5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–13

Understanding Image–Text Representations via Multi-Modal Information Bottleneck Attribution

Anonymous Authors
Anonymous Institution

Abstract

Information theory has become an important tool in understanding and interpreting deep neural networks. For example, information bottleneck theory has been applied to computer vision and natural language processing to improve model interpretability. However, despite the success of vision–language models, information bottleneck theory has not been applied to improve the interpretability of models used for multi-modal data, where interpretability requires understanding all modalities. In this paper, we propose a multi-modal information bottleneck (M2IB) objective for vision–language models that compresses irrelevant and noisy information while preserving crucial visual and textual features that are relevant to the respective other modality. We demonstrate how M2IB can be applied to attribution analysis of multimodal models and shed light on the decision-making mechanism of vision-language pretrained models, thereby improving the interpretability of such models when applied to safety-critical domains, such as medical diagnosis. Using CLIP as an example, we demonstrate the effectiveness of M2IB and show that it outperforms gradient-based attribution methods both qualitatively and quantitatively.

1. Introduction

Vision-Language Pretrained Models (VL-PMs) have gained significant attention due to their applicability to a wide range of vision–language tasks. These models are pretrained on extensive image–text datasets to enable robust and representative encoding of images and texts (Du et al., 2022). For example, the widely used CLIP model trains an image encoder and a text encoder on 400 million image–text pairs and has demonstrated remarkable performance on various vision–language tasks, such as Visual Question Answering and Visual Entailment (Radford et al., 2021; Shen et al., 2022).

Given the substantial computational costs of pretraining, common approaches to multi-model model construction involves finetuning VL-PMs and extracting image–text representations from pretrained VL-PMs and training smaller models on these features (Shen et al., 2022). The robustness and effectiveness of these representations are vital for learning good predictive models for downstream tasks. Moreover, improved understanding and interpretation of state-of-the-art VL-PMs is desirable for important real-world applications like medical diagnosis and image captioning for the visually impaired.

This paper introduces 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). This method extends previous work on information bottleneck attribution (IBA; Schulz et al. (2020b)) from uni-modal to multi-modal data, improving the interpretability of VL-PMs. Our key 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.
- Our proposed method outperforms existing gradient-based attribution methods—GradCam, HiResCam, XGradCam, and EigenCam—quantitatively and qualitatively.

© A. Authors.

