Towards More Explainable Multimodal Machine Learning

Yuanxin (Michael) Wang [1] **Jiaxin (Kelly) Shi** [1]

Abstract

Deep neural networks have been proven to be both effective and efficient in many tasks, however, the complex model architecture makes them difficult for us to understand their decision-making process. As human behaviors are intrinsically multimodal, multimodal deep neural networks have achieved better performance than unimodal baselines in most cases. With growing applications of multimodal deep neural networks based Artificial Intelligence, the interplay between heterogeneous and high-dimensional modalities makes explainability a key concern of multimodal neural networks, as its application in areas like healthcare requires trustworthy decisions. This paper introduces a gradient-based explanatory method that visualizes the contribution of features to both the cross-modal interaction and classification result of any black-box multimodal neural networks. This paper serves as a final progress report for our advanced multimodal machine learning class.

1. Introduction

Deep neural networks have made numerous breakthroughs in creating applications for different modalities including vision (Dosovitskiy et al., 2021) (He et al., 2021), language (Devlin et al., 2019) (Brown et al., 2020), and speech (Miao et al., 2015) (Rao et al., 2017). However, these complex models are blamed for a lack of trust and transparency given their intrinsic black-box nature. Correspondingly, the idea of explainable AI (XAI) has been adopted to probe the inner mechanisms of these unimodal architectures (Tjoa & Guan, 2021). Among a wide range of XAI approaches, the gradient-based explanatory method has achieved significant progress in understanding how each segment in the input contributes to the prediction results.

The design of explainable machine learning models is expected to be even more challenging when we move from unimodal to multimodal scenarios. Despite the recent increasing popularity in different domains (Baltruaitis et al., 2019), one key challenge in multimodal machine learning is to explain cross-modal interactions. Cross-modal feature extraction allows for richer representations, thus improving accuracy. However, the interplay between heterogeneous and high-dimensional modalities makes explainability a key concern of multimodal neural networks, as its application in areas like healthcare requires trustworthy decisions. Moreover, cross-modal interactions occur at multi-stages during the learning process, which makes it harder for us to measure how much cross-modal interactions are modeled. Specifically, the following questions should be answered to better explain cross-modal interaction: **(1)** what is the contribution of each modality to the prediction? **(2)** which input modality is dominant? **(3)** how to identify important input segments that facilitate the interaction between modalities?

Noticing that the aforementioned gradient-based explanatory method was mostly applied to the unimodal cases and has not been explored in the space of cross-modal interaction, we believe that the marriage between these two fields becomes a natural next step for both multimodal machine learning and XAI communities. In this work, we propose a novel gradient-based paradigm to uncover the extent of cross-modal interaction in a multimodal classification model and what role each input segment (e.g., image patches, language tokens, and audio clips) plays to enable such interaction. During the back-propagation on a trained multimodal model, the gradients flow back from one specific prediction label, through all intermediate representations and layers, and finally towards all input segments from different modalities. By completely or partially perturbing the input modalities, various cross-modal interaction patterns can be observed through the magnitude changes in the gradients from different layers. To the best of our knowledge, this paper is the first to explore the correlation between gradient variations and cross-modal interaction. Our code will be available at github.com/MichaelYxWang/CrossExplain.

2. Related Work

2.1. Explainable AI (XAI)

Artificial Intelligence systems based on deep neural networks are powerful in learning in many domains. Although deep neural networks have been proven to be effective and efficient for both unimodal and multimodal tasks, the complex model architecture and hidden layer processing make them difficult for us to understand their internal states and decision-making process. There are three major types of

XAI methods: intrinsic, distillation and visualization.

Intrinsic methods can be achieved through either attention mechanisms or joint training. Attention mechanisms (Vaswani et al., 2017a) focus on regions that are important for making predictions and can make the model inherently self-explainable. Jointly training the model for prediction and explanation can also make it intrinsically self-explanatory. For example, (Kanehira et al., 2019) jointly trains a classification module and an explanation module for predicting counterfactuality.

Distillation methods can be done through local approximation and model translation. LIME (Ribeiro et al., 2016) uses a simple and interpretable model to approximate the black-box model locally. It is model-agnostic, and it performs perturbations around a particular prediction and sees how the predictions change. The model translation method builds a surrogate model to simulate the original model. The surrogate model is usually inherently explainable. For example, (Kaya et al., 2017) uses decision trees to approximate a black-box model.

Visualization methods can be categorized into two types: backpropagation and perturbation methods. Backpropagation methods compute the feature relevance based on the gradients passed through the network. Some existing backpropagation works are Grad-CAM (Selvaraju et al., 2016), DeepLIFT (Shrikumar et al., 2019), and layer-wise relevance propagation (Binder et al., 2016). The perturbation method alters the input and compares the output with original and modified input to identify sensitive features for prediction (Zeiler & Fergus, 2013).

2.2. Visual Explanation

Understanding what regions contribute most to the final class label prediction is key to explainable computer vision models. Apart from directly observing intermediate convolutional activation maps, gradient-based methods such as Grad-CAM (Selvaraju et al., 2016) and class saliency map visualization (Simonyan et al., 2014), where the gradient of each raw pixel for a specific class label is computed and the resulting heatmap is used to show the relevance of each region. Most gradient-based methods can be universally applied to any Convolutional neural network architecture. Extensions to Grad-CAM including XGradCAM (Fu et al., 2020), EigenGradCAM (Muhammad & Yeasin, 2020), and LayerCAM (Jiang et al., 2021) have continuously increased the dimensions of computer vision explainability as well.

