5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–23

Visual Explanations of Image-Text Representations via Multi-Modal Information Bottleneck Attribution

Ying Wang*, *New York University* YW3076@NYU.EDU
Tim G. J. Rudner*, *New York University* TIM.RUDNER@NYU.EDU
Andrew Gordon Wilson, *New York University* ANDREWGW@CIMS.NYU.EDU

Abstract

Vision-language pretrained models have seen remarkable success, but their application to high-impact safety-critical settings is limited by their lack of interpretability. To improve the interpretability of vision-language models, we propose a multi-modal information bottleneck (M2IB) objective that compresses irrelevant and noisy information, while preserving relevant visual and textual features. We demonstrate how M2IB can be applied to attribution analysis of vision-language pretrained models, increasing attribution accuracy and improving the interpretability of such models when applied to safety-critical domains such as medical diagnosis. Unlike commonly used unimodal attribution methods, M2IB does not require ground truth labels, making it possible to audit representations of vision-language pretrained models when multiple modalities but no ground truth data is available. Using CLIP as an example, we demonstrate the effectiveness of M2IB attribution and show that it outperforms CAM-based attribution methods both qualitatively and quantitatively.

1. Introduction

Vision-Language Pretrained Models (VL-PMs), such as the CLIP (Radford et al., 2021), have shown impressive performance on various downstream tasks by leveraging their complex structures and numerous parameters (Shen et al., 2022). However, the complexity of these models reduces interpretability and obscures their decision-making process, which hinders its application in safety-critical applications like medical diagnosis. To enhance transparency and detect potential biases, attribution methods for post hoc interpretability have been proposed, assigning contribution scores to each input feature.

We introduce an attribution method for identifying critical features and improving our understanding of image-text representations in VL-PMs using the information bottleneck principle (Tishby and Zaslavsky, 2015). Unlike standard unimodal methods, the proposed multi-modal information bottleneck formulation (M2IB) *does not require access to ground-truth data*. Instead, we insert an information bottleneck (Tishby and Zaslavsky, 2015) into the trained neural network and aim to minimize the retained information in the target layer while preserving the relevance between image and text features. We perform a qualitative and quantitative empirical evaluation and find that M2IB is able to successfully identify key features relevant to *both* image and text inputs (Figure 1). Our contributions are as follows:

- We adopt information bottleneck attribution to multi-modal settings and propose an attribution method to interpret the image-text representation obtained from VL-PMs.
- To obtain a tractable objective function, we consider a Gaussian moment-matching procedure with empirical covariance matrices computed from the encoder representations.
- We demonstrate on several datasets, including safety-critical medical data, that the proposed method outperforms existing CAM-based attribution methods—GradCAM, GradCAM++, HiResCAM, EigenCAM—quantitatively and qualitatively.

* Corresponding authors.

© Y. Wang, T.G.J. Rudner & A.G. Wilson.

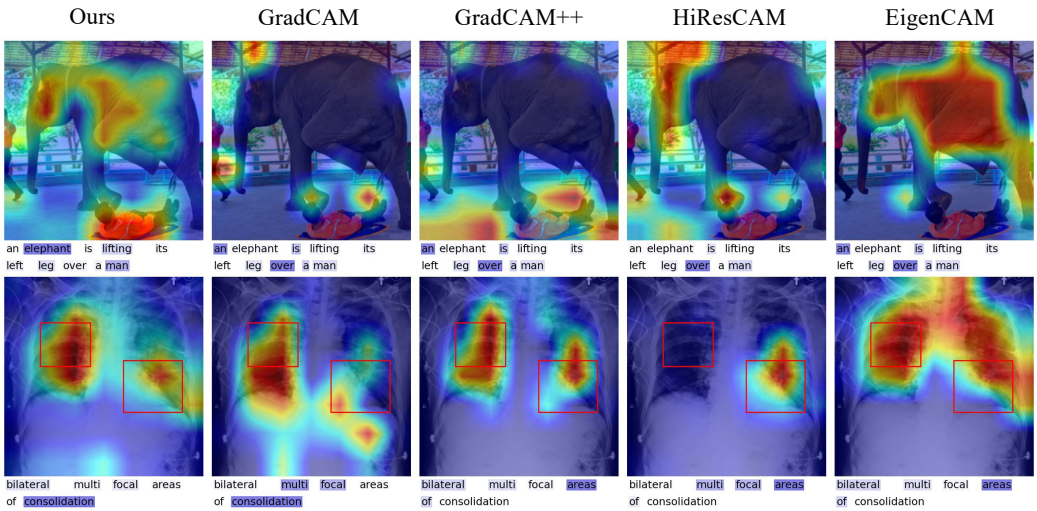


Figure 1: Example attribution maps on image and text. The red boxes on the second row are the ground-truth bounding boxes from MS-CXR (Boecking et al., 2022). Our method successfully identifies relevant objects in the given image and text.

2. Related Work

CAM-Based Attribution. Class Activation Mapping (CAM) (Zhou et al., 2015) is originally proposed for CNN on images, which generates a saliency map by weighting the activation maps of the last convolutional layer before the global pooling layer. Grad-CAM (Gradient-weighted Class Activation Mapping; Selvaraju et al. (2016)) extends CAM to a wider range of models by removing the constraint of a global pooling layer and using average pixel-wise gradients to weight the activation instead. However, GradCAM sometimes fails to capture multiple occurrences of the target and is not able to locate the entire object. To address the weakness of GradCAM, similar CAM methods are then proposed to increase the precision of the saliency maps. For example, Grad-CAM++ (Chattopadhay et al., 2018)) adopts pixel-wise weighting of the gradients, HiRes-CAM (High-Resolution Class Activation Mapping; Draelos and Carin (2021)) uses pixel-wise multiplication for gradients and activations, and EigenCAM (Eigen Class Activation Mapping; Muhammad and Yeasin (2020)) uses the principle components instead of gradients because gradients are usually noisy. This family of CAM methods can also be adapted to transformer-based models like ViT (Dosovitskiy et al., 2021) and Swin Transformers (Liu et al., 2021).[1]

Information-Theoretic Attribution. Schulz et al. (2020b) use information bottleneck attribution (IBA), where an information bottleneck is inserted into a layer of a trained neural network to distill the essential features for prediction. IBA is a model-agnostic method and shows impressive results on vision models including VGG-16 (Liu and Deng, 2015) and ResNet-50 He et al. (2016)). Subsequently, IBA is applied to language transformers and also outperforms other methods on this task (Jiang et al., 2020). However, IBA has thus far been focused on only one modality and has only been adopted in supervised learning. To the best of our knowledge, there is no previous research on applying the information bottleneck principle to aid our understanding of the inner mechanisms of VL-PMs.

1. https://github.com/jacobgil/pytorch-grad-cam

3. Attribution via a Multi-Modal Information Bottleneck Principle

In this section, we introduce a simple, multi-modal extension of the information bottleneck principle and explain how to adapt it to be used for feature attribution.

3.1. A Multi-Modal Information Bottleneck Principle

To make latent representation extract relevant information from input, the information bottleneck principle (Tishby and Zaslavsky, 2015) suggests two competing objectives during compression (minimize the mutual information between latent representation $Z \doteq f^{\ell}(X)$ at a given neural network layer ℓ and input X) and fitting (maximize the mutual information between latent representation Z and output Y). Here, we consider inserting an information bottleneck into layer m of a trained neural network and denote the output of the m-th layer after activation as Z. This results in the following loss function:

$$\mathcal{L} = I(Z, X) - \beta I(Z, Y), \tag{1}$$

where $I(\cdot, \cdot)$ is the mutual information function and β is a scaling hyperparameter.