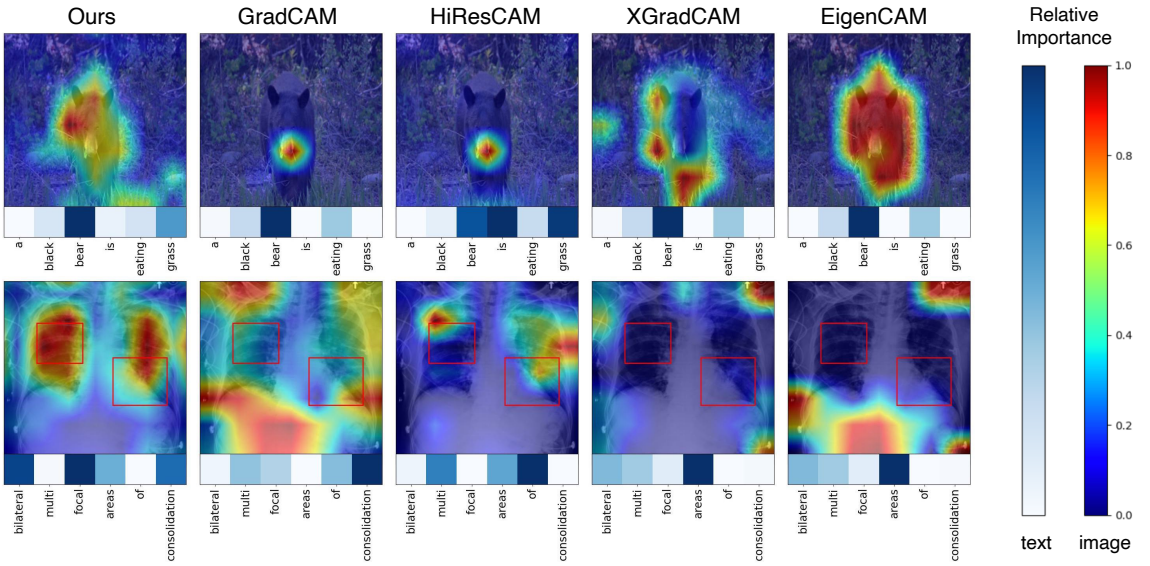

Figure 1: Example attribution maps on image and text. The red boxes on the lower image are the ground-truth bounding boxes associated with the text. Our method successfully identifies key objects in the given image and text ("bear" and "grass" in image and text of the upper figure, and both lungs and relevant words about bilateral lung consolidation in the lower figure), while other methods fail to extract all relevant information. Attribution in the images of the chest X-ray is less precise than attribution in common images since the model is not finetuned for object detection in chest X-rays.

2. Related Work

Gradient-Based Attribution. Gradient-based methods assign higher attribution scores to features with larger gradients. Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2016) is originally proposed for CNN on images, where average pixel-wise gradients are used to weight the activation. Inspired by this approach, similar gradient-based methods are then proposed to improve the performance. HiRes-CAM (High-Resolution Class Activation Mapping) (Draelos and Carin, 2021) adopts pixel-wise multiplication for gradients and activations, aiming for more precise localization of the relevant image regions. XGrad-CAM (CrossGrad-CAM) (Fu et al., 2020) uses normalized activations. EigenCAM (Eigen Class Activation Mapping) (Muhammad and Yeasin, 2020) utilizes the principle components since gradients are usually noisy.

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-50He et al. (2016)). Subsequently, IBA is applied to language transformers and also outperforms other methods on this tasks (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.

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

Information bottleneck (Tishby and Zaslavsky, 2015) provides a rich framework for understanding the inner mechanism of deep neural networks and has been widely applied in interpretable machine learning. To make latent representation extract relevant information from input, it suggests two competing objectives during compression (minimize the mutual information between latent representation Z 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, we cannot simply apply the traditional information bottleneck as above because the output is unclear and the definition of relevancy is insufficient. To obtain an image–text corpus that is as large as possible, we usually only have (text, image) pairs and conduct self-supervised learning with a contrastive loss function (Baevski et al., 2022; Mu et al., 2021). This approach differs from supervised 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 image–text representation learning, however, the label is usually converted into a sentence such as "This is a picture of a bear" (L'_{bear}) (Radford et al., 2021). In this setting, both V_{bear} and L'_{bear} are inputs, and the output label is no longer valid. Another common dataset is an image–caption dataset where we none of the data points are labeled. In fact, we would like to make the outputs "undefined" because we want a task-agnostic image–text representation, whereas the outputs are closely dependent on the specific downstream task. As a result, how to define the fitting term $I(Z, Y)$ is unclear.

Fortunately, there is a natural proxy for the relevance of information in multi-modal data. If the image and text 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:

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

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

Note that the second terms (excluding β) in Equations 2 and 3 are the same because mutual information is symmetric.

3.2. Information Bottleneck for Attribution

To obtain an attribution map of image and text, we modify the information bottleneck attribution (IBA; Schulz et al. (2020a)) method for usage in multi-modal settings. To restrict the information flow in the selected layer, noise is added to the layer output Z to obtain a masked output T,

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

where $\odot$ is the Hadamard product and $\mathbf{1}$ is an all-one matrix having the same dimension as Z and λ. In the 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 is purely noise. To ensure the masked output has the same magnitude as the original feature, we let $\epsilon \sim N(\mu_Z, \sigma_Z^2)$. We learn λ for image and text representations, respectively, by minimizing

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

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

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 $T|Z = z$ 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 $P(T)$ as $q_T(t) = \mathcal{N}(t; \mu_Z, \sigma_Z)$, 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 proof 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, 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.