2.3. Textual Explanation

Compared to the direct computation of the raw pixel-level gradient to understand computer vision outputs, explaining word/token importance in natural language processing is a different story. The mismatch between words represented as continuous embedding vectors and word-level importance scores represented as scalar values has motivated the sensitivity analysis (SA-based) method. SA-based methods operate directly on either the raw value (Nguyen, 2018), L1-norm (Li et al., 2016), or L2-norm (Arras et al., 2016) of the gradient vectors for each word and use the variants of these gradient vectors as importance scores. The limitation of the SA-based method is that it can only measure the absolute value of word importance instead of distinguishing between positive and negative effects. Gradient × Input (GI) method (Denil et al., 2014) and its variants (Arras et al., 2019) (Ancona et al., 2018), where the dot product between the word vector and the gradient vector is performed to compute the word importance score, are proposed to solve this issue. In addition, layer-wise relevance propagation, where the local relevance redistribution of each input is set proportionally to its contribution in the forward pass, is applied in natural language processing to tell the differences between words that are used to support or oppose the classification decisions (Arras et al., 2016).

However, researchers have also noticed that gradient-based explanatory methods in natural language processing are not always reliable (Wang et al., 2020): gradients can be manipulated without affecting the prediction of the model in adversarial settings and can lead to completely unreasonable word importance scores (e.g., focusing only on stop words). Va NLP explainability can also be tackled by non-gradient-based methods effectively. Inspired by how convolutional neural networks can be used to represent sentences (Kim, 2014), word importance scores can be also obtained by performing average pooling on each row of the sentence matrix and summing up the pooled vectors across different feature maps for each word (Lee et al., 2018).

With the prevalence of BERT-family models, (Vaswani et al., 2017b) (Devlin et al., 2019), explainable transformer architectures have aroused increasing interest from the research community. In addition to attention head inspection and representation latent space visualization methods (Coenen et al., 2019) (Rogers et al., 2020), direct analysis of per-token local relevance can be achieved by propagating relevance scores based on the deep Taylor Decomposition principle through attention layers and skip connections (Chefer et al., 2021).

2.4. Multimodal Explanation

A growing number of deep neural network-based models have made significant progress on multimodal tasks based on several modalities such as vision, textual data, and language. Several studies have applied explainability methods to determine the relative importance of each modality in making decisions on multimodal datasets. We could categorize them by the stage explanation modeling is introduced

to the deep neural networks: during modeling and post-modeling phase.

During the modeling phase, models are inherently developed to provide explanations by employing the intrinsic method. For generating a multimodal explanation for VQA, (Wu & Mooney, 2018) uses the model-agnostic explainer LIME (Ribeiro et al., 2016) to determine the segmented objects that influenced the decision the most; then it learns to embed the question, answer, and the VQA-attended features to generate textual explanations and measure how well the object referenced in the generated explanation matches the segments highlighted by LIME. There are also hybrid models that joint prediction and explanation. Contextual Explanation Networks (Al-Shedivat et al., 2017) is a class of probabilistic models that learns to predict by generating and leveraging intermediate context-specific explanations. Self-explanatory Neural Networks (Alvarez-Melis & Jaakkola, 2018) consists of three components: a concept encoder that transforms the input into a small set of interpretable basis features, an input-dependent parameterizer that generates relevance scores, and an aggregation function that makes predictions. Its robustness loss on the parameterizer encourages the full model to behave as a linear function locally, which yields an immediate interpretation of both concepts and relevance. For giving explanations in the post-hoc phase, the explanation method is implemented after the model is trained through backpropagation methods (Selvaraju et al., 2016) and proxy models (Kaya et al., 2017). These methods are model-agnostic thus can be applied to any trained models to improve their explainability.

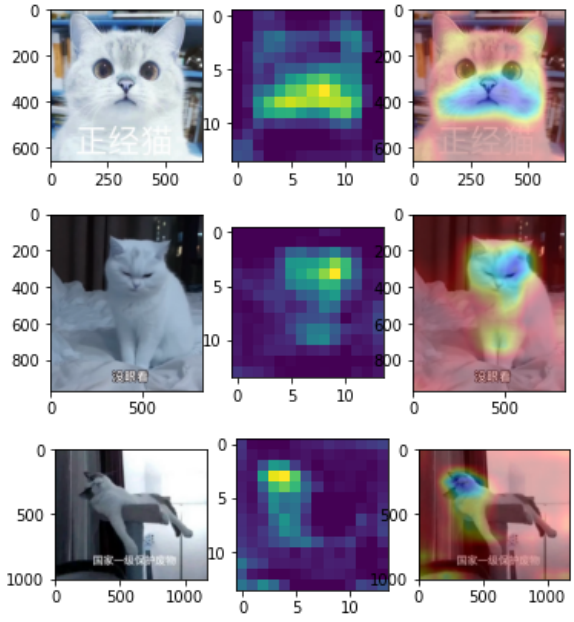


Figure 1. Unimodal results for image classification Grad-CAM visualization

3. Proposed Method

Given a trained multimodal classification model, a post-hoc gradient-based explanation method is developed to investigate which segments of the input contribute significantly to the final class prediction and which segments contributing more to the cross-modal interaction. To identify such segments for each input example, four rounds of forward-backward passes with different input perturbations are needed. For simplicity, our illustrated example uses a multimodal classification model that takes one image and one piece of text as input, and outputs a Softmax probability across various classes.

In the first round, as shown in Figure 4, both text and image modalities are forward passed through the model and the vector containing the gradients for each text token representation is denoted as $text\_grad\_no\_mask$. This gradient vector is expected to identify the tokens (e.g., token 2 and token n in Figure 4) that play a more important role in predicting a specific class label compared to other tokens.

In the second round illustrated in Figure 5, we mask the image completely and only the text modality is forward passed. We denote the same gradient vector for all token representations as $text\_grad\_mask\_all$. We compare the norms between this new vector with $text\_grad\_no\_mask$ to identify whether the model is learning from a dominant modality or from cross-modal interaction. If there is no change in the predicted class and no significant difference between the norms, it implies that the image modality is not contributing and the text modality is dominant; otherwise, there are two possible explanations: **(1)** the image modality is dominant **(2)** cross-modal interaction is necessary for this prediction. Explanation **(1)** can be relatively straightforward to examine by simply masking the text completely and observing the gradient norm changes in the image side. Assuming that we confirm that cross-modal interaction actually happened during the prediction, then intuitively the tokens that receive the largest gradients in this vector (e.g., token n-1 and token n in Figure 5) should be regarded as bimodally-important.

