/yshinya6/xbm/
Extended version — https://arxiv.org/abs/2409.17663

1 Introduction

Although deep learning models can achieve remarkable performance on many applications, they are black-box, i.e., their output predictions are not interpretable for humans. Introducing concept bottleneck models (CBMs, Koh et al. (2020)) is a promising approach to interpreting the output of deep models. In contrast to black-box models that directly predict output labels from input in an end-to-end fashion, CBMs first predict *concept* labels from input and then predict final target class labels from the predicted concepts. Since the predicted concepts represent semantic input ingredients, this two-staged prediction enables users to know the reasons for the final target label predictions and interactively intervene in the decision-making process for critical applications such as healthcare (Chauhan et al. 2023).

However, the existing CBMs depend on the fixed pre-defined concept sets to predict final labels. In other words, they can not provide interpretability to any other than the pre-defined concepts. We argue that this limitation presents a fundamental challenge for CBMs in achieving interpretable deep models. Although recent CBM variants leveraging pre-trained large language models (Yuksekgonul, Wang, and Zou 2023; Oikarinen et al. 2023) enable to express concepts of arbitrary target classes, the interpretability is still restricted to a fixed and small number of concepts. This is because a large number of concept labels are difficult to learn due to their long-tail distribution and are less interpretable by the limitation of human perception (Ramaswamy et al. 2023). In fact, the prior works restrict the number of concepts by filtering with the similarity between concepts and training images to maintain the performance and interpretability (Oikarinen et al. 2023; Yang et al. 2023). Therefore, as long as they depend on pre-defined concepts, CBMs are restricted in the number of interpretable concepts and are insufficient to explain the output of deep models.

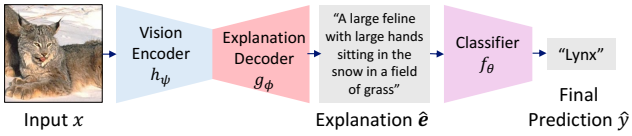

Figure 1: Explanation bottleneck models (XBMs). We propose an interpretable model that generates text explanations for the input embedding with respect to target tasks and then predicts final task labels from the explanations.

This paper tackles a research problem where we do not assume pre-defined concept sets for constructing interpretable deep neural networks. To this end, we propose a novel family of interpretable models called *explanation bottleneck models* (XBMs), which leverage pre-trained multi-modal encoder-decoder models that can generate text descriptions from input data (e.g., BLIP (Li et al. 2022, 2023)). Leveraging pre-trained multi-modal encoder-decoder enables capturing concepts that actually appeared in the input beyond pre-defined concept sets. Our key idea is to decode concepts as text explanations from input and then predict the final label with a classifier that takes the decoded explanations (Fig. 1). In contrast to CBMs, which make predictions based on pre-defined concepts, XBMs make predictions based on concepts actually appeared in the input data through the decoded explanations and can provide an intuitive interpretation of the final prediction tied to the input. Through end-to-end training, XBMs

*shinya.yamaguchi@ntt.com

aim to generate explanations focusing on the textual features for solving the target task.

A major challenge for XBMs is forgetting the text generation capability during training on target tasks. Since target datasets usually lack ground-truth text labels, it is challenging to avoid catastrophic forgetting. To generate high-quality explanations, we introduce a training technique called *explanation distillation*, which penalizes the text decoders by the reference explanations generated by frozen pre-trained text decoders. Solving target tasks with explanation distillation enables XBMs to decode explanations from input data in natural sentences without corruption.

We conduct experiments to evaluate XBMs on multiple datasets by comparing them to existing CBMs and black-box baselines regarding interpretability and target task performance. Our experiments show that XBMs can provide a more relevant explanation to input than the pre-defined concepts of existing CBMs while achieving competitive performance to black-box baselines and largely outperforming CBMs in target test accuracy. We also show that training XBMs can enhance the multi-modal understanding capability of backbone vision-language models by focusing on the target-related vocabulary. Further, we confirm the reliability and practicality of the XBMs' explanations through the experiments intervening with the random texts and the ground-truth explanations.

2 Explanation Bottleneck Models

This section introduces the principle of explanation bottleneck models (XBMs). XBMs are interpretable deep learning models that predict a final label from the generated explanation text from XBMs themselves. Since the predicted final labels are based on the generated explanation of input images, we can naturally interpret the explanation as the reason for the prediction of XBMs. Figure 2 illustrates the overview of training an XBM. An XBM consists of a visual encoder h_ψ, an explanation decoder g_ϕ, and a classifier f_θ for predicting final target labels. Among them, h_ψ and g_ϕ are initialized by an arbitrary pre-trained multi-modal encoder-decoder like BLIP (Li et al. 2022). f_θ is a multi-modal classifier built on a transformer that takes the generated explanations as input and conditions the cross-attention layers with image embeddings; this design is inspired by hybrid post-hoc CBMs (Yuksekgonul, Wang, and Zou 2023) that uses input embeddings to complement missing concepts not in the predicted concepts. We also confirm the practicality when using a text classifier in Section 3.4. In this section, we mainly describe XBMs with a multi-modal classifier. XBMs are trained by the target classification loss in an end-to-end manner. Since naïve training leads to collapse in generated text explanation, we avoid the collapse by *explanation distillation*. Explanation distillation penalizes the explanation decoder with a reference text generated from a frozen pre-trained text decoder g_{ϕ_p} to prevent the decoders from forgetting the text generation capability.