In the context of image-text representation learning, however, the loss function above is not suitable, since—instead of using task-specific labels—we wish to extract information from (text feature, image feature) pairs instead from (feature, label) pairs, making the above definition of relevancy insufficient. Doing so requires a multi-modal information bottleneck principle more akin to self-supervised learning methods for image–test representation learning that only uses (text, image) pairs for learning (Baevski et al., 2022; Mu et al., 2021).

This learning problem fundamentally differs from supervised attribution map learning for uni-modal tasks. For example, we may have an image V_{bear} of a bear and its corresponding label L_{bear} "bear". For an image classification task, we can simply minimize the $I(V_{\text{bear}}; Z_{\text{text}})$ and maximize $I(L_{\text{bear}}; Z_{\text{image}})$ where Z_{image} and Z_{text} are the latent representations. In contrast, in image-text representation learning, we typically have text descriptions, such as "This is a picture of a bear" (L'_{bear}), instead of labels (Radford et al., 2021). In this setting, both V_{bear} and L'_{bear} are "inputs" without a pre-defined corresponding label. To obtain a task-agnostic image-text representation independent from any task-specific ground-truth labels, we would like to use both input modalities and define a multi-modal information bottleneck principle and whereas the outputs are closely dependent on the specific downstream task. This requires defining an alternative to the "fitting term" $I(Z, Y)$.

Fortunately, there is a natural proxy for the relevance of information in multi-modal data. If image and text inputs are related, a good image encoding should contain information about the text, while a good text encoding should include information about the image Based on this intuition, we can express a multi-modal information bottleneck for image and text representations, respectively:

$$\mathcal{L}_{\text{image}} = I(Z_{\text{image}}, X_{\text{image}}) - \beta_{\text{image}} I(Z_{\text{image}}, Z_{\text{text}}) \tag{2}$$

$$\mathcal{L}_{\text{text}} = I(Z_{\text{text}}, X_{\text{text}}) - \beta_{\text{text}} I(Z_{\text{text}}, Z_{\text{image}}). \tag{3}$$

3.2. Information Bottleneck for Attribution

To obtain an attribution map of image and text, we define an information bottleneck attribution method for multi-modal data. To restrict the information flow, we change the

"bottleneck terms" $I(Z_{\text{text}}, Z_{\text{image}}) = I(Z_{\text{image}}, Z_{\text{text}})$ by adapting the masking approach in Schulz et al. (2020b). Specifically, we add *input-indepdendent* noise to the layer output Z to obtain a masked output T,

$$T(\lambda_X) = \lambda_X \odot Z + (\mathbf{1} - \lambda_X) \odot \epsilon, \tag{4}$$

where $\odot$ is the Hadamard product, $\mathbf{1}$ is an all-one matrix having the same dimension as Z and λ_X, and we let $\epsilon \sim N(\epsilon; \mu_\epsilon, \sigma_\epsilon^2)$ for some input-independent μ_Z and σ_Z to ensure that the masked output has the same magnitude as the original feature. In extreme cases, $\lambda_i = 1$ means no noise is added at index i, so T_i will be the same as the original Z_i, whereas $\lambda_i = 0$ means T_i will be pure noise.

We then obtain multi-modal information bottleneck attribution (M2IB) objectives

$$\mathcal{L}_{\text{image}}(\lambda_{X_{\text{image}}}) = I(T_{\text{image}}(\lambda_{X_{\text{image}}}), Z_{\text{image}}) - \beta_{\text{image}} I(T_{\text{image}}(\lambda_{X_{\text{image}}}), Z_{\text{text}}) \tag{5}$$

$$\mathcal{L}_{\text{text}}(\lambda_{X_{\text{text}}}) = I(T_{\text{text}}(\lambda_{X_{\text{text}}}), Z_{\text{text}}) - \beta_{\text{text}} I(T_{\text{text}}(\lambda_{X_{\text{text}}}), Z_{\text{image}}), \tag{6}$$

which we optimize with respect to a input-specific parameters $\lambda_{X_{\text{image}}}$ and $\lambda_{X_{\text{text}}}$ for the image and text representations, respectively, and β_{image} and β_{text} are hyperparameters. To make these objectives more tractable, we use a standard trick and replace the first term in the above equations with a variational upper bound. Note that $I(T, Z) = \mathbb{E}[\mathbb{D}_{\text{KL}}(p_{T|Z} \parallel p_T)]$, where $T|Z$ can be sampled empirically whereas p_T does not have an analytic expression because the integral $p_T(t) = \int p_{T|Z}(t|z) p_Z(z) \, \mathrm{d}z$ is intractable. Thus, we approximate $p_T(t)$ by $q_T(t) = \mathcal{N}(t; m_T, s_T)$, which assumes all dimensions of T independently follow a Gaussian distribution. This approximation leads to an upper bound on $I(T, Z)$:

$$I(T, Z) = \mathbb{E}[\mathbb{D}_{\text{KL}}(p_{T|Z} \parallel q_T) - \mathbb{D}_{\text{KL}}(p_T \parallel q_T) \leq \mathbb{E}[\mathbb{D}_{\text{KL}}(p_{T|Z} \parallel q_T)]. \tag{7}$$

A derivation of this upper bound can be found in Appendix 1. Thus, we obtain a more tractable upper bound on the compression term. Since we wish to minimize the mutual information, minimizing the upper bound has a similar effect as direct minimization.

While uni-modal information bottleneck attribution uses ground-truth labels and compute the "fitting term" in the objective via a cross-entropy loss, the objectives for multi-modal information bottleneck (M2IB) attribution for vision–text given above require computing the mutual information between the masked embedding of one modality and the unmasked embedding of the other modality, which is not in general tractable and we require further approximation to obtain a tractable objective function.

For VL-PMs like CLIP, image-text representations are jointly trained and aligned in one embedding space. To approximate the mutual information between T and Z, consider three different estimation procedures: The analytically tractable mutual information between two moment-matched Gaussian distributions, the Pearson correlation coefficient, and the cosine similarity. For details on the Pearson correlation coefficient and cosine similarity estimators, see Appendix 1. We include an empirical comparison in Appendix 2.

We take advantage of the fact that image-text representations are aligned in one embedding space to obtain a tractable estimator of $I(T_1(\lambda_{X_1}), Z_2)$, where the subscripts denote different modalities. In particular, we approximate the distributions over $T_1(\lambda_{X_1})$, Z_2, and $[T_1(\lambda_{X_1})^\top, Z_2^\top]^\top$ by $\tilde{p}_{T_1} = \mathcal{N}(\mu_{T_1}, \Sigma_{T_1})$, $\tilde{p}_{Z_2} = \mathcal{N}(\mu_{Z_2}, \Sigma_{Z_2})$, and $\tilde{p}_{T_1,Z_2} = \mathcal{N}(\mu_{T_1,Z_2}, \Sigma_{T_1,Z_2})$, respectively, with covariance matrices given by

$$\Sigma_{T_1} \doteq T_1 T_1^\top + \varepsilon_{T_1} \mathbf{I} \text{ and } \Sigma_{Z_2} \doteq Z_2 Z_2^\top + \varepsilon_{Z_2} \mathbf{I} \text{ and } \Sigma_{T_1,Z_2} \doteq [T_1^\top, Z_2^\top]^\top [T_1^\top, Z_2^\top] + \varepsilon_{T_1,Z_2} \mathbf{I}, \tag{8}$$

respectively, where ε_{T_1}, ε_{Z_2}, and ε_{T_1,Z_2} are diagonal offsets that ensure the resulting matrices are positive semi-definite. We then obtain an estimator for the mutual information, given by

$$\hat{I}(T_1, Z_2) = 0.5(\ln(\det(\Sigma_{T_1}) \det(\Sigma_{Z_2})) - \ln\det(\Sigma_{T_1,Z_2})) \tag{9}$$