To make this assumption more convincing, the third round of forward-backward pass is performed with the text modality completely masked to get the image gradient matrix denoted as $img\_grad\_mask\_all$ and a fourth round is carried out with only the pre-assumed bimodally-important tokens (e.g., token n-1 and token n in Figure 5) masked to get the image gradient matrix denoted as $img\_grad\_mask\_partial$, which are illustrated in Figure 6 and Figure 7, respectively. If these two matrices are similar to each other, we can conclude that the pre-assumed bimodally-important tokens play such a key role in the bimodal interaction that masking them only takes the same effect as masking the whole text modality.

By repeating a similar process on the image end, we can also find the image regions that are important for the prediction of a specific label as well as cross-modal interaction. Apart from masking input modalities, other modality-specific perturbation methods such as re-ordering the tokens in the text and adding Gaussian noise to the image pixels, can also be integrated into the pipeline.

4. Experimental Setup

4.1. Dataset

For real-world dataset, we use the Flickr 8K Dataset, which consists of 8000 images that are each paired with five different captions describing salient entities and events. We reformulate the task from image captioning to binary classification, where matched image-text pairs are annotated as positive and randomly sampled image-text pairs are annotated as negative (negative sampling).

Given the complex nature of real-world multimodal dataset, it is hard to control and to measure the cross-modal interactions between modalities. To better understand the effectiveness of our methods capturing how much cross-modal interaction the model have learned, we built a synthetic dataset upon the MNIST Dataset.

This synthetic dataset contains 8000 datapoints of 2 modalities: text and image. For the image modality, we simply collage 4 images randomly chosen from the MNIST Dataset. The text modality contains one-sentence descriptions of each collaged image. There are 4 types of descriptions:(1) The position of the largest digit in the image; (2) the position of the smallest digit in the image; (3) the number of even numbers in the image; (4) The number of odd numbers in the image. The label is 1 if the text description matches the image, and 0 otherwise. Figure 2 shows 4 datapoints each demonstrating a task describes above. To have a balanced dataset between two classes(0 and 1), we randomly choose 50% of the dataset and changed the text descriptions to make them mismatch the corresponding images. To analyze how the gradient and how cross-modal interactions change with different levels of text information granularity, we randomly decide if the text description is precise of more general. For example, if the text description in on the largest or the smallest digit in the image, the more precise description gives both vertical(upper or bottom) and horizontal(left or right) positions, and the less precise description gives either vertical or horizontal position. For text description on the number of even or odd digits in the image, the more precise description gives both the count and the position of the first even or odd number while the less precise description only give the count.

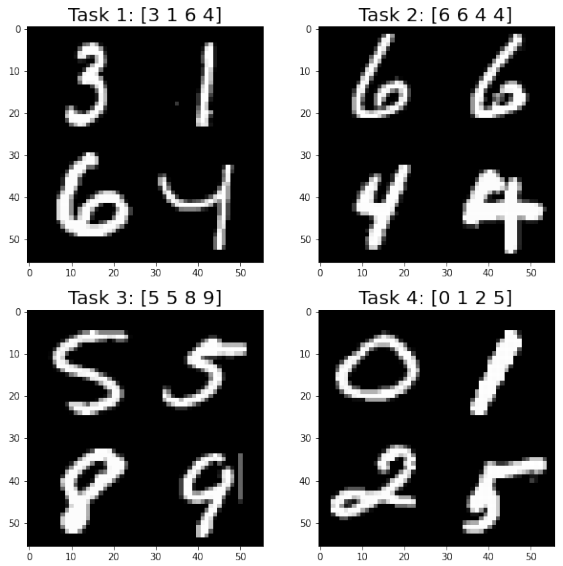


Figure 2. These are demonstrations of 4 tasks. (1) Text: The largest number is at the bottom right corner. Label: 0; (2) Text: The smallest number is on the upper half. Label: 1; (3) Text: There are 1 even number(s) in the image. Label: 1; (4) Text: There is no odd number in the image. Label: 0

4.2. Experiment Models

For the synthetic dataset, we start with a very simple CNN-LSTM based early fusion architecture, where the image and the text are processed by a 6-layer CNN and 2-layer LSTM respectively to unimodal representations, and then these representations are concatenated as a joint representation and passed through a few dense layers to make prediction. We decide to use a simple model here since this synthetic dataset is expected to be easy to learn.

For the real-world captioning dataset, we build a VGG-BERT based early fusion architecture. Similar to the CNN-LSTM model, a joint representation is learned by combining two unimodal representations and is passed through several dense layers. Given that the captioning dataset is far more complex to learn compared to the synthetic dataset, we pick VGG and BERT model to better capture unimodal characteristics.

We also tried using the CLIP (Radford et al., 2021) encoder for both image and text given CLIP is intrinsically designed for multimodal matching task.