For VL-PMs like CLIP, its image–text representation is jointly trained and is aligned in one embedding space. To approximate the mutual information between T and Z for different modalities, we calculate the cosine similarities of the image embedding and text embedding as a crude but effective approximation. Denoting the original text and image embeddings by $\mathbf{e_t}$ and $\mathbf{e_i}$, respectively, and the embedding after inserting the information bottleneck by $\mathbf{e'_t}$ and $\mathbf{e'_i}$, respectively, we compute

$$\hat{I}(T_{\text{image}}, Z_{\text{text}}) = s(\mathbf{e'_i}, \mathbf{e_t}) = \frac{\mathbf{e'_i} \cdot \mathbf{e_t}}{||\mathbf{e'_i}|| \cdot ||\mathbf{e_t}||} \quad \text{and} \quad \hat{I}(Z_{\text{image}}, T_{\text{text}}) = s(\mathbf{e_i}, \mathbf{e'_t}) = \frac{\mathbf{e_i} \cdot \mathbf{e'_t}}{||\mathbf{e_i}|| \cdot ||\mathbf{e'_t}||}. \tag{8}$$

Combining Equations 7 and 8, we obtain the loss function

$$\mathcal{L} = \mathbb{E}[\mathbb{D}_{\text{KL}}(p_{T|Z} \parallel q_T)] - \beta s(T, Z). \tag{9}$$

For image and text, we thus have the objective functions

$$\mathcal{L}_{\text{image}}(\lambda) = \mathbb{E}[\mathbb{D}_{\text{KL}}(P(T_{\text{image}}|Z_{\text{image}}) \parallel Q(T_{\text{image}}))] - \beta_{\text{image}} s(T_{\text{image}}, Z_{\text{text}})) \tag{10}$$

$$\mathcal{L}_{\text{text}}(\lambda) = \beta_{\text{text}} \mathbb{E}[\mathbb{D}_{\text{KL}}(P(T_{\text{text}}|Z_{\text{text}}) \parallel Q(T_{\text{text}}))] - \beta_{\text{text}} s(T_{\text{text}}, Z_{\text{image}})). \tag{11}$$

4. Empirical Evaluation

We evaluate the proposed attribution method using CLIP (Radford et al., 2021) on two image–caption datasets: **(i)** Conceptual Captions (Sharma et al., 2018), a dataset consisting of diverse images and captions from the web, and **(ii)** MS-CXR (Boecking et al., 2022), which contains chest X-rays, texts describing radiological findings and bounding boxes.

4.1. Experiment Setup

For all experiments, we use HuggingFace's pretrained CLIP models with ViT-B/32 (Dosovitskiy et al., 2021) as the image encoder. For Conceptual Captions, we load weights of `openai/clip-vit-base-patch32`.[1] Since MS-CXR is a domain-specific dataset while CLIP is not exposed to this kind of data during pretraining, we use a finetuned version of CLIP, CLIPMD[2], that is trained on the radiology dataset ROCO (Radiology Objects in COntext; Pelka et al. (2018)) containing radiology image-caption pairs from the open-access biomedical literature database PubMed Central. For each {image, caption} pair, we insert an information bottleneck into the given layer m of the text encoder and image encoder of CLIP separately, then train the bottleneck for 10 epochs using the Adam optimizer and 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 and the weight β, as shown in Figure 2. For baseline methods, 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 2 (also in Appendix 2).

4.2. Evaluation Metrics

We use four evaluation metrics to compare our method with baselines. We sample 100 pairs from each dataset and run five experiments for each metric (Table 1). The first three metrics are implemented by `pytorch-gradcam`.[4] **(1)** Drop in Confidence: An ideal attribution map 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. Formally, we define this score by $\frac{1}{N}\sum \max(0, o_i - s_i)$ where o_i is the cosine similarity of the original image and text and s_i is the cosine similarity when one modality is distilled according to the attribution map. **(2)** Increase in Confidence: Similarly, removing noisy information in the input might increase the model's confidence. We compute $\frac{1}{N}\sum \mathbb{1}(o_i < s_i)$ where $\mathbb{1}$ is the indicator function. **(3)** Remove and Debias (ROAD; Rong et al. (2022)): This method replaces pixels by the average of their neighbors. We use a combined score implemented by `pytorch-gradcam`, which calculates (Least Relevant First - Most Relevant First)/2, across thresholds $[20, 40, 60, 80]$. Since the average of token ids is meaningless, we only use this

1. https://huggingface.co/openai/clip-vit-base-patch32
2. https://huggingface.co/Idan0405/ClipMD
3. https://github.com/jacobgil/pytorch-grad-cam
4. GradCAM and its variants are usually applied to image classification and use the softmax outputs of each class as confidence score. Since our setting doesn't 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).