Combining Equation (7) and Equation (9), we obtain the loss function estimator

$$\mathcal{L}(\lambda) = \mathbb{E}[\mathbb{D}_{\mathrm{KL}}(p_{T|Z} \parallel q_T)] - \beta \hat{I}(T_1, Z_2). \tag{10}$$

For image and text, we thus have the objective functions

$$\mathcal{L}_{\text{image}}(\lambda) = \mathbb{E}[\mathbb{D}_{\mathrm{KL}}(P_{T_{\text{image}}|Z_{\text{image}})} \parallel Q_{T_{\text{image}})})] - \beta_{\text{image}} \hat{I}(T_{\text{image}}, Z_{\text{text}})) \tag{11}$$

$$\mathcal{L}_{\text{text}}(\lambda) = \mathbb{E}[\mathbb{D}_{\mathrm{KL}}(P_{T_{\text{text}}|Z_{\text{text}})} \parallel Q_{T_{\text{text}})})] - \beta_{\text{text}} \hat{I}(T_{\text{text}}, Z_{\text{image}})), \tag{12}$$

which, for a given pair of data points (X_{image}, X_{text}), are optimized with respect to $\lambda_{X_{\text{image}}}$ and $\lambda_{X_{\text{text}}}$, and collectively have hyperparameters $\{\beta_{\text{image}}, \beta_{\text{text}}, \varepsilon_{T_1}, \varepsilon_{Z_2}, \varepsilon_{T_1,Z_2}, \mu_\epsilon, \sigma_\epsilon\}$.

4. Empirical Evaluation

We evaluate the proposed attribution method using CLIP (Radford et al., 2021) on **(i)** Conceptual Captions (Sharma et al., 2018) consisting of diverse images and captions from the web, and**(ii)** MS-CXR (Local Alignment Chest X-ray dataset; Boecking et al. (2022)), which contains chest X-rays and texts describing radiological findings, complementing MIMIC-CXR (MIMIC Chest X-ray; Johnson et al. (2019)) by improving the bounding boxes and captions.

4.1. Experiment Setup

For all experiments, we use pretrained CLIP model with ViT-B/32 (Dosovitskiy et al., 2021) as the image encoder and a 12-layer self-attention transformer as the text encoder. For Conceptual Captions, we use the pretrained weights of `openai/clip-vit-base-patch32`.[2] For the MIMIC-CXR dataset, we compare the results of pretrained CLIP and CLIP that is finetuned on MIMIC-CXR and compare the impact of finetuning in Appendix 5. For each {image, caption} pair, we insert an information bottleneck into the given layer of the text encoder and image encoder of CLIP separately, then train the bottleneck using the same setup as the *Per-Sample Bottleneck* of original IBA (Schulz et al., 2020a), which duplicates a single sample for 10 times to stabilize training and runs 10 iterations using the Adam optimizer with a learning rate of 1. Experiments show no significant difference between different learning rates and more training steps. The crucial hyper-parameters here are the index of the layer m, the scaling factor β, and the variance σ^2, as shown in Figure 3. For CAM baselines, we use the implementation in `pytorch-gradcam` library.[3] For a discussion of hyperparameters in the multi-modal information bottleneck objective, see Appendix 2. For an ablation study on the effect of variations in the hyperparameters, see Figure 3 (also in Appendix 2). For an ablation study on the effect of variations in the mutual information estimators, see Appendix 2.2.

2. See `https://huggingface.co/openai/clip-vit-base-patch32`.
3. See `https://github.com/jacobgil/pytorch-grad-cam`.

Table 1: Quantitative Results. The bold number in grey is the best in the row.

	Methods	GradCAM	GradCAM++	HiResCAM	EigenCAM	Ours
CC image	% Conf. Drop ↓	4.35 ± 0.25	5.06 ± 0.51	4.48 ± 0.52	4.05 ± 4.05	**1.11** ± 0.14
	% Conf. Incr. ↑	20.6 ± 4.22	16.00 ± 3.85	13.2 ± 2.99	13.00 ± 2.97	**35.60** ± 3.61
	% ROAD Comb. ↑	0.86 ± 0.21	0.96 ± 0.09	0.92 ± 0.21	0.01 ± 0.09	**1.49** ± 0.14
	% ROAR+ ↑	23.23 ± 25.43	13.03 ± 36.83	-3.43 ±23.75	-2.85 ± 22.25	**32.90** ± 32.93
CC text	% Conf. Drop ↓	7.28 ± 0.19	7.04 ± 0.36	7.24 ± 0.34	7.05 ± 0.34	**3.47** ± 0.32
	% Conf. Incr ↑	4.60 ± 1.20	3.00 ± 1.41	4.40 ± 1.50	3.00 ± 1.41	**14.6** ± 3.01
	% ROAR+ ↑	16.58 ± 51.6	24.40 ± 54.26	28.84 ± 25.89	15.43 ± 48.17	**33.35** ± 56.79
MSCXR image	% Conf. Drop ↓	1.61 ± 0.13	2.52 ± 0.17	1.69 ± 0.26	2.66 ± 0.26	**0.51** ± 0.05
	% Conf. Incr. ↑	32.25 ± 1.79	20.25 ± 1.92	26.00 ± 5.24	15.25 ± 1.48	**43.8** ± 3.87
	% ROAD Comb. ↑	0.30 ± 0.09	0.49 ± 0.13	0.46 ± 0.10	-0.10 ± 0.05	**0.78** ± 0.10
	% ROAR+ ↑	30.38 ± 3.02	30.76 ± 5.52	32.52 ± 4.51	35.86 ±2.05	**50.54** ± 4.12
	% Localization ↑	8.95 ± 0.41	7.18 ± 0.67	13.07 ± 0.95	8.47 ± 0.43	**17.60** ± 0.83
MSCXR text	% Conf. Drop ↓	10.99 ± 0.47	8.31 ± 0.28	10.40 ± 0.60	8.44 ± 0.29	**4.37** ± 0.12
	% Conf. Incr. ↑	3.25 ± 1.09	4.50 ± 1.50	4.50 ± 2.96	4.50 ± 1.50	**5.6** ± 0.80
	% ROAR+ ↑	15.29 ± 7.41	13.30 ± 6.62	12.16 ± 6.34	9.56 ± 4.66	**15.42** ± 10.43

4.2. Results

We compare our method with 4 CAM-based methods (GradCAM (Selvaraju et al., 2016), GradCAM++ (Chattopadhay et al., 2018), HiResCAM (Draelos and Carin, 2021), EigenCAM (Muhammad and Yeasin, 2020)). As shown in Figure 1 and Appendix 3, our method is able to capture all relevant objects appearing in both modalities, while other methods tend to focus on one major object.

To quantitatively evaluate the performance of our model, we conduct the following localization and degradation tests and summarize the results in Table 1. Given the model under evaluation achieves state-of-the-art performance, a good attribution method should be able to detect the relevant objects in the image according to the text (i.e., zero-shot localization, see Appendix 4.1). Moreover, removing pixels or tokens with lower attribution scores according to our method generally increases the mutual information with the other modality (quantified by the evaluation metrics "Confidence Drop", "Confidence Increase", and "Remove and Debias" (ROAD), which are described in detail in Appendix 4.2), while masking by our attribution map generally decreases the relevance with the other modality and makes the model perform worse after retraining (ROAR+, see Appendix 4.2). Our method outperforms all CAM-based methods in all numerical metrics.

To further confirm the effectiveness and understand the limitation of our method, we include a sanity check in Appendix 5 and an error analysis in Appendix 6.