2.1 Problem Setting

We consider a K-class image classification task as the target task. We train neural network models $h_\psi : \mathcal{X} \to \mathbb{R}^{d_\mathcal{X}}$, $g_\phi : \mathbb{R}^{d_\mathcal{X}} \to \mathcal{E}$, and $f_\theta : (\mathbb{R}^{d_\mathcal{X}}, \mathcal{E}) \to \mathcal{Y}$ on a labeled target dataset $\mathcal{D} = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^N$, where $\mathcal{X}$, $\mathcal{E}$, and $\mathcal{Y}$ are the input, text explanation, and output label spaces, respectively. The text explanation space consists of token sequences of the length L with token vocabulary $\mathcal{V}$, i.e., $\mathcal{E} = \mathcal{V}^L$. h_ψ is a vision encoder, which embeds an input x into $d_\mathcal{X}$ dimensional space, g_ϕ is an auto-regressive text decoder that generates a text explanation $\boldsymbol{e} \in \mathcal{E}$ from an input embedding $h_\psi(x)$, and f_θ is a classifier that predicts a final target task label y. We assume that h_ψ and g_ϕ are initialized by pre-trained multi-modal model's parameters ψ_p and ϕ_p, which are pre-trained on large-scale text-image paired datasets with an existing method such as BLIP (Li et al. 2022) and LLaVA (Liu et al. 2023). Note that we do not assume ground truth text explanation set $\{\boldsymbol{e}^i\}_{i=1}^N$ in $\mathcal{D}$ for training g_ϕ.

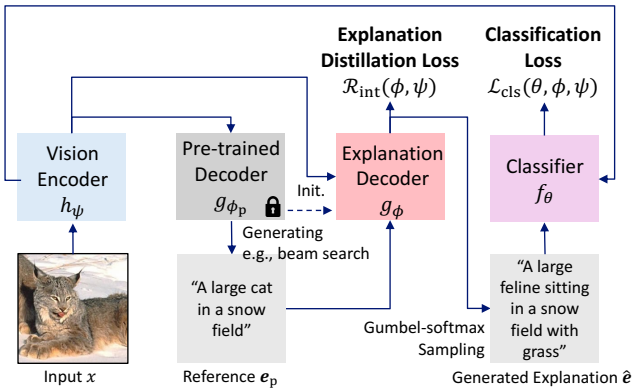

Figure 2: Training of XBMs. An XBM is optimized by the target task loss with explanation distillation. Explanation distillation leverages a reference explanation $\boldsymbol{e}_\mathrm{p}$ generated from a pre-trained text decoder g_{ϕ_p} for penalizing the output distribution of an explanation decoder g_ϕ to maintain the interpretable text generation capability of g_ϕ.

This setting is similar to that of concept bottleneck models (CBMs, Koh et al. (2020)), where a model predicts a final label y from a set of concepts $\{c^j \in \mathcal{C}\}_{j=1}^M$ decoded from input x instead of using $\boldsymbol{e}$. The major difference is in the assumption of pre-defined concept sets: our setting does not explicitly specify the words and phrases for the explanations, whereas CBMs explain the model's output based on the words and phrases in a pre-defined concept set $\{c^j\}$.

2.2 Objective Function

XBMs aim to achieve high target classification accuracy while providing interpretable explanations of the predictions. To this end, XBMs solve an optimization problem with a regularization term defined by the following objective function.

$$\min_{\theta,\phi,\psi} \mathcal{L}_{\mathrm{cls}}(\theta, \phi, \psi) + \lambda \mathcal{R}_{\mathrm{int}}(\phi, \psi), \quad (1)$$

$$\mathcal{L}_{\mathrm{cls}}(\theta, \phi, \psi) = \mathbb{E}_{(x,y)\in\mathcal{D}}\, \ell_{\mathrm{CE}}(f_\theta \circ g_\phi \circ h_\psi(x), y), \quad (2)$$

where $\mathcal{R}_{\mathrm{int}}(\cdot)$ is a regularization term that guarantees the fluency of the explanations generated from g_ϕ, λ is a hyperparameter for balancing $\mathcal{L}_{\mathrm{cls}}$ and $\mathcal{R}_{\mathrm{int}}$, and ℓ_{CE} is cross-entropy loss. Through this objective, the text decoder g_ϕ is trained to focus on the textual features that are useful

for minimizing $\mathcal{L}_{\text{cls}}$ while keeping the interpretability by $\mathcal{R}_{\text{int}}$. We found that g_ϕ easily collapses their output without $\mathcal{R}_{\text{int}}$. Thus, the design of $\mathcal{R}_{\text{int}}$ is crucial for training XBMs. However, since we often do not have the ground truth explanation sets in a real-world target dataset $\mathcal{D}$, we can not directly penalize g_ϕ with supervised losses as $\mathcal{R}_{\text{int}}$. To overcome this challenge, we introduce a distillation-based approach using pre-trained text decoders in the next section.

2.3 Explanation Distillation

XBMs utilize pre-trained multi-modal models as the initial parameters of the text (explanation) decoder g_ϕ. As an auto-regressive sequence model, the pre-trained text decoder $g_{\phi_{\text{p}}}$ can learn a conditional distribution $q(\boldsymbol{e}|x)$ as

$$q(\boldsymbol{e}|x) = \prod_{l=1}^{L} q(e_l|x, \boldsymbol{e}_{<l}), \tag{3}$$

where L is the maximum token length, e_l is the l-th token, and $\boldsymbol{e}_{<l}$ is the text sequence before e_l. Since $g_{\phi_{\text{p}}}$ is trained on large-scale text-image pairs, $q(\boldsymbol{e}|x)$ is expected to be able to generate a token sequence describing important information of various inputs x.