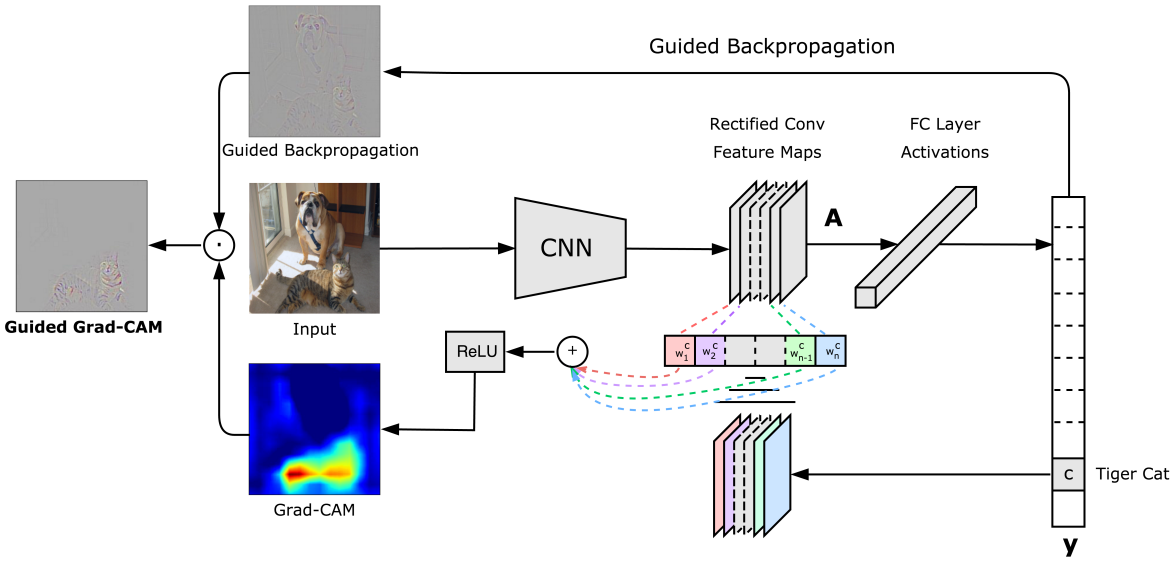


Figure 3. Grad-CAM illustration (Selvaraju et al., 2016)

5. Results and Discussion

5.1. Unimodal Results

Building explainable multimodal models are generally more difficult than unimodal ones due to the added complexity from more modalities. Therefore, it is crucial to test the gradient-based method for unimodal cases first. Currently, we have implemented gradient-based explanatory methods for two unimodal models: a VGG-based image classification model and a BERT-based sentiment classification model.

For the vision side, we implement the Grad-CAM algorithm as illustrated in Figure 3. The gradients are pooled across each channel of the last convolutional feature map to several scalar values, and a weighted sum is performed between these pooled scalar values and the activation maps from the corresponding channel in the last convolutional layer to generate a class saliency heatmap. Initial results, as shown in Figure 1, show that the model can successfully classify cat images and highlight important regions (e.g., cat faces and mouths) to justify their predictions.

For the language side, we incorporate a gradient-based idea that is similar to Grad-CAM, where the gradients are back-propagated to the contextualized representations of each token and then element-wise multiplied with the representation vectors / activations themselves. We use the norm of the resulting vector product to illustrate token importance to the final class prediction. We qualitatively compare the explainability effects produced by the gradient only versus the multiplication between gradient and activations. As shown in Figure 18, there are cases where the two methods both agree on the same set of salient words (example 2); there are also cases where only one of the two methods works (example 1 and 3); interestingly, in the last example, there are two reasonable sets of salient words and each method captures one of them.

5.2. Multimodal Results on Synthetic Dataset

Our model achieves 75% accuracy on our synthetic dataset. We then ran our gradient-based method to analyze the model performance. We demonstrate that our gradient-based method can visualize what raw features are paid attention to by the model, therefore improves the explainability of the model's decision process.

5.2.1. Explainability

Figure 8 is an example of a correct prediction output by the model, with relatively higher confidence score of 0.88. Since the text is describing the position of the largest number in the image, we could see that the most important area indicated by the gradient flow is where the largest number is. The second important area is the upper right corner mentioned in the text. On the text modality side, the 3 most important words indicated by the gradient are "largest", "at", "the". We have observed that gradient is not able to capture all key words in most cases, but they are sensitive to the words related to the types of descriptions, such as "largest", "smallest", even" and "odd".

Figure 9 is a correct prediction with a relatively low confidence score(0.56). The text describes the position of the

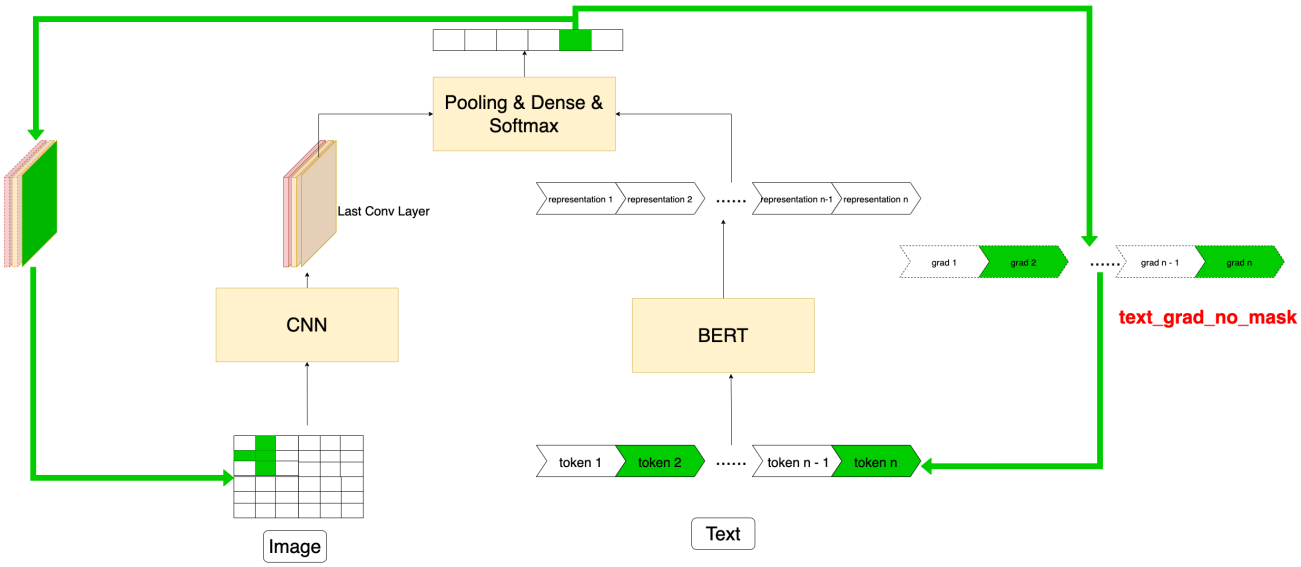


Figure 4. Forward and backward pass two full modalities through the model and get text_grad_no_mask.

first even number in the image. The most important words indicated by the gradient are: "number", "first", "even". We could see that no actual key words is captured. Moreover, the most important area in the image is the zero at the bottom right corner, which is not the position of the first even number in the image. The gradient at the position of the first even number(upper left corner) is slightly highlighted. Therefore, the model still made a correct prediction, but with low confidence score.