5. Discussion and Conclusions

The proposed information-theoretic approach can be extended easily to other vision-language models, and even beyond image-text representations. Its main limitation is that the features of all modalities must be projected into a shared embedding space—which has been widely adopted as a convention in state-of-the-art multi-modal models. We provided evidence that M2IB increases attribution accuracy and improves the interpretability of complex machine learning models. We hope that this work encourages further research into multi-modal information-theoretic attribution methods that can help introduce modern machine learning methods into safety-critical domains where interpretability is critical.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *CoRR*, abs/1810.03292, 2018. URL http://arxiv.org/abs/1810.03292.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. *CoRR*, abs/2202.03555, 2022. URL https://arxiv.org/abs/2202.03555.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 1–21, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20059-5.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks, 2021.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. Inserting Information Bottlenecks for Attribution in Transformers. In *Findings of the Association for Computational*

Linguistics: EMNLP 2020, pages 3850–3857, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.343. URL https://aclanthology.org/2020.findings-emnlp.343.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019. URL http://arxiv.org/abs/1901.07042.

Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015. doi: 10.1109/ACPR.2015.7486599.

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 (ICCV)*, pages 10012–10022, October 2021.

Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. *CoRR*, abs/2112.12750, 2021. URL https://arxiv.org/abs/2112.12750.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *CoRR*, abs/2008.00299, 2020. URL https://arxiv.org/abs/2008.00299.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18770–18795. PMLR, 2022.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020a. URL https://iclr.cc/virtual_2020/poster_S1xWh1rYwB.html.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020b. URL https://iclr.cc/virtual_2020/poster_S1xWh1rYwB.html.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL http://arxiv.org/abs/1610.02391.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zf_Ll3HZWgy.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015. URL http://arxiv.org/abs/1512.04150.

Appendix

Visual Explanations of Image-Text Representations via Multi-Modal Information Bottleneck Attribution

Table of Contents

1 Derivations **11**

2 Experimental Details and Further Experimental Results **12**
2.1 Hyperparameters 12
2.2 Comparison of Different Estimators for fitting term 15

3 Attribution Maps for Additional Examples **17**

4 Quantiative Metrics **19**
4.1 Localization Test 19
4.2 Degradation Test 19

5 Sanity Check **21**

6 Error Analysis and Limitations **22**

1. Derivations

Upper bound on mutual information:

$$
\begin{aligned}
I(T,Z) &= \mathbb{E}[\mathbb{D}_{\mathrm{KL}}(p_{T|Z} \parallel p_T)] && (1.1)\\
&= \int_Z p(z) \left(\int_T p(t|z) \log \frac{p(t|z)}{p(t)} \, \mathrm{d}t \right) \, \mathrm{d}z && (1.2)\\
&= \int_Z \int_T p(t,z) \log \frac{p(t|z)}{p(t)} \, \mathrm{d}t \, \mathrm{d}z && (1.3)\\
&= \int_Z \int_T p(t,z) \log \frac{p(t|z)}{p(t)} \frac{q(t)}{q(t)} \, \mathrm{d}t \, \mathrm{d}z && (1.4)\\
&= \int_Z \int_T p(t,z) \log \frac{p(t|z)}{q(t)} \, \mathrm{d}t \, \mathrm{d}z + \int_Z \int_T p(t,z) \log \frac{q(t)}{p(t)} \, \mathrm{d}t \, \mathrm{d}z && (1.5)\\
&= \int_Z \int_T p(t,z) \log \frac{p(t|z)}{q(t)} \, \mathrm{d}t \, \mathrm{d}z + \int_T p(t) \left(\int_Z p(z|t) dz \right) \log \frac{q(t)}{p(t)} \, \mathrm{d}t \, \mathrm{d}z && (1.6)\\
&= \mathbb{E}[\mathbb{D}_{\mathrm{KL}}(p_{T|Z} \parallel q_T)] - \mathbb{D}_{\mathrm{KL}}(p_T \parallel q_T) && (1.7)\\
&\leq \mathbb{E}[\mathbb{D}_{\mathrm{KL}}(p_{T|Z} \parallel q_T)] && (1.8)
\end{aligned}
$$

2. Experimental Details and Further Experimental Results

2.1. Hyperparameters

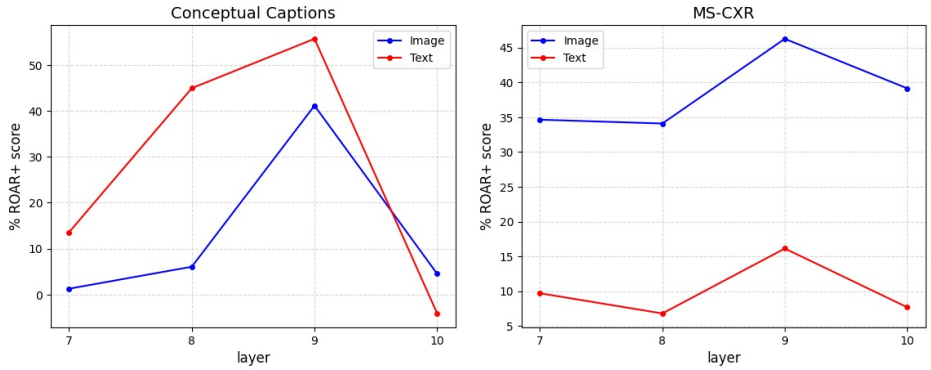
ε_{T_1}, ε_{Z_2}, and ε_{T_1,Z_2} are diagonal jitter terms added on the covariance matrics Σ_{T_1}, Σ_{Z_2}, and Σ_{T_1,Z_2}. We use the smallest value ($\varepsilon_{T_1} = 1$, $\varepsilon_{Z_2} = 1$, $\varepsilon_{T_1,Z_2} = 1$) such that the resulting matrices are positive semi-definite.

β controls the relative importance of the fitting term. The larger the β is, the more information is allowed to flow through this layer. As shown in Figure 3(a), too large and too small β generates similar attribution maps in terms of relative importance. However, too large β allows nearly everything through the bottleneck, whereas too small β nearly discards everything.

μ_ϵ and σ_ϵ control the values of the noise added to the intermediate representations. Since we insert the information bottleneck after layer normalization, we fix μ_ϵ to be 0. When σ_ϵ is very small, the values of the noise will be close to 0, thus having minimal impact on the intermediate representations. This effect is similar to the situation when β is very large and IB will add almost no noise (Figure 3(b)). σ_ϵ also directly affect the compression term as smaller σ_ϵ will lead to higher KL divergence. Thus, σ_ϵ and β are correlated with each other and we perform a grid search to find the best combination.

Layer m where the information bottleneck is inserted also impacts the attribution. Inserting the bottleneck too early will prevent the model from learning informative features while inserting the bottleneck too late reduces the impact (Figure 3(c)). We also observe that the attribution of texts is usually more stable than images.

These hyperparameters can be chosen according to numerical metrics mentioned in Appendix 4.2. We perform a grid search for the best combination of $\beta = \{1, 10, 100, 1000\}$ and $\sigma_\epsilon^2 = \{1, 0.1, 0.01, 0.001\}$ and $l = \{7, 8, 9, 10\}$ (indexing from 0), and find the best layer index is 9 for both datasets (2). Then, we fix the layer index and perform a grid search for the best combination of β and σ_ϵ, as shown in Table 2.



(a) (b)

Figure 2: Layer index hyperparameter selection. The plots show the average ROAR+ score for different layer indices over 3 random seeds. Layer index 9 gives the best score for both modalities of both datasets.