	Methods	GradCAM	HiResCAM	XGradCAM	EigenCAM	Ours
CC image	% Conf. Drop ↓	4.00 ± 0.51	3.88 ± 0.20	3.08 ± 0.19	3.59 ± 0.23	**1.11** ± 0.05
	% Conf. Incr. ↑	20.60 ± 1.95	16.40 ± 3.36	20.00 ± 4.47	16.60 ± 2.97	**41.20** ± 5.26
	% ROAD ↑	1.03 ± 0.14	0.94 ± 0.10	0.31 ± 0.13	0.07 ± 0.12	**1.53** ± 0.03
	% ROAR+ ↑	14.51 ± 6.07	5.07 ± 6.93	6.50 ± 2.80	6.30 ± 1.88	**28.89** ± 4.44
CC text	% Conf. Drop ↓	9.29 ± 0.30	7.98 ± 0.44	9.18 ± 0.56	8.92 ± 0.53	**5.36** ± 0.37
	% Conf. Incr ↑	3.40 ± 1.14	2.60 ± 1.52	3.00 ± 1.00	2.40 ± 1.67	**5.80** ± 2.68
MSCXR image	% Conf. Drop ↓	5.40 ± 1.40	24.64 ± 1.22	19.37 ± 1.69	20.73 ± 0.79	**1.21** ± 0.23
	% Conf. Incr. ↑	54.60 ± 3.36	12.40 ± 2.30	12.20 ± 2.17	17.60 ± 1.95	**69.40** ± 4.39
	% ROAD ↑	1.79 ± 0.91	3.71 ± 0.28	-0.38 ± 0.92	-1.65 ± 0.45	**9.57** ± 0.13
	% ROAR+ ↑	33.59 ± 5.16	59.78 ± 4.01	84.81 ± 3.16	140.78 ± 21.36	**202.87** ± 8.97
MSCXR text	% Conf. Drop ↓	26.96 ± 1.28	26.80 ± 1.32	28.22 ± 1.03	**21.46** ± 1.44	26.59 ± 0.69
	% Conf. Incr. ↑	7.20 ± 2.59	**7.80** ± 3.27	4.60 ± 2.88	7.60 ± 3.44	7.20 ± 2.39

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

metric for images. **(4)** 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 for images and spaces for texts), 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. Results & Discussion

We compared M2IB to GradCAM, HiResCAM, XGradCAM, EigenCAM, and find that it is generally better at capturing the shared information in images and texts. As shown in Figure 1, GradCAM, HiResCAM, and EigenCAM successfully identify the key object "bear" mentioned both in the image and text but fail to highlight "grass". More examples are provided in Figures 4 and 5. M2IB also outperforms existing models in almost all numerical evaluation metrics (Table 1). Removing pixels or tokens with lower attribution scores according to our method generally increases the cosine similarity with the other modality while masking by our attribution map generally decreases the relevance with the other modality and makes the model perform worse after retraining. This confirms that our method generates correct attribution maps.

We provided evidence that M2IB is able to shed light on the decision-making mechanism of VL-PMs and improve the interpretability of complex machine learning models. We hope that this work paves the way for further work on multi-modal information-theoretic attribution methods that can help introduce modern machine learning methods into safety-critical domains where interpretability is critical.

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.

References

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.

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.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models, 2022. URL https://arxiv.org/abs/2202.10936.

Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns, 2020.

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.

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.

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.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and C. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *CVII-STENT/LABELS@MICCAI*, 2018.

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.

Supplementary Material

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

2.1. Hyperparameters

β controls the relative importance of the fitting term. The larger the β is, the more information is allowed to flow through this layer. Note that cosine similarity is bounded between 0 and 1, while KL divergence doesn't have an upper bound. Thus, β needs to be sufficiently large to make the fitting term at the same scale as the compression term. As shown in the upper part of Figure 2, 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. The 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. We also observe that the attribution of texts is usually more stable than images, which is probably because texts convey less information than images. These hyperparameters can be chosen according to numerical metrics mentioned in 4.2 (results in Appendix 2.2). The best β is 1e4 and the best layer index is 9 for both datasets, which is consistent with the findings in Figure 2.