Figure 10 is an incorrect prediction output by the model. The text describes the position of the smallest number in the image, and the most important words captured by the gradient are: "smallest", "is", "number". We hope on the image side, the area where the smallest number is should be heavily highlighted, but it was only slightly highlighted. Since the gradient on the text side did not catch all key words, and the gradient on the image side is highlighting the wrong area, it makes this incorrect prediction explainable.

5.2.2. Cross-Modal Interactions

We will use the example shown in Figure 8 to visualize which features are important to modeling cross-model interactions. Since this example is a correct prediction with relatively high confidence, we would expect sufficient cross-modal interactions are captured.

First we mask part or all image to see how gradient on the text side changes. However, no matter how large the area we mask, the gradients on the text side all become close to zero. We will analyze and give our hypothesis on this issue in Section 5.4.

Then we mask text to see how gradients on the image chanes. Figure 11, 12, 13 shows the gradients on image after masking all text, the most important word, top three most important words, respectively. We could see that in all three cases, the gradients barely changed. Among all three changes, masking out all text content affected the gradient on the image the most. From the experiment we did, we suspect that the model is not learning enough cross-model interactions, and makes predictions mostly based on the image. More discussion on the dominance of image modality will be in section 5.4.

5.3. Multimodal Results on Real-World Dataset

Since we spent much time fine-tuning the VGG-BERT and CLIP models on the captioning dataset to get reasonable performance, only initial visualizations without masking schemes are available at this time. However, even with the initial visualizations, several interesting patterns can be found for both models.

We first found that only in very few cases, the VGG-BERT model takes advantage of both image and text information to make the decision. As shown in Figure 14, the model looks at both the dog and the water in the image while pointing to the "water" word in the caption. This can also be regarded as an initial trend of cross-modal interaction.

However, it is also found that in most cases, the VGG-BERT model solely relies on image to make a decision. This can be proved by the fact that corresponding image regions are highlighted while the contribution from text is attributed to random tokens such as $\langle PAD \rangle$ and $\langle SEP \rangle$. For instance,

as shown in Figure 15, the model can successfully capture the key regions such as "woman", "helmet", "bike", and "car" but the corresponding keywords in the caption are not highlighted.

Another finding is that we can utilize this visualization to detect multimodal classification failure modes. For example, as shown in Figure 16, the model mistakenly matches the image of someone riding a motorcycle with the text "This is a black dog splashing in the water". However, if we observe the highlighted region in the image, the model's prediction could be reasonable because the shape of the motorcycle resembles a dog and its blue color is related to water.

For CLIP-based model, it can achieve much better matching accuracy (95%) compared to the VGG-BERT model (75%). However, given that GradCAM algorithm only works for CNN-based architecture and CLIP is based on vision transformer architecture, we are still working on a way to dissect this large model and find an appropriate place to capture the image gradients. Currently, we can only back-propagate the gradients to the raw pixels, and as shown in Figure 17 the visualization does not show much meaningful information. This is as expected since the raw pixels do not contain high-level semantic information compared to later layers.

5.4. Discussions

For both VGG-BERT on the captioning dataset and CNN-LSTM on the synthetic dataset, we observe that image seems to be the dominant modality. However, we argue that it might be hasty to come to this conclusion for several reasons: (1) First, due to the limit of time, we did not train a perfect binary classification model for both datasets and only used very naive random negative sampling methods for the captioning dataset, resulting in some unreliability in the labels and predictions (2) For the image side, we use a mature and well-acknowledged GradCAM algorithm which multiplies the gradients with its activations as discussed in previous sections, but for the text side, there is no prior research indicating either gradient, activation, or the heuristic-based multiplication between them should also work well (3) Compared to the number of layers in VGG network, BERT architecture is far more nested and deeper, resulting in potential vanishing gradient problem and unsatisfactory text explanations.

6. Future Work

For immediate next steps, our plan is as follows: (1) further justify whether image is the dominant modality for both the synthetic and captioning datasets (2) perform four rounds of masking experiments on VGG-BERT model (3) replace the vision transformer to a CNN-like architecture in CLIP and incorporate GradCAM idea into it.

In the long term, we plan to design more robust quantitative evaluation metrics for our paradigm. In addition, we plan to extend our work from image-text models to video-text models and answer more complex multimodal explanatory questions such as **(1)** Which chunk of the video contributes most to the bimodal or trimodal interactions? **(2)** If conversations are involved in the video, which speaker turns contributes most to the interactions? **(3)** If multi-party conversations are involved in the video, which party contributes most to the interactions? We believe that our gradient-based framework is flexible for the extension to more modalities with temporal complexity.

References

Al-Shedivat, M., Dubey, A., and Xing, E. P. Contextual explanation networks. *CoRR*, abs/1705.10301, 2017. URL http://arxiv.org/abs/1705.10301.

Alvarez-Melis, D. and Jaakkola, T. S. Towards robust interpretability with self-explaining neural networks. *CoRR*, abs/1806.07538, 2018. URL http://arxiv.org/abs/1806.07538.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. 01 2018.

Arras, L., Horn, F., Montavon, G., Müller, K.-R., and Samek, W. Explaining predictions of non-linear classifiers in nlp. *ArXiv*, abs/1606.07298, 2016.

Arras, L., Osman, A., Müller, K.-R., and Samek, W. Evaluating recurrent neural network explanations. In *BlackboxNLP@ACL*, 2019.

Baltruaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41: 423–443, 2019.

Binder, A., Montavon, G., Bach, S., Müller, K.-R., and Samek, W. Layer-wise relevance propagation for neural networks with local renormalization layers, 2016.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., 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. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Chefer, H., Gur, S., and Wolf, L. Transformer interpretability beyond attention visualization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, 2021.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F. B., and Wattenberg, M. Visualizing and measuring the geometry of bert. In *NeurIPS*, 2019.