Table 2: Hyperparameter tuning results for β and σ_ϵ: We calculate the ROAR+ score for different combinations of β and σ_ϵ for 3 runs and report the average. For each table, the highest score is in bold and indicates the performance is optimal for this set of hyperparameters. Since ROAR+ uses binarized saliency maps (75% threshold for image pixels and 50% threshold for text tokens, it mainly focuses on the features with high attribution scores and neglects the change in the attribution for less important features. Thus, sometimes hyperparameters with slightly lower ROAR+ scores might generate more visually appealing results, as shown in 3. But the numerical results are consistent with qualitative examples in general.

(a) Conceptual Captions - Image

	σ_ϵ^2=1	σ_ϵ^2=0.1	σ_ϵ^2=1e-2	σ_ϵ^2=1e-3
β=1	27.71	61.16	-49.40	-17.36
β=10	25.34	-2.57	-0.73	-31.00
β=1e2	18.10	8.24	**117.78**	48.37
β=1e3	-15.13	0.45	28.84	-8.18

(b) Conceptual Captions - Text

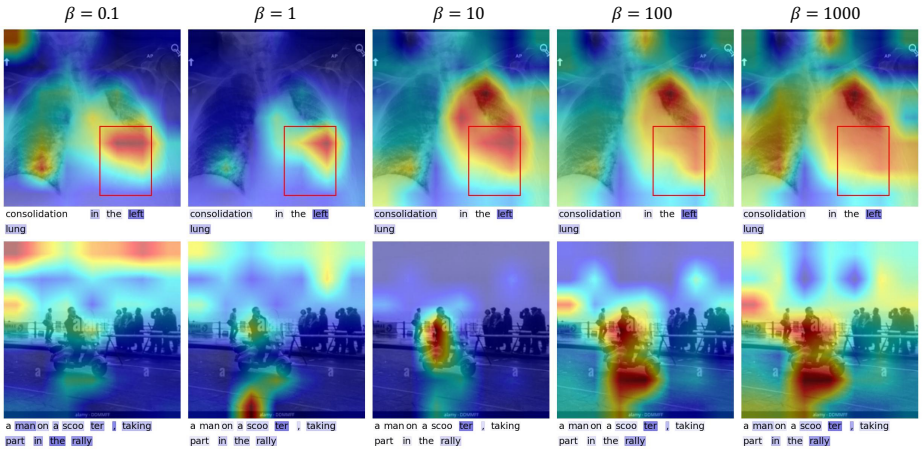
	σ_ϵ^2=1	σ_ϵ^2=0.1	σ_ϵ^2=1e-2	σ_ϵ^2=1e-3
β=1	10.22	-4.76	-21.36	-7.89
β=10	39.14	-2.61	57.80	30.03
β=1e2	**65.16**	31.42	24.76	-38.70
β=1e3	8.82	0.46	13.63	-1.09

(c) MSCXR - Image

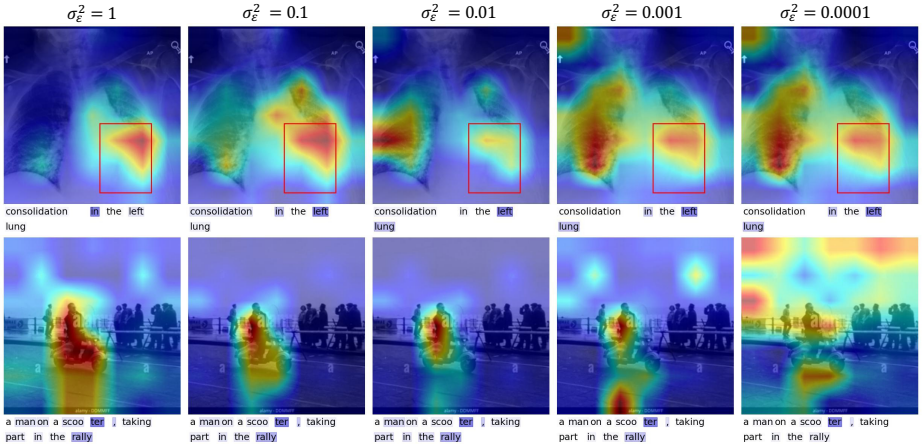
	σ_ϵ^2=1	σ_ϵ^2=0.1	σ_ϵ^2=1e-2	σ_ϵ^2=1e-3
β=1	44.01	27.50	38.68	55.08
β=10	43.98	46.74	34.63	31.38
β=1e2	**56.00**	45.29	30.06	33.73
β=1e3	40.45	52.58	27.53	30.99

(d) MSCXR - Text

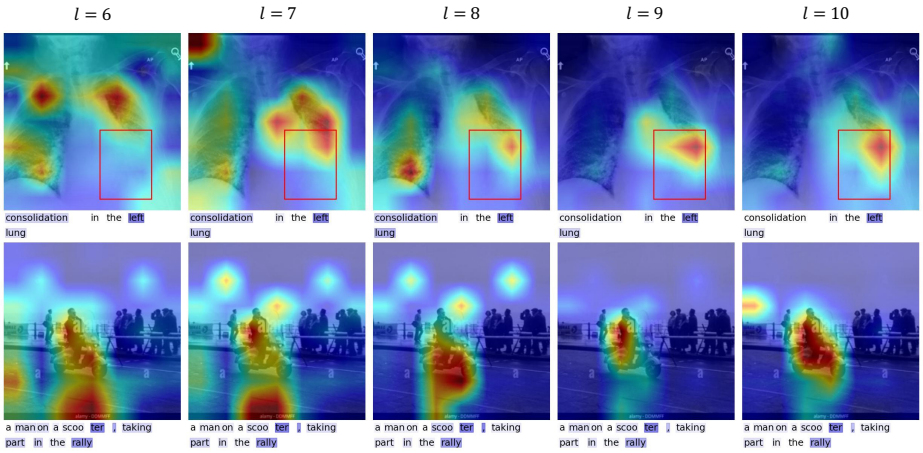
	σ_ϵ^2=1	σ_ϵ^2=0.1	σ_ϵ^2=1e-2	σ_ϵ^2=1e-3
β=1	10.18	5.46	**21.87**	-3.10
β=10	6.50	6.14	8.77	9.21
β=1e2	5.56	6.85	11.27	0.96
β=1e3	9.34	14.63	18.75	18.71



(a) Impact of scaling factor β: Higher β means higher weight of the fitting term.



(b) Impact of noise variance σ_ϵ^2: Smaller σ_ϵ^2 means lower impact of compression.



(c) Impact of the index of the layer where IB is inserted l: Larger l means deeper layer in the network.

Figure 3: Visualization of the impact of different hyperparameters. β and σ_ϵ^2 that make fitting term and compression at a similar scale and deeper layer l, usually give better performance. Note that the attribution score is assigned to each token, instead of each word, due to tokenization.

2.2. Comparison of Different Estimators for fitting term

Since image and text features are projected into a shared embedding space, we consider the following three estimators for the fitting term $I(T, Z)$. We have presented a single-sample multivariate Gaussian estimator in Section 3.2, where the latent space dimension determines the dimensionality of the multivariate Gaussian distribution. Alternatively, we also consider an estimator where each latent space dimension is viewed as a sample of one-dimensional representations. Computing the empirical variance and again making a Gaussian moment matching assumption, we obtain the mutual information estimator

$$\hat{I}(T_1, Z_2) = \frac{1}{2} \ln \left(\frac{1}{1 - r(T_1, Z_2)^2} \right) \tag{2.9}$$

where $r(T_1, Z_2)$ is the Pearson correlation coefficient between the representations T_1 and Z_2. Lastly, we also consider approximating the mutual information by the cosine similarity between the representations T_1 and Z_2, which directly measures the similarity between text and image features. The very ad-hoc estimator is given by

$$\hat{I}(T_1, Z_2) = \frac{T_1 \cdot Z_2}{||T_1|| \cdot ||Z_2||}. \tag{2.10}$$

The qualitative and quantitative results are shown in Figure 4 and Table 3. Perhaps surprisingly, these three estimators yield similar results.