Our key idea is to leverage $q(\boldsymbol{e}|x)$ as the reference distribution for maintaining the interpretability of the generated explanation $\hat{\boldsymbol{e}} \sim p_\phi(\boldsymbol{e}|x)$, where $p_\phi(\boldsymbol{e}|x)$ is the model distribution of g_ϕ. If $p_\phi(\boldsymbol{e}|x)$ and $q(\boldsymbol{e}|x)$ are sufficiently close, it can be guaranteed that the interpretability of the sequence generated by $p_\phi(\boldsymbol{e}|x)$ approximate to that by $q(\boldsymbol{e}|x)$. Concretely, we compute the KL divergence between $p_\phi(\boldsymbol{e}|x)$ and $q(\boldsymbol{e}|x)$ as the regularization term $\mathcal{R}_{\text{int}}$ in Eq. (1).

$$\begin{aligned}\mathcal{R}_{\text{int}}(\phi, \psi) &= D_{\text{KL}}(q \| p_\phi) \\ &= \mathbb{E}_{\boldsymbol{e}\sim q(\boldsymbol{e}|x)} \log \left(\frac{q(\boldsymbol{e}|x)}{p_\phi(\boldsymbol{e}|x)} \right).\end{aligned} \tag{4}$$

However, $D_{\text{KL}}(q\|p_\phi)$ is computationally intractable because it requires multiple sequential sampling over $\mathcal{E} = \mathcal{V}^L$ from $q(\boldsymbol{e}|x)$ and the back-propagation through all sampling processes of $p_\phi(e_l|x, \boldsymbol{e}_{<l})$. To approximate Eq. (4), we focus on the connection to knowledge distillation (Hinton, Vinyals, and Dean 2015). That is, minimizing Eq. (4) can be seen as a knowledge distillation from $g_{\phi_{\text{p}}}$ to g_ϕ. In such a sense, the approximation is

$$\begin{aligned}\mathcal{R}_{\text{int}}(\phi, \psi) &\approx -\sum_{\boldsymbol{e}\in\mathcal{E}} \mathbb{I}_{\boldsymbol{e}=\boldsymbol{e}_{\text{p}}} \log p_\phi(\boldsymbol{e}|x) \\ &= -\log p_\phi(\boldsymbol{e} = \boldsymbol{e}_{\text{p}}|x),\end{aligned} \tag{5}$$

where $\boldsymbol{e}_{\text{p}}$ is the sample from $q(\boldsymbol{e}|x)$ and $\mathbb{I}$ is the indicator function returning one when $\boldsymbol{e}$ equals to $\boldsymbol{e}_{\text{p}}$ or returning zero otherwise; we omit the constant terms from the approximation for the simplicity. As a concrete procedure, we first generate $\boldsymbol{e}_{\text{p}}$ from $g_{\phi_{\text{p}}}$ and then penalize the output logits of g_ϕ through the cross-entropy loss for each output token in a next token prediction task. This approximation technique is well-known as sequence-level knowledge distillation (Kim and Rush 2016) in the field of neural machine translation, and it works well in the knowledge distillation of auto-regressive sequence models. Sequence-level knowledge distillation corresponds to matching the modes of p and q and omits to transfer the uncertainty represented by the entropy $H(q)$ (Kim and Rush 2016). Nevertheless, we consider that this is sufficient for XBMs because the goal of XBMs is to provide interpretable explanations for target task predictions, not to replicate the pre-trained models perfectly. We call the regularization with Eq. (5) *explanation distillation*, and introduce it in training XBMs to maintain the text generation capability.

Algorithm 1: Training of XBMs

Require: Training dataset $\mathcal{D}$, vision encoder h_ψ, text decoder g_ϕ, classifier f_θ, pre-trained parameters $(\phi_{\text{p}}, \theta_{\text{p}})$, training batch-size B, step size η, trade-off parameter λ
Ensure: Trained models $(h_\psi, g_\phi, f_\theta)$

```
# Initialize parameters
φ ← φ_p,  ψ ← ψ_p
while not converged do
    {(x^i, y^i)}_{i=1}^B ~ D
    # Generating reference explanation
    {e_p^i}_i^B ← {generate(g_φp, h_p(x^i))}_i^B
    # Gumbel-softmax sampling
    {ê^i}_i^B ← {g_sampling(g_φ, h_ψ(x^i))}_i^B
    # Computing batch-mean losses
    L_cls^B ← (1/B) Σ_{i=1}^B ℓ_CE(f_θ(h_ψ(x^i), ê^i), y^i)
    R_int^B ← (1/B) Σ_{i=1}^B ℓ_CE(g_φ ∘ h_ψ(x^i), e_p^i)
    # Updating parameters via backprop.
    θ ← θ − η∇_θ(L_cls^B + λR_int^B), φ ← φ − η∇_φ(L_cls^B + λR_int^B),
    ψ ← θ − η∇_ψ(L_cls^B + λR_int^B)
end while
```