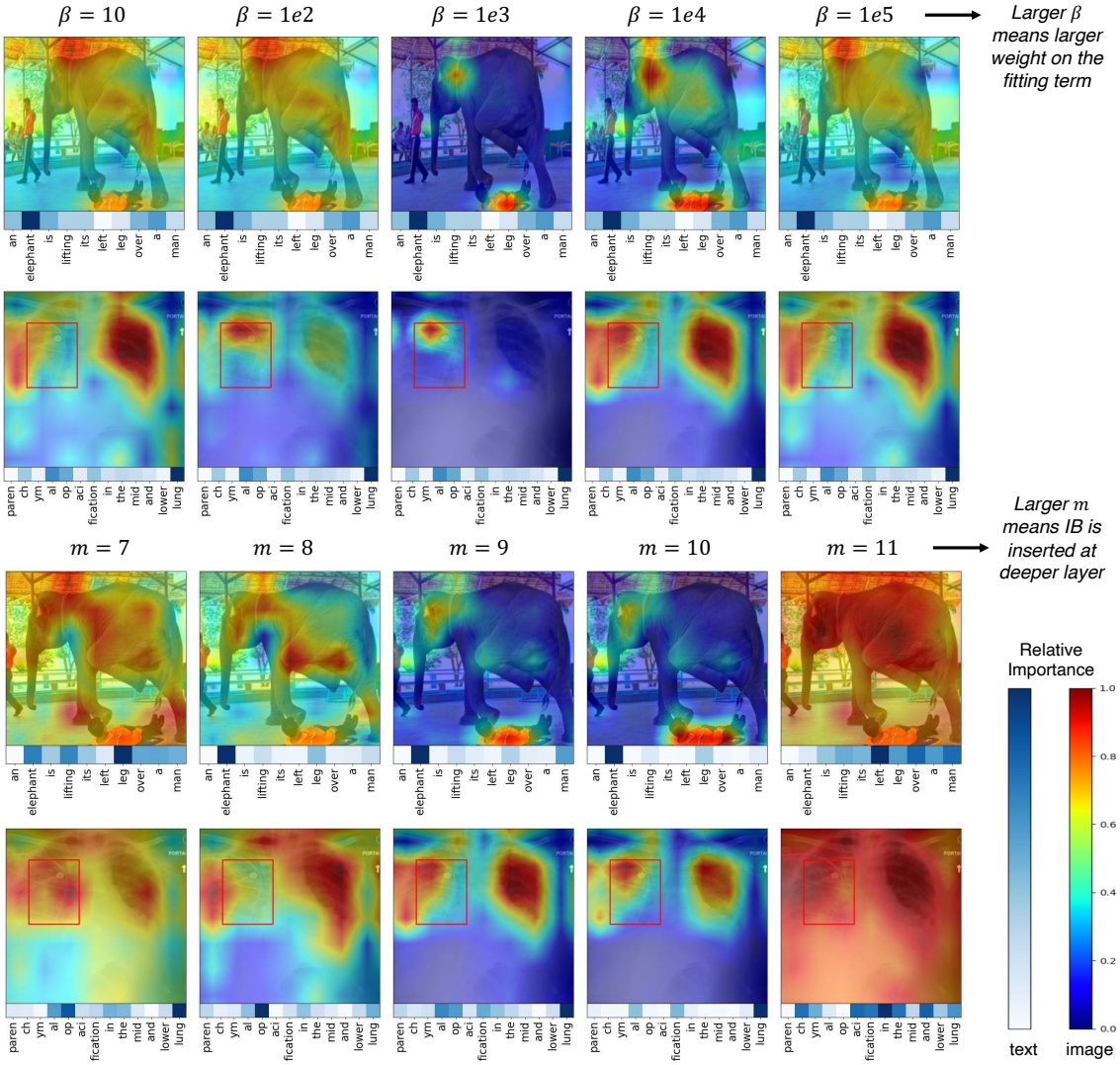

Figure 2: Visualization of the impact of different hyperparameters. β that makes fitting term and compression at a similar scale, and layer locates in the middle of the second half of the encoder, usually gives better performance. Note that the attribution score is assigned to each token, instead of each word, due to tokenization. Thus, subwords are observed for some terminologies in the MS-CXR dataset.