Denil, M., Demiraj, A., and de Freitas, N. Extraction of salient sentences from labelled documents. *ArXiv*, abs/1412.6815, 2014.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

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. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., and Li, B. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *ArXiv*, abs/2008.02312, 2020.

He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., and Girshick, R. B. Masked autoencoders are scalable vision learners. *ArXiv*, abs/2111.06377, 2021.

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., and Wei, Y. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

Kanehira, A., Takemoto, K., Inayoshi, S., and Harada, T. Multimodal explanations by predicting counterfactuality in videos, 2019.

Kaya, H., Gürpinar, F., and Salah, A. A. Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1651–1659, 2017.

Kim, Y. Convolutional neural networks for sentence classification. In *EMNLP*, 2014.

Lee, G., Jeong, J., Seo, S., Kim, C., and Kang, P. Sentiment classification with word attention based on weakly supervised learning with a convolutional neural network. *ArXiv*, abs/1709.09885, 2018.

Li, J., Chen, X., Hovy, E. H., and Jurafsky, D. Visualizing and understanding neural models in nlp. In *HLT-NAACL*, 2016.

Miao, Y., Gowayyed, M. A., and Metze, F. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 167–174, 2015.

Muhammad, M. B. and Yeasin, M. Eigen-cam: Class activation map using principal components. *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2020.

Nguyen, D. Comparing automatic and human evaluation of local explanations for text classification. In *NAACL*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark,

J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Rao, K., Sak, H., and Prabhavalkar, R. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 193–199, 2017. doi: 10.1109/ASRU.2017.8268935.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

Rogers, A., Kovaleva, O., and Rumshisky, A. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. 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.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences, 2019.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.

Tjoa, E. and Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32:4793–4813, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017a.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *ArXiv*, abs/1706.03762, 2017b.

Wang, J., Tuyls, J., Wallace, E., and Singh, S. Gradient-based analysis of nlp models is manipulable. *ArXiv*, abs/2010.05419, 2020.

Wu, J. and Mooney, R. J. Faithful multimodal explanation for visual question answering. *CoRR*, abs/1809.02805, 2018. URL http://arxiv.org/abs/1809.02805.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks, 2013.

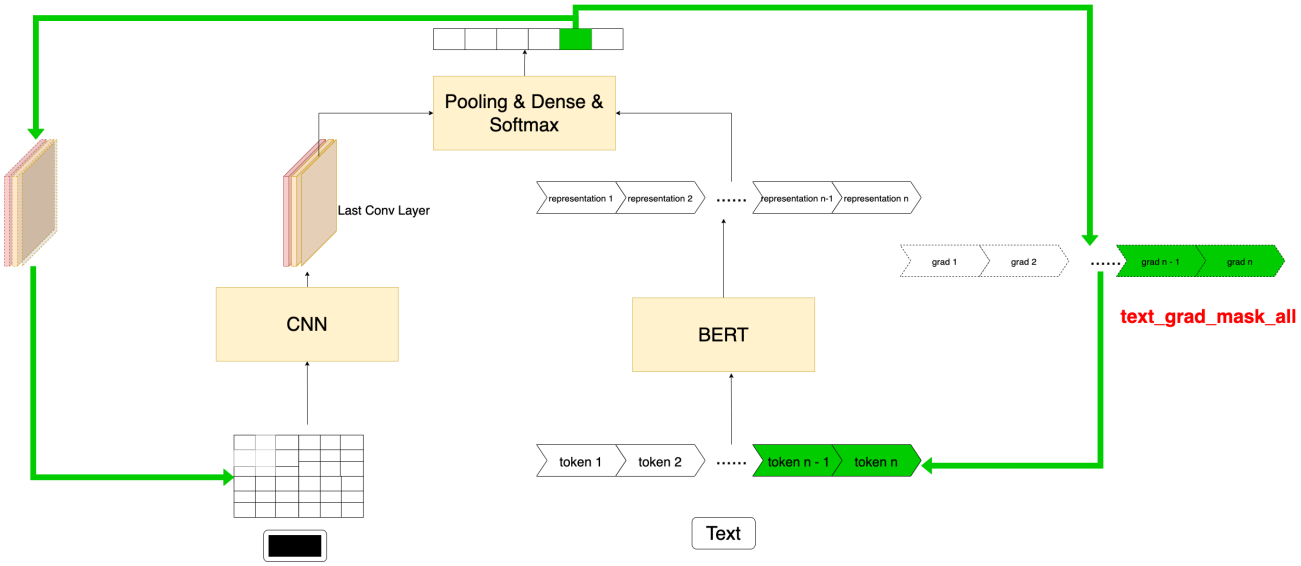


Figure 5. Mask the image completely, forward and backward pass only the text modality through the model and get text_grad_mask_all.