2.4 Algorithm

Training We show the training procedure in Algorithm 1. In the training loop, we first generate the reference and predicted explanations $\boldsymbol{e}_{\text{p}}$ and $\hat{\boldsymbol{e}}$ by $\text{generate}(\cdot)$ and $\text{g\_sampling}(\cdot)$, respectively (line 4 and 5). To approximate the mode of $q(\boldsymbol{e}|x)$ and ensure the quality as the reference, we generate $\boldsymbol{e}_{\text{p}}$ from frozen $g_{\phi_{\text{p}}}$ by beam search following the previous work (Kim and Rush 2016). For sampling $\hat{\boldsymbol{e}}$, we introduce the Gumbel-softmax trick (Jang, Gu, and Poole 2017) to retain the computation graph for the end-to-end training with back-propagation. The l-th token can be approximately sampled by

$$e_l = \text{softmax}((\log(g_\phi(h_\psi(x))) + \mathbf{g})/\tau), \tag{6}$$

where $\mathbf{g} = \{\text{g}_1, ..., \text{g}_{|\mathcal{V}|}\}$ is a vector of length $|\mathcal{V}|$ where each element is sampled from $\text{Gumbel}(0, 1)$ and τ is the temperature parameter. Intuitively, the temperature τ controls the diversity of the token outputs from g_ϕ; larger τ stimulates more diverse outputs. To obtain diverse and accurate tokens for describing input, we apply exponential annealing to the temperature values according to the training steps, i.e., $\tau^{(i+1)} = \tau^{(0)} \exp(-r_{\text{a}} i)$, where i and r_{a} are training step and annealing rate. This allows XBMs to focus on the diversity of the output tokens in the early training steps and on the quality in the later steps. We evaluate this design choice in Appendix E.1. After sampling $\boldsymbol{e}_{\text{p}}$ and $\hat{\boldsymbol{e}}$, we update all trainable parameters according to the objective function Eq. (1).

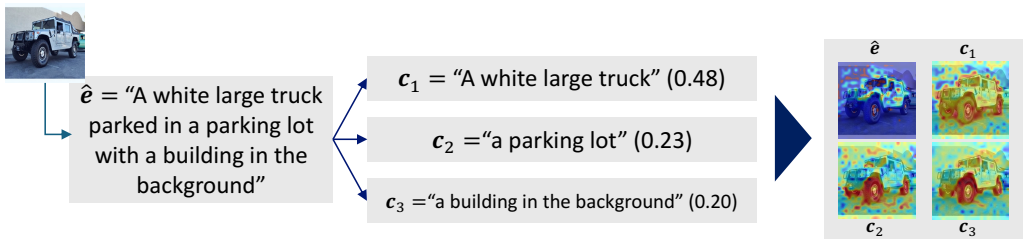

Figure 3: Explanation styles provided by XBMs. XBMs can output (i) text explanation directly generated from the explanation decoder, (ii) concept phrases with self-attention scores in the classifier, and (iii) cross-attention heatmap for the entire text explanation and each concept phrase. Concept phrases are constructed by a natural language parser, and the self-attention scores are computed in a middle layer of the classifier with respect to the `[CLS]` token for each concept phrase. Cross-attention heatmaps are the heatmap visualizations of cross-attention scores between input text tokens and image embedding tokens in the middle layer of the multi-modal classifier (a redder means a higher score).

Inference For the inference of test input x, we generate $\hat{\boldsymbol{e}}$ by beam search instead of the Gumbel-softmax trick, i.e., $\hat{\boldsymbol{e}} \leftarrow \text{generate}(g_\phi, h_\psi(x))$. Finally, we return the target label prediction $\hat{y} \leftarrow f_\theta(h_\psi(x), \hat{\boldsymbol{e}})$ and the explanation $\hat{\boldsymbol{e}}$ to users. Optionally, XBMs provide the other styles of explanation in addition to $\hat{\boldsymbol{e}}$ (Fig. 3). A *concept phrase* $\boldsymbol{c}$ is a noun phrase that compose $\hat{\boldsymbol{e}}$, which can be extracted by natural language parser automatically (Feng et al. 2022). Similar to the concept outputs of CBMs, $\boldsymbol{c}$ provides contributions of noun phrases in text explanations for the prediction. For example, if the classifier f_θ is implemented with transformer families with attention layers, we can interpret the contribution of $\boldsymbol{c}$ for the target prediction $\hat{y}$ via its self-attention scores as in Fig. 3 (ii). Furthermore, we can visualize the cross-attention scores between the text explanations and visual tokens as a heatmap, suggesting what the model perceives as a concept in input data (Fig. 3 (iii)).

3 Experiment

We evaluate XBMs on multiple visual classification tasks and pre-training models. We conduct qualitative and quantitative experiments on the explanation outputs of XBMs to evaluate the target performance and the interpretability. We also provide a more detailed analysis, including varying hyperparameters λ, τ and comparing explanation distillation with an alternative regularization loss in Appendix E.

3.1 Setting

Implementation Our basic implementation of XBMs is based on BLIP (Li et al. 2022) because of its simplicity; we denote this model as XBM-BLIP. That is, as the visual encoder h_ψ, we used the ViT-B/32 (Dosovitskiy et al. 2021). For the classifier f_θ, we used a BERT-base transformer (Devlin et al. 2019); we input $h_\psi(x)$ into the cross-attention layers when using a multi-modal classifier inspired by BLIP (Li et al. 2022). We initialized ϕ and ψ by the BLIP model pre-trained on image captioning tasks in the official repository[1]. We also report the results using larger pre-trained multi-modal models of LLaVA (Liu et al. 2023). We used v1.5 and v1.6 of LLaVA with multiple language model backbones (LLaMA2-7B (Touvron et al. 2023), Vicuna-7B (Chiang et al. 2023), and Mistral-7B (Jiang et al. 2023)); we denote these models as XBM-LLaVA. We provide detailed training settings in Appendix A.