2.2. Hyperparameter Tuning

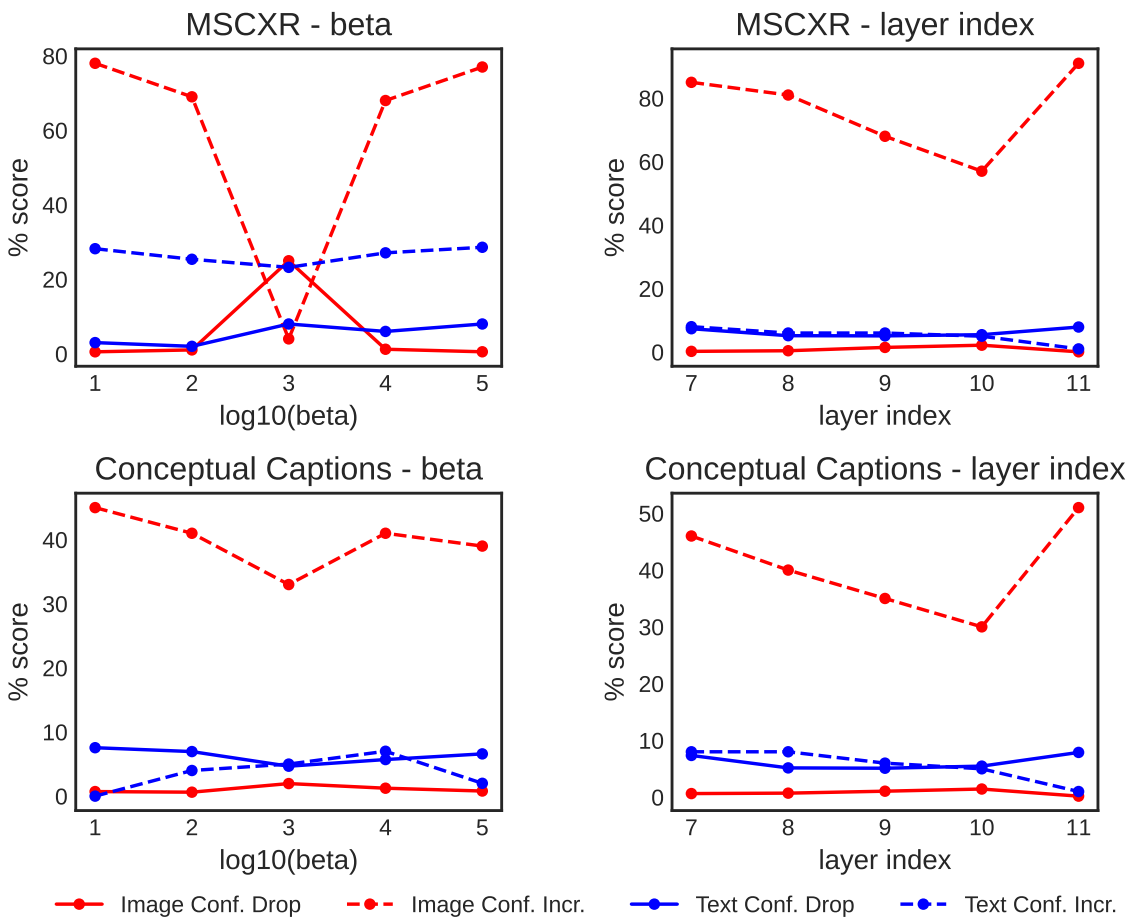

Figure 3: Hyperparameter tuning results. A lower Confidence Drop score (in solid lines) and a higher Confidence Increase score (in dashed lines) indicate better performance. Although we can choose different hyperparameters for image and text because we insert the information bottleneck separately, we use the same hyperparameters for both modalities here to save time. For both datasets, β of 1e4 (or 1e2) and layer index of 9 lead to the best performance, which is consistent with the visualization results in Figure 2. We observe there is a large drop in the Confidence Increase score of images when $\beta = 3$. When inspecting Figure 2, we find that $\beta = 3$ highlights a much smaller area, which will cause the model to remove too much information when evaluating on this metric. However, this behavior might be desirable when we want to focus on a few essential pixels. There is no convention on evaluating attribution method, we suggest taking into consideration of various evaluation metrics and visualization of examples.