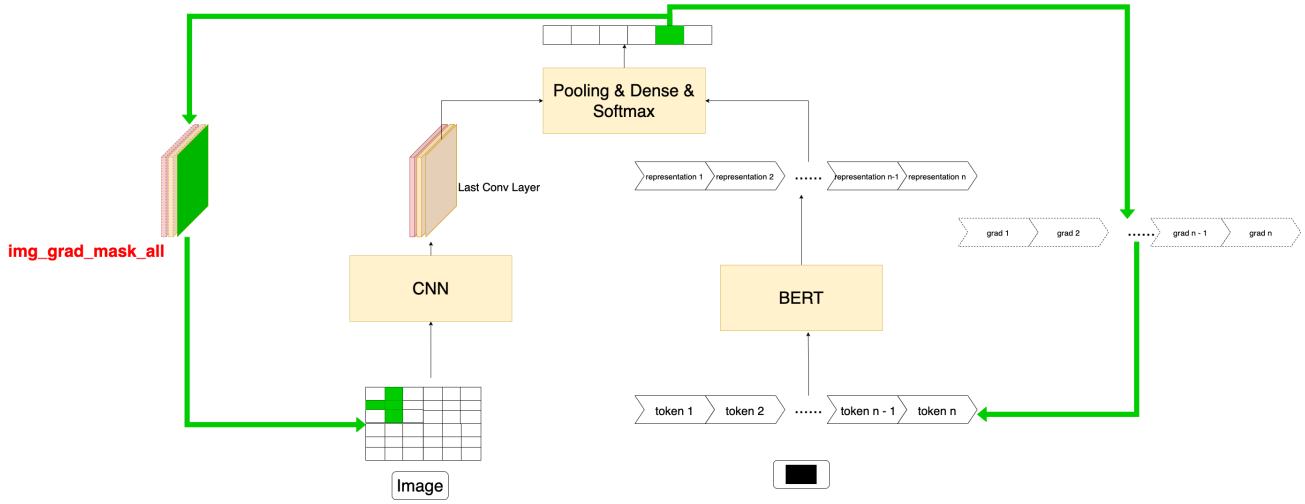


Figure 6. Mask the text completely, forward and backward pass only the image modality through the model and get img_grad_mask_all.

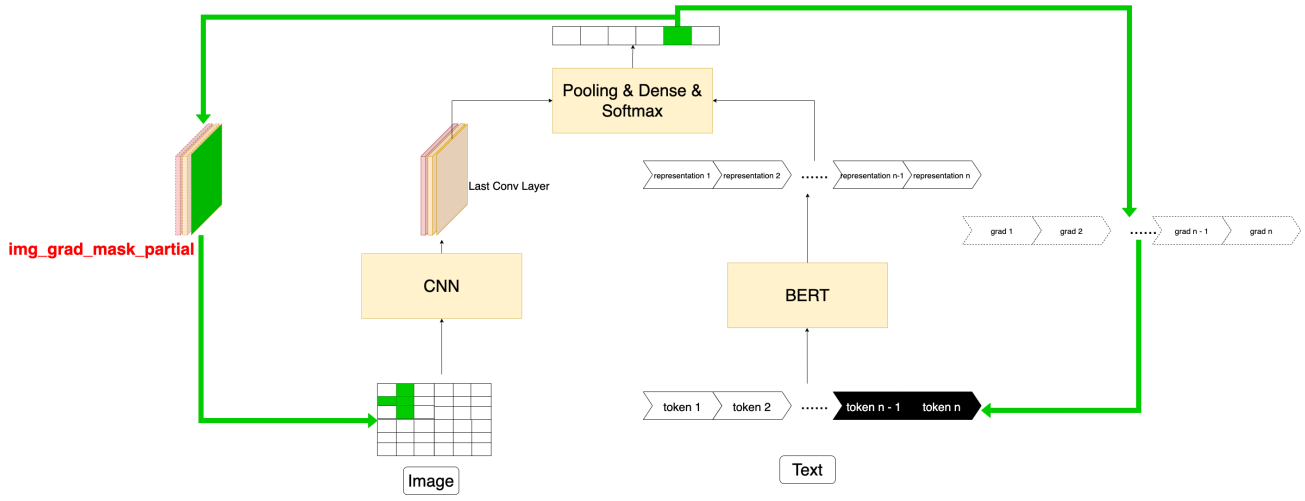


Figure 7. Mask the pre-selected bimodally-relevant tokens, forward and backward pass both modalities through the model and get

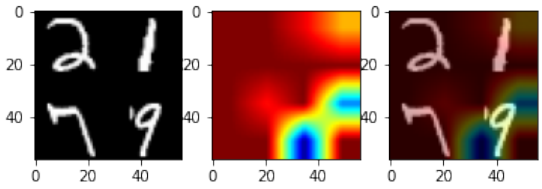


Figure 8. Label is 0; text is "The largest number is at the upper right corner". Model prediction is 0.0 with confidence 0.8802282139658928. Three most important words from gradient: largest, at, the

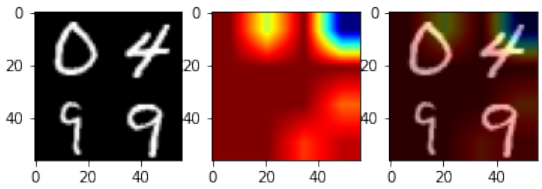


Figure 10. Label is 0; text is "The smallest number is on the bottom half". Model prediction is 1.0 with confidence 0.6061593890190125. Three most important words from gradient: smallest, is, number

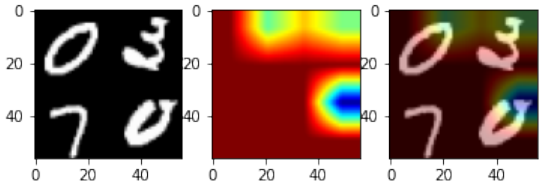


Figure 9. Label is 1; text is "The first even number is at the upper left corner". Model prediction is 1.0 with confidence 0.5645593404769897. Three most important words from gradient: number, first, even

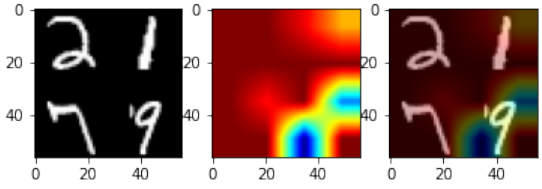


Figure 11. Label is 1; text is "Gradient on image after masking all words in the text.

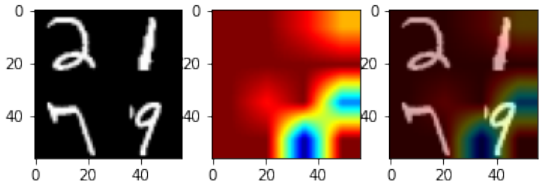


Figure 12. Label is 1; text is ”Gradient on image after masking the single most important word(”smallest”) in the text.