Baselines We compare XBMs to black-box and interpretable baselines in performance and interpretability. **Fine-tuned BLIP-ViT** is the black-box baseline, which directly optimizes the visual encoder of BLIP via fine-tuning. **Label-free CBM** (Oikarinen et al. 2023) is a state-of-the-art concept bottleneck model, which automatically constructs pre-defined concept sets from ConceptNet (Speer, Chin, and Havasi 2017) or GPT-3 (Brown et al. 2020a) and then constructs concept embedding matrix via CLIP vision and text encoder. We used BLIP-ViT as the backbone vision encoder of label-free CBMs. **Frozen BLIP** baselines use frozen BLIP to generate text explanations and predict final labels by a multi-modal $f_\theta(h_\psi(x), \hat{\boldsymbol{e}})$ or text classifier $f_\theta(\hat{\boldsymbol{e}})$. We also show the results of **XBM w/o** $\mathcal{R}_{\text{int}}$, which updates g_ϕ only on the classification loss Eq. (2).

Datasets We used four image datasets for classification tasks in various domains: **Aircraft** (Maji et al. 2013), **Bird** (Welinder et al. 2010), **Car** (Krause et al. 2013), and **ImageNet** (Russakovsky et al. 2015). Aircraft, Bird, and Car are fine-grained image datasets, and ImageNet is a large-scale general image dataset. For datasets other than ImageNet, we randomly split a dataset into $9:1$ and used the former as the training set and the latter as the validation set. For ImageNet, we set the split ratio $99:1$ and used the official validation set as the test dataset.

Evaluation Metrics We report test accuracy as the target task performance. For the interpretability evaluations, we introduce **CLIP-Score** (Radford et al. 2021; Hessel et al. 2021), which is based on the cosine similarity between image embeddings and text embeddings on CLIP, i.e., higher is better. CLIP-score was originally used to evaluate image captioning based on the relevance of the output captions to the input images. Since it is highly sensitive to the hallucinations in the captions as reported in (Hessel et al. 2021), CLIP-score can be used to assess the factuality of explanations. For

[1] `model_base_caption_capfilt_large.pth` in https://github.com/salesforce/BLIP

	Aircraft			Bird		
	Test Acc. (↑)	CLIP-Score (↑)	GPT-2 Perplexity (↓)	Test Acc. (↑)	CLIP-Score (↑)	GPT-2 Perplexity (↓)
Fine-tuned BLIP-ViT	77.86±.30	N/A	N/A	83.48±.15	N/A	N/A
Label-free CBM (ConceptNet)	15.37±.17	0.5356	N/A	17.67±.40	0.6025	N/A
Label-free CBM (GPT-3)	44.47±.34	0.6153	N/A	77.74±.43	0.6904	N/A
Frozen BLIP + $f_\theta(h_\psi(x), \hat{\boldsymbol{e}})$	45.23±.32	0.6824	155.8	68.03±.10	0.7535	173.5
XBM w/o $\mathcal{R}_{\text{int}}$	70.78±.48	0.4730	322.6	61.94±.13	0.5137	431.0
XBM (Ours)	**74.09±.07**	**0.7151**	**129.8**	**80.99±.18**	**0.7942**	**166.8**
	Car			ImageNet		
	Test Acc. (↑)	CLIP-Score (↑)	GPT-2 Perplexity (↓)	Test Acc. (↑)	CLIP-Score (↑)	GPT-2 Perplexity (↓)
Fine-tuned BLIP-ViT	90.08±.35	N/A	N/A	65.21±.14	N/A	N/A
Label-free CBM (ConceptNet)	15.27±.13	0.5561	N/A	60.07±.42	0.6826	N/A
Label-free CBM (GPT-3)	77.91±.21	0.6091	N/A	64.28±.09	0.7026	N/A
Frozen BLIP + $f_\theta(h_\psi(x), \hat{\boldsymbol{e}})$	80.53±.29	0.6555	168.8	56.04±.49	0.7732	199.5
XBM w/o $\mathcal{R}_{\text{int}}$	86.59±.11	0.4792	415.3	66.58±.30	0.5020	517.1
XBM (Ours)	**89.47±.10**	**0.7173**	**131.8**	**67.83±.33**	**0.7920**	**122.8**

Table 1: Performance and Interpretability Evaluation of XBMs on multiple target datasets.

XBMs, we measured averaged CLIP-Scores between test inputs and the output explanations. For Label-free CBMs, we measured averaged CLIP-Scores between test inputs and the output concept texts with the binary output of the concept bottleneck layer greater than 0.05; this threshold follows Oikarinen et al. (2023). We also introduce **GPT-2 Perplexity** as a measure of fluency in XBM's output explanations. In general, perplexity scores on language models are calculated by the averaged cross-entropy of the next token probabilities and thus represent the fluency of the generated texts because the lower perplexity means that the sentence is composed of words that are likely to occur probabilistically. Inspired by Chan et al. (2023), we computed perplexity scores of explanations on GPT-2 (Radford et al. 2019). That is, the generated explanations are unbiasedly evaluated by an external language model. GPT-2 perplexity is helpful as a metric of the fluency of explanations because it shows the proximity to the natural text distribution learned by GPT-2. We used open-sourced GPT-2 in huggingface transformers (Wolf et al. 2019) to maintain reproducibility.