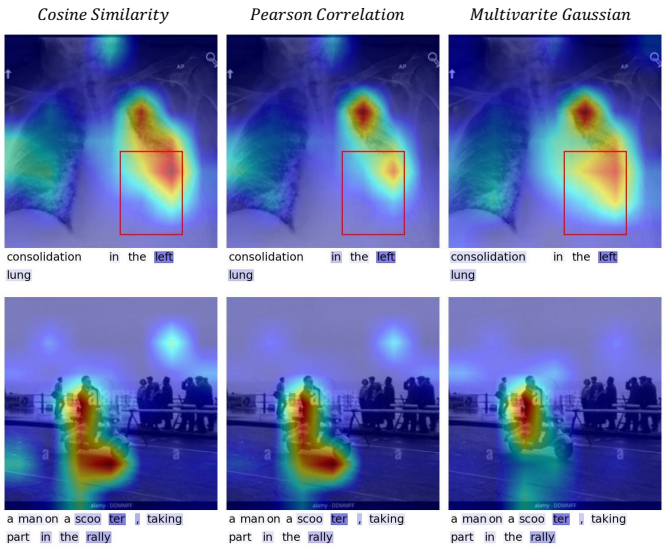


Figure 4: Impact of different estimators for the fitting term.

Table 3: Comparison of performance of different estimators.

		Cosine Similarity	**Pearson Correlation**	**Multivariate Gaussian**
CC image	% Conf. Drop ↓	0.85 ± 0.05	0.88 ± 0.14	1.11 ± 0.14
	% Conf. Incr. ↑	37.80 ± 3.66	41.6 ± 5.54	35.60 ± 3.61
	% ROAD Comb. ↑	1.78 ± 0.05	1.73 ± 0.13	1.49 ± 0.14
	% ROAR+ ↑	-2.70 ± 21.32	27.97 ± 48.03	32.90 ± 32.93
CC text	% Conf. Drop ↓	2.84 ± 0.26	2.9 ± 0.3	3.47 ± 0.32
	% Conf. Incr ↑	20.00 ± 2.00	19.8 ± 4.87	14.6 ± 3.01
	% ROAR+ ↑	25.68 ± 62.94	6.98 ± 22.32	33.35 ± 56.79
MSCXR image	% Conf. Drop ↓	0.45 ± 0.04	0.51 ± 0.09	0.51 ± 0.05
	% Conf. Incr. ↑	48.8 ± 2.64	43.0 ± 3.1	43.80 ± 3.87
	% ROAD Comb. ↑	0.81 ± 0.06	0.8 ± 0.05	0.78 ± 0.10
	% ROAR+ ↑	36.95 ± 6.00	34.40 ± 6.93	50.54 ± 4.12
	% Localization ↑	17.65 ± 1.50	18.22 ± 1.52	17.60 ± 0.83
MSCXR text	% Conf. Drop ↓	4.19 ± 0.07	4.78 ±0.21	4.37 ± 0.12
	% Conf. Incr. ↑	6.20 ± 2.14	5.40 ± 2.24	5.60 ± 0.80
	% ROAR+ ↑	18.05 ± 18.20	4.44 ± 1.88	15.42 ± 10.43

3. Attribution Maps for Additional Examples

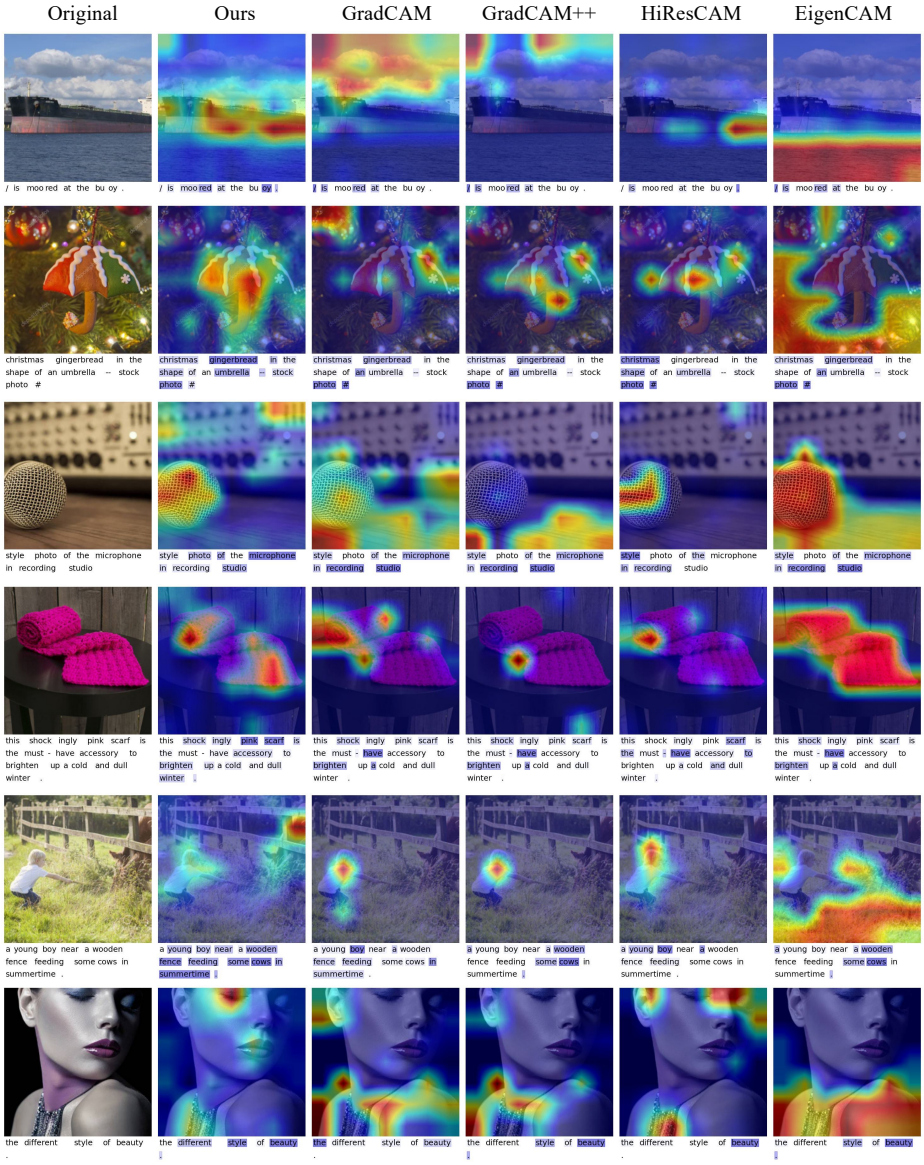


Figure 5: Attribution maps for randomly picked examples from the Conceptual Captions dataset