3. Additional Examples

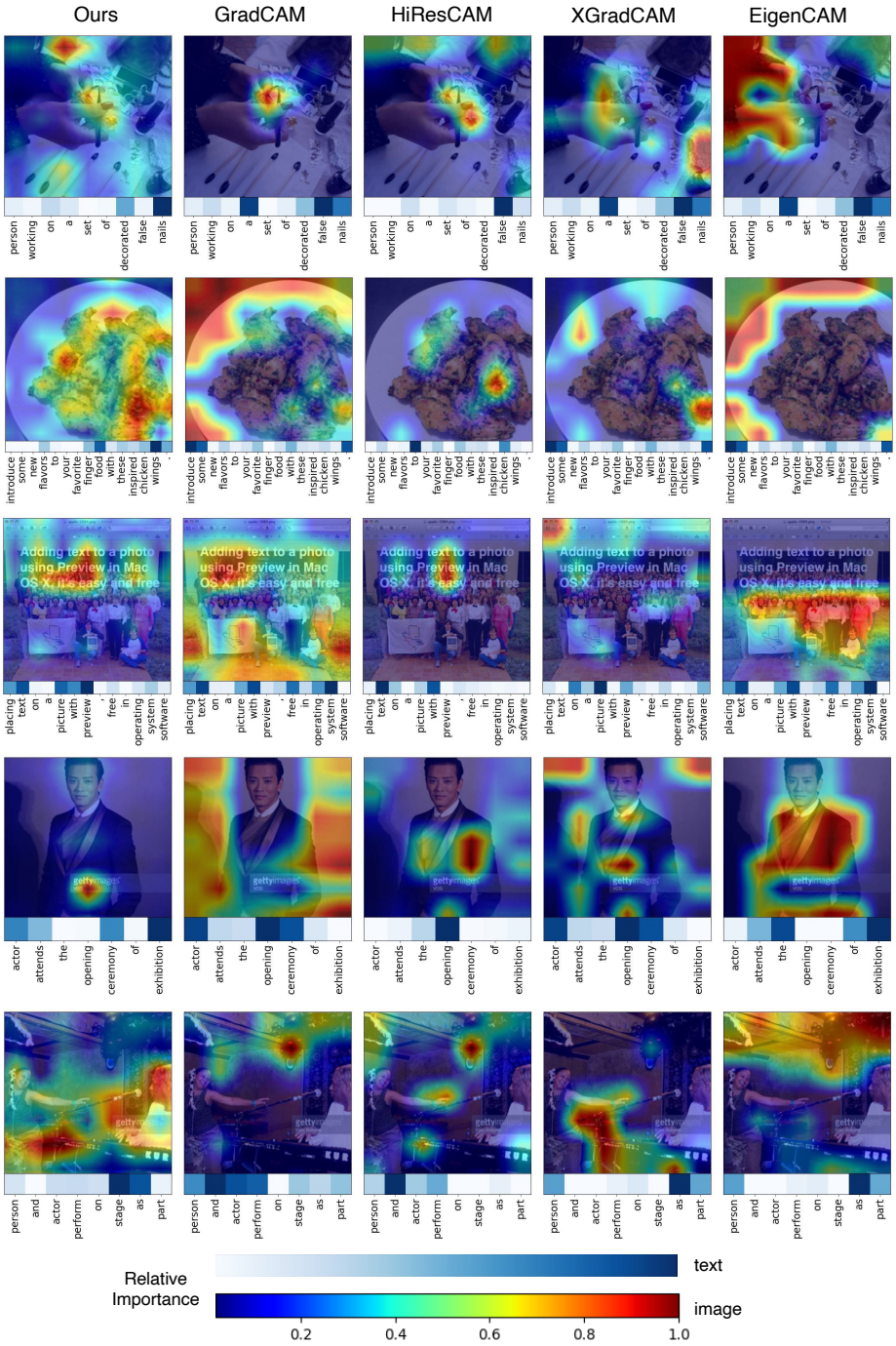

Figure 4: Randomly picked examples from