3.2 Design Evaluation of XBMs

Quantitative Evaluation Table 1 demonstrates the quantitative performance and interpretability of XBM-BLIP on the four target datasets. For the target performance, our XBMs outperformed the Label-free CBM baselines and achieved competitive performance with the black-box baseline in the test accuracy. In particular, XBM achieved high performance on datasets where label-free CBM did not perform well (i.e., Aircraft and Car). This can be caused by insufficient pre-defined concepts due to the limited vocabulary in ConceptNet and GPT-3 about describing objects in these datasets, whereas XBMs promote multi-modal understanding by training the explanation decoder to describe arbitrary objects useful for the target dataset with unlimited vocabulary. For the interpretability, XBMs outperformed CBMs in CLIP-Score. This indicates that the explanations from XBMs are more factual to the input images than the concept outputs of CBMs, which are in pre-defined concept sets.

Furthermore, the ablation study in the bottom rows of Table 1 shows that the objective function in Eq. (1) works effectively as we expected. Compared to the frozen BLIP baselines, which simply apply fixed pre-trained BLIP to generate text captions, our XBM significantly improved all of the test accuracy, CLIP-Score, and GPT-2 Perplexity. This suggests that optimizing text decoders with respect to target tasks guides the generated explanation to be informative and target-related for solving the task. We also confirm that the regularization term $\mathcal{R}_{\text{int}}$ by explanation distillation (Eq. (5)) is crucial to generate meaningful explanation; XBM w/o $\mathcal{R}_{\text{int}}$ catastrophically degraded CLIP-Score and GPT-2 Perplexity.

Qualitative Evaluation Table 2 shows the qualitative studies of explanations generated from XBMs; we also show the other examples in Appendix B. We computed the self-/cross attention scores in the middle of the transformer layers by following Zhang* et al. (2020). For comparison, we also show the top-3 concept outputs of CBMs and the generated captions of pre-trained BLIP, i.e., the initial states of XBMs. The text explanations of XBMs contain more detailed information than pre-trained BLIP. This is because the target classification loss $\mathcal{L}_{\text{cls}}$ forces the text decoders to describe target-related visual information to solve the task. Importantly, XBMs without explanation distillation $\mathcal{R}_{\text{int}}$ generate totally broken explanations, indicating the objective function of XBMs succeed in training the models to focus on the tokens related to the target task without the collapse of explanations. Meanwhile, the concept phrase explanations show the contributions to the final outputs (i.e., self-attention scores) for each noun phrase in the text explanations. In contrast to CBM's concepts, the concept phrases tend to be aligned with visual features appearing in input images rather than describing input by pre-defined knowledge. This is easy for humans to understand when interpreting the output of the models. Finally, the cross-attention heatmaps intuitively localize where the generated text explanations correspond to the input image spaces. We confirm that the heatmaps concentrate on objects through optimization and facilitate a

	Bird (Yellow Bellied Flycatcher) 
Pre-trained BLIP (Caption)	A bird perched on a wire fence with leaves on the ground and a blurry background.
Label-free CBMs (Top-3 Concept)	olive-colored sides (0.77) green head (0.55) a small, green body (0.52)
XBMs w/o $\mathcal{R}_{\text{int}}$ (Text Explanation)	2222222222222 2222222222
XBMs (Text Explanation)	A small green and yellow bird perched on a wire fence with leaves on the side.
XBMs (Top-3 Concept Phrase)	a small green and yellow bird (0.39) leaves on the side (0.32) a wire fence (0.21)
XBMs (Cross-Attn. Heatmap)	

Table 2: Qualitative evaluation of explanation outputs.

multi-modal understanding of the image in Section 3.5.

We also analyze the transition of the generated explanations in Fig. 4. We print the text explanation of XBMs and the top-10 word occurrence for all classes and the input class at 0, 20, and 40 epochs. According to the training epoch, the explanations and words progressively focus on detailed and target-related information in images. Concretely, in this example, the XBM is optimized to describe "yellow beak (mouth)", a key feature of California Gull. These suggest that XBMs can provide interpretable and useful explanations for humans.

3.3 XBMs with Large Vision-Language Models

Here, we evaluate the scalability and practicality of XBMs by combining them with larger vision-language models than BLIP. Instead of BLIP, we used the LLaVA models with various language model backbones (Liu et al. 2023). Table 3 shows that leveraging the high-performance vision-language model in XBMs yields better performance and interpretability scores, suggesting that the XBM's objective function can enhance the multi-modal understanding ability even if using the large vision-language models pre-trained on massive image-text pairs. This emphasizes the flexibility of XBM, consisting of arbitrary vision-language models.

3.4 XBMs with Text Classifier

Table 3 also evaluates XBMs with a text classifier $f_\theta(\hat{e})$, which relies only on text information for the final predictions. Although XBM-BLIP with $f_\theta(\hat{e})$ drops the performance from one with a multi-modal classifier $f_\theta(h_\psi(x), \hat{e})$, switching the backbone from BLIP to LLaVA (Liu et al. 2023) resolves the performance gap. This indicates that more sophisticated vision-language models make XBMs generate informative text explanations, and they can achieve practical performance even when not using input features $h_\psi(x)$. Appendix C further shows the results on the other datasets.