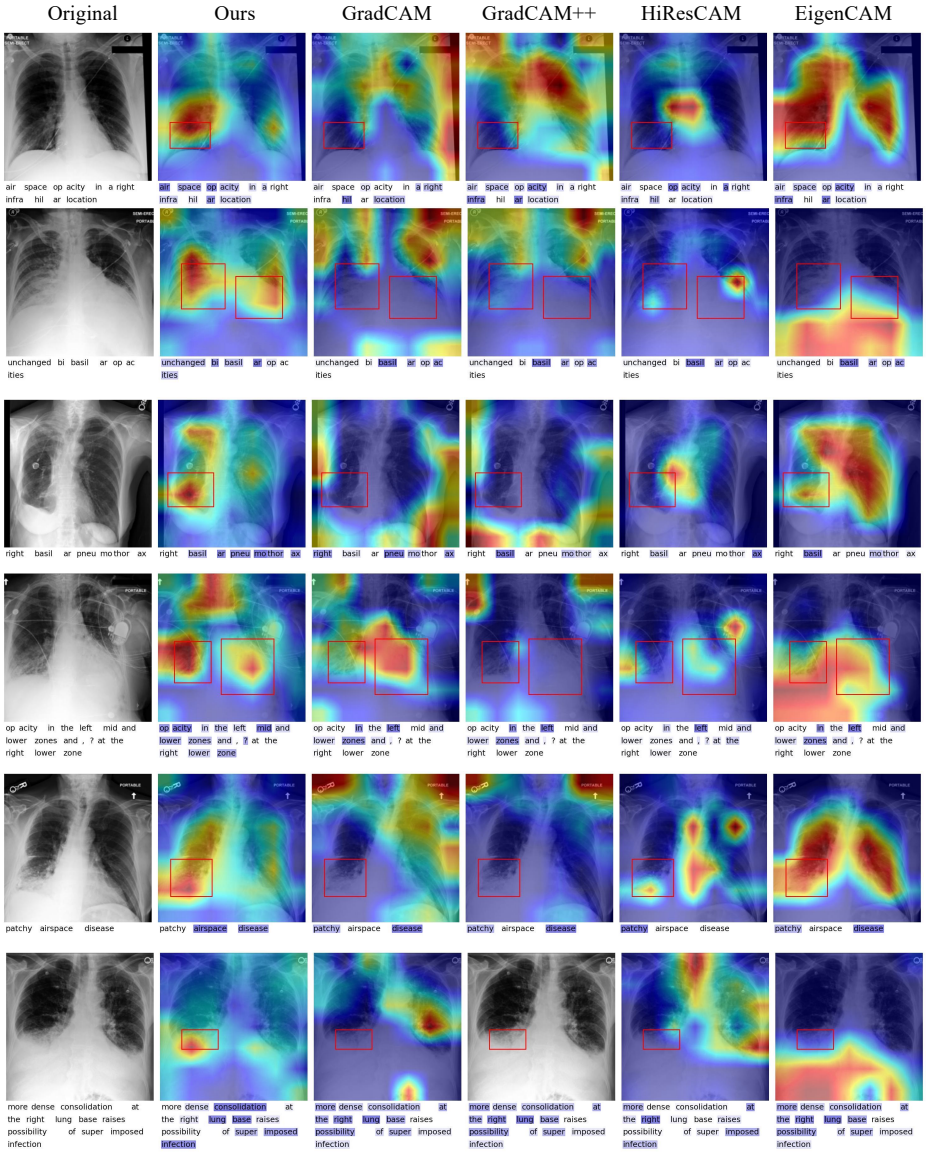


Figure 6: Attribution maps for randomly picked examples from the MS-CXR chest x-ray dataset.

4. Quantiative Metrics

4.1. Localization Test

We quantitatively measure the effectiveness of our proposed attribution method by evaluating its accuracy in zero-shot detection for images. We binarize the saliency map such that the area with scores higher than the threshold (75%) is assigned 1 while the rest is assigned 0. We denote the resulting binary map as S_{pred}. We also construct a ground-truth binary map, S_{gt}, using the bounding boxes provided by MS-CXR (Boecking et al., 2022), where the region inside the bounding boxes is assigned to 1 while the outside is assigned to 0. Note that some samples have multiple bounding boxes and we should consider all of them to test the method's multi-occurrence detection ability. Then, we calculate the IoU (Intersection over Union) of S_{pred} and S_{gt}. Namely, for images with a height of n and a width of m, the score is calculated by

$$\text{Localization} = \frac{\sum_{i=1}^{n}\sum_{i=1}^{m} \mathbb{1}_{S^{ij}_{\text{pred}} \wedge S^{ij}_{gt}}}{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{S^{ij}_{\text{pred}} \parallel S^{ij}_{gt}}} \tag{4.11}$$

where $\mathbb{1}$ is the indicator function, $\wedge$ is the logical and $\parallel$ is the logical or.

As a result, our method obtains 17.60% average IoU for this zero-shot detection task, which is significantly larger than CAM-based methods in comparison (Table 1). Despite outperforming other models, the score is still lower than expected, which is probably caused by the following two factors. **(i)** Our method indeed generates segmentation instead of bounding boxes, so evaluation by bounding boxes might underestimate the quality of the saliency map. **(ii)** The model under evaluation is CXR-RePaiR (Endo et al., 2021), which is a chest x-ray report generation model that is trained on MIMIC-CXR (Johnson et al., 2019). Since medical diagnosis is a challenging task and CXR-RePaiR is not finetuned for detection, the learned image-text representation might be less useful.

4.2. Degradation Test

Although the localization test suggests the potential of the attribution method as zero-shot detection and segmentation tool, it might underestimate the accuracy of attribution methods. Even a perfect attribution method can produce a low localization score because the model under evaluation is poor at extracting useful information, which is very likely for challenging tasks like chest X-ray classification.

Therefore, we use the following three evaluation metrics to further compare our method with baselines. The idea is that removing features with high attribution scores should decrease the performance while discarding features with low attribution scores can improve the performance (because noisy information is ignored). We randomly sample 100 pairs from each dataset and run five experiments for each metric (Table 1). The first two metrics are implemented by `pytorch-gradcam`.[4]

4. GradCAM and its variants are usually applied to image classification and use the softmax outputs of each class as confidence scores. Since our setting does not contain any labels, we use cosine similarity with the other modality instead. The idea is similar to `pytorch-gradcam`'s official tutorial where cosine similarity is used as targets in attribution for image embedding (`https://github.com/jacobgil/pytorch-grad-cam/blob/master/tutorials/Pixel%20Attribution%20for%20embeddings.ipynb`).

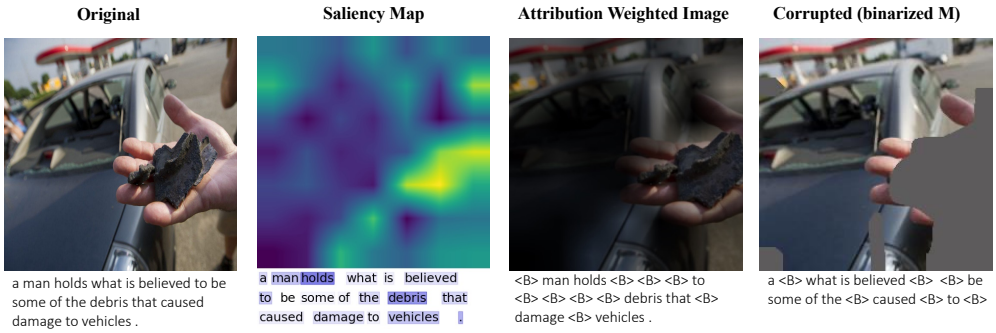


Figure 7: Visualization of saliency map and degradation. The third column is obtained by calculating the element-wise product of the original image and saliency map, while the text with attribution scores lower than 50% percentile is masked by a blank token ¡B¿. It is used in the *Increase in Confidence* metric and its inverse is used in *Drop in Confidence*. The fourth column is an example of the training data in ROAR+. We replace the image pixels with attribution scores higher than 75 % percentile by the channel mean and replace the text tokens with attribution scores higher than 50 % by a blank token ¡B¿. The results in Table 1 use a space as the blank token.