Conceptual Captions

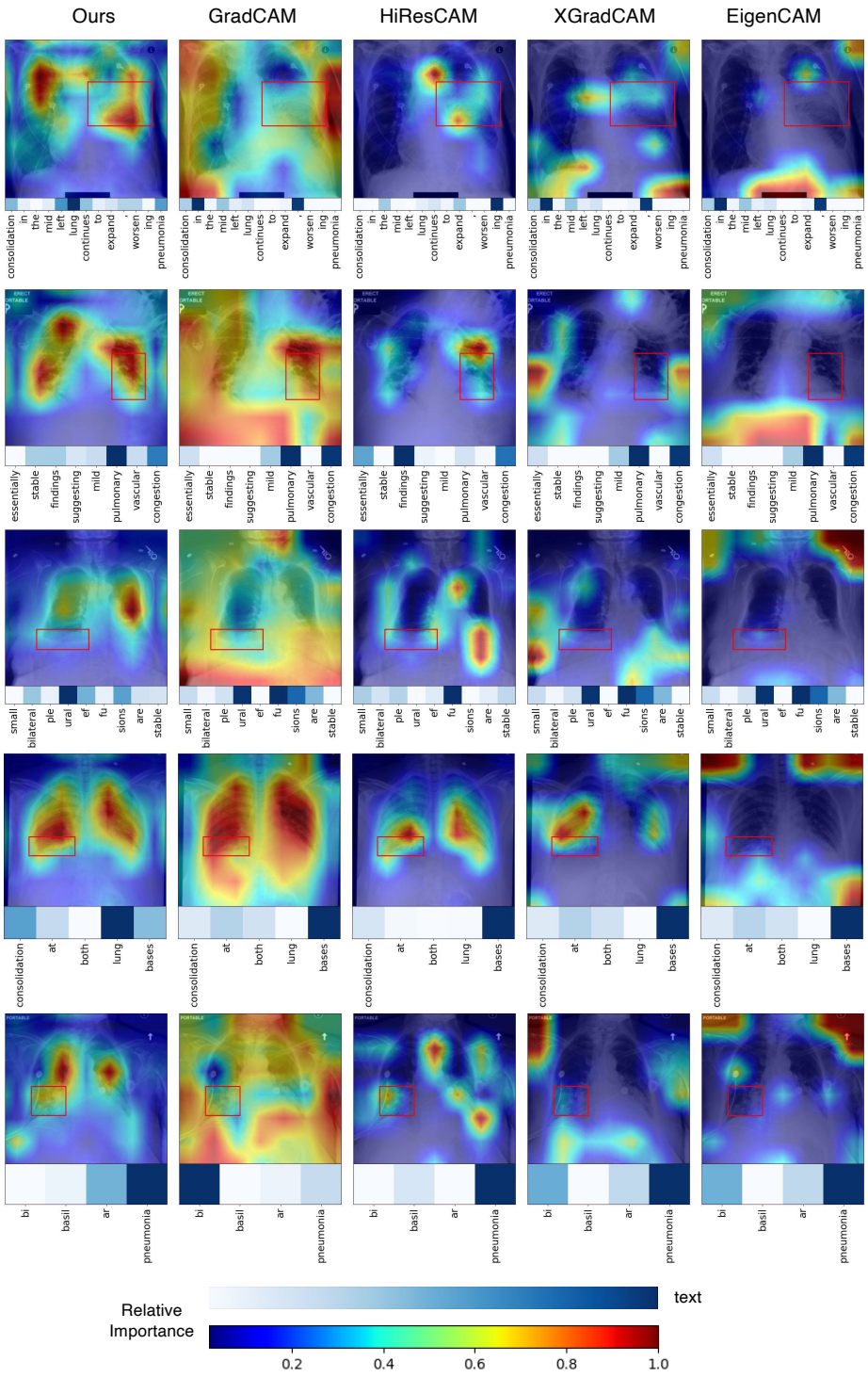

Figure 5: Randomly picked examples from MS-CXR