3.5 Evaluations of Cross-Attention Heatmap

The cross-attention heatmap explanation of XBMs visualizes the local input space regions correlated to the text explanation

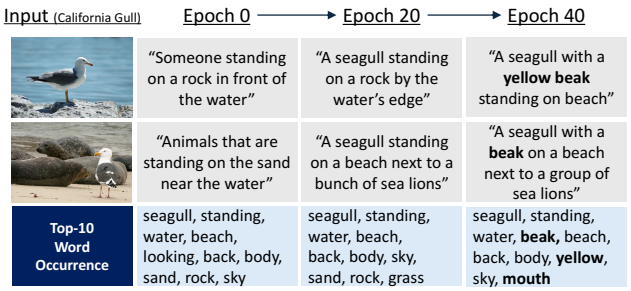

Figure 4: Transition of XBM's explanations during training.

in the classifier. To assess the validity of XBMs on improving multi-modal understanding, we evaluate the generated heatmaps on the ImageNet segmentation task by following Chefer, Gur, and Wolf (2021) and Gandelsman, Efros, and Steinhardt (2024). That is, we generate the heatmaps on the test set of ImageNet Segmentation (Guillaumin, Küttel, and Ferrari 2014) and compute the pixel accuracy, mean IoU (mIoU), and mean average precision (mAP) with the ground truth segmentation masks. Through this evaluation, we can evaluate how heatmaps cover the object of target classes in the pixel spaces. Table **??** shows the results. Compared to the frozen BLIP, XBM-BLIP improved all of the segmentation metrics. This means that the training objective of XBMs encourages the multi-modal understanding of target class objects on the models. In Appendix D, we further compare the XBM's heat maps with existing attribution methods, such as GradCAM (Selvaraju et al. 2017).

3.6 Reliability Evaluation via Human Intervention

CBMs allow the debugging of the model behavior through human intervention in the predicted concepts (Koh et al. 2020). Similarly, we can debug the behavior of XBMs by intervening in the generated explanations. Here, we show examples of an intervention in which all explanations are replaced to check the effect of the explanation quality on the final classification results. At inference, we replace the generated explanations from the explanation decoder with modified explanations. We tested two types of interventions: (i) randomized and (ii) ground-truth explanations. For randomized explanation, we used a token sequence uniformly sampled from the vocabulary space for the length of the originally generated explanation. For ground-truth explanation, we used the extended annotation set for Bird proposed by Reed et al. (2016). Table 5 shows the performance of the intervened XBM-BLIP models. The intervened explanations with randomized explanations significantly degraded the performance of XBM-BLIP, indicating that the generated explanations are essential to achieving high performance. In contrast, the intervention with ground-truth explanations largely improved the performance. This suggests that higher-quality explanations can yield higher performance, and intervening with human explanations is helpful for XBMs to improve their performance. In other words, the final prediction of XBMs largely depends on the content of the generated explanation $\hat{e}$, indicating that $\hat{e}$ is a reliable explanation for the final prediction.

	Text Classifier $f_\theta(\boldsymbol{e})$			Multi-modal Classifier $f_\theta(h_\psi(x), \boldsymbol{e})$		
	Test Acc. (↑)	CLIP-Score (↑)	GPT-2 Perplexity (↓)	Test Acc. (↑)	CLIP-Score (↑)	GPT-2 Perplexity (↓)
Frozen BLIP	9.97±.12	0.7732	199.5	56.04±.49	0.7732	199.5
XBM-BLIP	18.26±.31	0.8007	148.1	67.83±.33	0.7920	122.8
Frozen LLaVA-v1.5-LLaMA-7B	64.01±.46	0.7773	236.8	70.21±.18	0.7773	100.8
XBM-LLaVA-v1.5-LLaMA-7B	71.41±.25	0.8008	127.2	72.95±.16	0.7998	82.6
XBM-LLaVA-v1.6-Vicuna-7B	73.73±.30	0.8140	36.74	74.42±.23	0.8037	32.3
XBM-LLaVA-v1.6-Mistral-7B	72.14±.27	0.8037	20.67	74.04±.11	0.8130	21.7

Table 3: Evaluation of XBMs with text and multi-modal classifiers built on large vision-language models on ImageNet.

	Pixel Acc. (↑)	mIoU (↑)	mAP (↑)
Frozen BLIP	78.67	57.90	79.72
XBM-BLIP	80.90	60.80	80.18

Table 4: Evaluation of cross-attention map of XBMs on ImageNet Segmentation.

	Test Acc. (↑)	CLIP-Score (↑)	GPT-2 Perp. (↓)
XBM-BLIP	80.99	0.7942	166.8
Randomized Intervention	44.42	0.4497	4631.1
Ground-Truth Intervention	82.21	0.8179	104.5

Table 5: Evaluation of Intervened XBMs on Bird.

To conclude, these results support the debuggability of XBMs and the reliability of the generated explanations.

4 Related Work

The main research directions of the interpretability of black-box deep neural networks are briefly divided into attribution-based and concept-based methods. Attribution-based methods such as CAM (Zhou et al. 2016) and GradCAM (Selvaraju et al. 2017) generate a localization map representing important regions for the model predictions for specific classes. However, since the maps generated by attribution-based methods do not have information other than that they responded to the predictions, they are less interpretable regarding what semantic input features contribute to the output. In contrast to these methods, our XBMs can generate semantically interpretable heatmaps via cross-attention between image and text explanations, which can be decomposed at the level of noun phrases.