Drop in Confidence (Chattopadhay et al., 2018). An ideal attribution method should only assign high scores to important features, thus we should not observe a drop in performance if only the high-attribution parts are allowed in the input. For images, we use point-wise multiplication of the saliency map and the image input. Since scaling token ids is meaningless, we use binarization similar to Wang et al. (2020) where only tokens with attribution scores in the top 50% are kept. Formally, we define this score by

$$\text{Confidence Drop} = \frac{1}{N} \sum \max(0, o_i - s_i) \tag{4.12}$$

where o_i is the cosine similarity of features of original images and texts, and s_i is the new cosine similarity when one modality is distilled according to the attribution. The lower this metric is, the better the attribution method is.

Increase in Confidence (Chattopadhay et al., 2018). Similarly, removing noisy information in the input might increase the model's confidence. We compute

$$\text{Confidence Increase} = \frac{1}{N} \sum \mathbb{1}(o_i < s_i) \tag{4.13}$$

where $\mathbb{1}$ is the indicator function and the definition of o_i and s_i is the same as above. A higher value indicates better performance.

Remove and Debias (ROAD; Rong et al. (2022)). This method replaces pixels with the average of their neighbors and has two metrics: LoRF (Least Relevant First) representing the target score when removing the least relevant features, and MoRF (Most Relevant First) representing the target score when most relevant features are removed. Similar to the above, we use cosine similarity as the target score. We use a combined score implemented by `pytorch-gradcam`, which calculates (LoRF(t) - MoRF(t))/2, across

thresholds $t = [20, 40, 60, 80]$. Since the average of token ids is meaningless, we only use this metric for images.

Remove and Retrain + (ROAR+, extending the original ROAR; Hooker et al. (2019)). We finetune the base model on the degraded images and texts where the most important parts are replaced by the uninformative values (channel mean of images or spaces for text) and evaluate on a validation set of original inputs.[5] If the attribution method is accurate, a sharp decrease in performance is expected because all useful features are removed and the model cannot learn anything relevant from the degraded data. We sample 500 image-text pairs from each dataset with 80% for training and 20% for validation. We use the same contrastive loss as CLIP uses in pretraining, and define the score by $(l_c - l_o)/l_o$ where l_o is the validation losses of retraining using original data and l_c is that with corrupted data.

5. Sanity Check

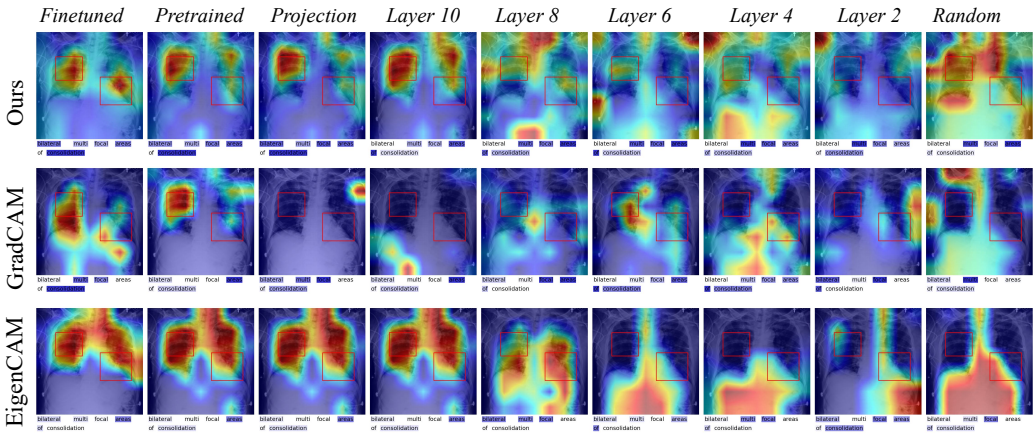


Figure 8: Saliency maps for sanity checks. For our method, we insert M2IB at layer 9. "Finetuned" represents the model that is finetuned on MIMIC-CXR (Johnson et al., 2019), a Chest X-ray dataset. "Pretrained" represents pretrained CLIP (Radford et al., 2021) from OpenAI. "Projection" represents the final layer of the image encoder and test encoder of CLIP that projects image and text features into a shared embedding space. The rest columns represent models with weights randomized starting from the last to the first layer, where "random" means that all parameters in the model are randomly initiated. The outcomes of our method indicate that the saliency maps are sensitive to model weights and thus our method passes the sanity check. However, the output of EigenCAM does not vary much for the last few layers when weights are randomized.

We conduct a sanity check on our method to ensure our method will produce different results if the model parameters changes. We follow the sanity check proposed by Adebayo et al. (2018) where parameters in the model are randomized starting from the last to the first layer. As shown in Figure 8, our method passes the sanity check as the attribution scores of image pixels and text tokens change as the model weights change. Our method also

5. The original ROAR also corrupts the validation data. However, different methods will have different training and validation data under this setting, which makes it hard to compare. We use the same original data as the validation data for all methods instead.

produces more accurate saliency maps for finetuned models compared to pretrained models, which further confirms that the resulting attribution can successfully reflect the quality of the model. Since we restrict the information to a selected layer, the randomization of the previous layer appears to have a larger influence on the output.

6. Error Analysis and Limitations

We notice that our proposed attribution method generally performs well on text, but sometimes shows less satisfying performance on images. By inspecting the qualitative examples, we observe that the proposed method sometimes fails to detect the entire relevant regions in images. As shown in Figure 9 and the fourth ("scarf" example) and sixth ("beauty" example) rows in Figure 5, our method only highlights a fraction of the object in the image though it should include the whole foreground. This kind of error is particularly common in the ROCO dataset where the caption usually refers to the major objects in the image. This is probably because the model under evaluation only relies on a few patterns in the image to make its prediction. Increasing the relative importance of the fitting term (i.e. using larger β) helps to enlarge the highlighted area. However, we don't suggest using too large β because it will break the balance between fitting term and compression and thus make information bottleneck unable to squeeze information.

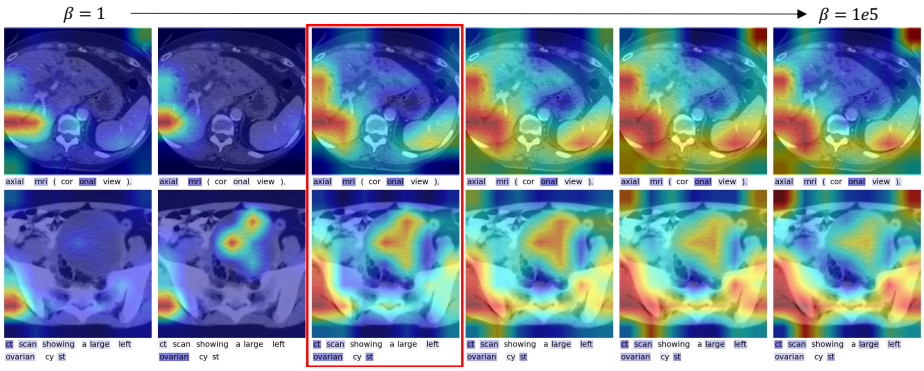


Figure 9: Attribution maps for MRI and CT examples from the ROCO dataset. The red box indicates the visualization generated by the selected hyperparameters according to our hyperparameter search in 2.1. Irrelevant background is included when β is too large.

We also note that the most significant limitation of M2IB is the sensitivity to hyperparameters. As discussed in 2.1, different combinations of hyperparameters will generate different saliency maps. We show how to use the ROAR+ score to systematically select the optimal hyperparameters and also provide visualization to illustrate the effect of different hyperparameters. Since there is no convention on evaluating the attribution method, we suggest taking into consideration of various evaluation metrics, visualization of examples, and the goal of the attribution task when choosing hyperparameters. We emphasize our method should be used with caution since attributing the success or failure of a model solely

to a set of features can be overly simplistic and different attribution methods might have very different results.