On the other hand, concept-based methods such as TCAV (Kim et al. 2018) and CBMs (Koh et al. 2020) compute contribution scores for pre-defined concepts on intermediate outputs of models. Among them, CBMs are highly relevant to our XBMs since both have interpretable intermediate layers in models. CBMs predict concept labels and then predict final class labels from the predicted concepts. The original CBMs have the challenge of requiring human annotations of concept labels (Zarlenga et al. 2022; Moayeri et al. 2023; Xu et al. 2024). Post-hoc CBMs (Yuksekgonul, Wang, and Zou 2023) and Label-free CBMs (Oikarinen et al. 2023) addressed this challenge by automatically collecting concepts corresponding to target task labels by querying large language models (e.g., GPT-3 (Brown et al. 2020b)) or existing concept banks (e.g., ConceptNet (Speer, Chin, and Havasi 2017)). However, CBMs' explanations are still restricted to pre-defined concepts, and they are not necessarily reliable because CBMs often predict the concepts without mapping to corresponding input regions (Huang et al. 2024). On the contrary, our XBMs directly generate natural language explanations to interpret the model outputs without pre-defined concepts.

Similar to our work, a few works attempted to generate linguistic explanations for target classification models (Hendricks et al. 2016; Nishida, Nishida, and Nishioka 2022). However, these methods require ground truth text explanations for training models, which are expensive and restrict applications. Our XBMs address this limitation by learning explanation generation by the classification loss and explanation distillation using a pre-trained text decoder.

5 Limitation

One of the limitations of XBMs is that they can not generate explanations based on user-defined concepts, which can be expressed by CBMs. In other words, XBMs are good at fluently explaining outputs in a general vocabulary because of their language model backbone but have difficulty giving interpretations for fixed concepts based on expert knowledge. A promising direction of future work is to associate the fluent explanations with user-defined concepts.

6 Conclusion

In this paper, we presented a novel interpretable deep neural networks called explanation bottleneck models (XBMs). By leveraging pre-trained vision-language models, XBMs generate explanations corresponding to input and output in natural language description, concept phrases with contribution scores, and cross-attention heatmaps on input spaces. To ensure both the target task performance and the explanation quality, XBMs are optimized by the target task loss with explanation distillation, which penalizes the divergence between the distributions of the training and pre-trained text decoders. Experiments show that XBMs can achieve both high target task performance and accurate and fluent explanations; they achieve competitive performance to black-box baselines and outperform CBMs in target test accuracy. Further, we found that XBMs' training can enhance the multi-modal understanding capability of vision-language models even when using large vision-language models pre-trained on massive image-text pairs. We believe that this work introduces a new perspective on natural language explanations and advances the study of interpretable deep models to the next paradigm.

Acknowledgements

We thank the members of the Kashima Laboratory at Kyoto University and our NTT colleagues, especially Han Bao and Daiki Chijiwa, for their helpful feedback on this research.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020a. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*.

Chan, D.; Petryk, S.; Gonzalez, J.; Darrell, T.; and Canny, J. 2023. CLAIR: Evaluating Image Captions with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Chauhan, K.; Tiwari, R.; Freyberg, J.; Shenoy, P.; and Dvijotham, K. 2023. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5948–5955.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 782–791.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A. R.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2022. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations*.

Gandelsman, Y.; Efros, A. A.; and Steinhardt, J. 2024. Interpreting CLIP's Image Representation via Text-Based Decomposition. In *International Conference on Learning Representations*.

Guillaumin, M.; Küttel, D.; and Ferrari, V. 2014. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110: 328–348.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, Q.; Song, J.; Hu, J.; Zhang, H.; Wang, Y.; and Song, M. 2024. On the Concept Trustworthiness in Concept Bottleneck Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21161–21168.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representation*.

Jiang, A.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B (2023). *arXiv preprint arXiv:2310.06825*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*.

Kim, Y.; and Rush, A. M. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *4th International IEEE Workshop on 3D Representation and Recognition*. Sydney, Australia.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*.

Maji, S.; Kannala, J.; Rahtu, E.; Blaschko, M.; and Vedaldi, A. 2013. Fine-Grained Visual Classification of Aircraft. *arXiv*.

Moayeri, M.; Rezaei, K.; Sanjabi, M.; and Feizi, S. 2023. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning*.

Nishida, K.; Nishida, K.; and Nishioka, S. 2022. Improving Few-Shot Image Classification Using Machine-and User-Generated Natural Language Descriptions. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 1421–1430.

Oikarinen, T.; Das, S.; Nguyen, L. M.; and Weng, T.-W. 2023. Label-free Concept Bottleneck Models. In *International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Ramaswamy, V. V.; Kim, S. S.; Fong, R.; and Russakovsky, O. 2023. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3).

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD Birds 200. Technical report, California Institute of Technology.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xu, X.; Qin, Y.; Mi, L.; Wang, H.; and Li, X. 2024. Energy-based concept bottleneck models: unifying prediction, concept intervention, and conditional interpretations. In *International Conference on Learning Representations*.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2023. Post-hoc Concept Bottleneck Models. In *International Conference on Learning Representations*.

Zarlenga, M. E.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Precioso, F.; Melacci, S.; Weller, A.; Lio, P.; et al. 2022. Concept embedding models. In *Advances in Neural Information Processing Systems*.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.