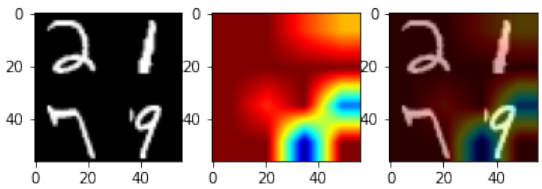
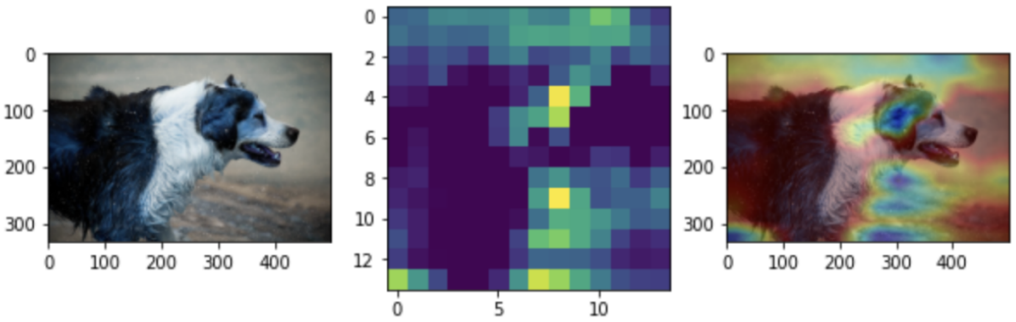
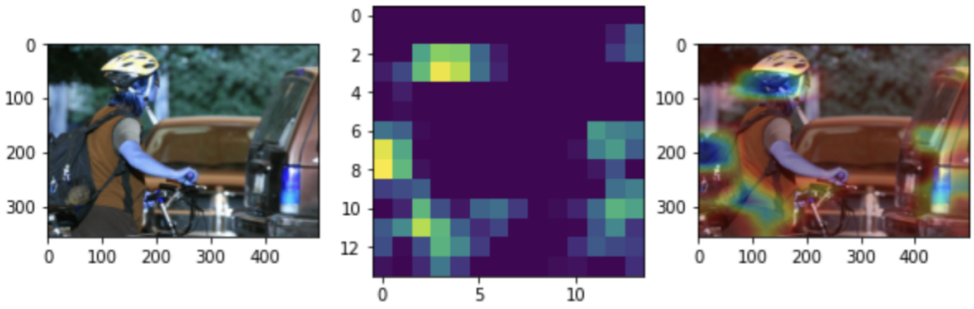


Figure 13. Label is 1; text is ”Gradient on image after masking top three most important word(”smallest”,”is”,”number”) in the text.



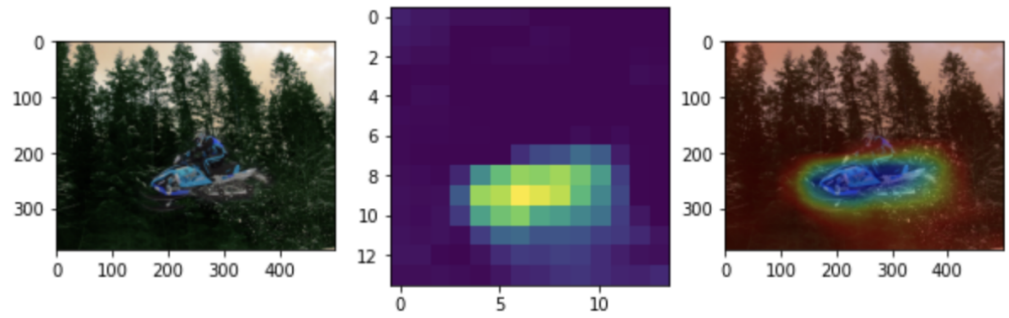
```
Image label is:  1
Image caption is:  A dog is looking at something near the water .
Prediction is 1 with confidence 0.5590145587921143
Most important tokens: ['the', 'water', '.', '[SEP]', '[PAD]', '[PAD]']
```

Figure 14. An example showing VGG-BERT model successfully making the correct prediction, attributing the contribution to the correct image region as well as correct text segment



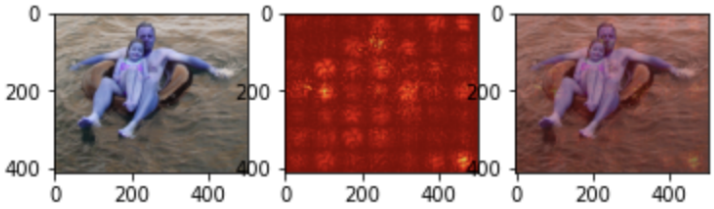
```
Image label is:  1
Image caption is:  A woman in a helmet rides her bike behind a car .
Prediction is 1 with confidence 0.5119242072105408
Most important tokens: ['.', '[SEP]', '[PAD]', '[PAD]', '[PAD]', '[PAD]']
```

Figure 15. An example showing VGG-BERT model successfully making the correct prediction, attributing the contribution to the correct image region, but failing to find important text contribution



```
Image label is:  0
Image caption is:  This is a black dog splashing in the water .
Prediction is 1 with confidence 0.5612204670906067
Most important tokens: ['[PAD]', '[PAD]', '[PAD]', '[PAD]']
```

Figure 16. An example showing VGG-BERT model making wrong prediction but this can be explained by focused image region



```
Image label is:  1
Image caption is:  A man and a small girl floating on an innertube
Prediction is 1.0 with confidence 0.9865541458129883
```

Figure 17. An example showing CLIP-based model making correct prediction but looking at random regions in the image

Input Sentence:
```
It's been great! A great to do list app
```

Ground Truth:
```
Positive.
```

Predict:
```
Positive.
```

Most important context given by backward gradient:
```
['great', '!', 'A', 'great']
```

Most important context given by forward activation x backward gradient:
```
['to', 'do', 'list', 'app']
```

Input Sentence:
```
Ruined a perfectly good app. Add task above/below features have been
removed. When I create a new task it goes all the way to the bottom of the
list. I have to then drag it all the way back up to the top of the list of
100+ tasks. App is now useless.
```

Ground Truth:
```
Negative.
```

Predict:
```
Negative.
```

Most important context given by backward gradient:
```
['is', 'now', 'useless', '.']
```

Most important context given by forward activation x backward gradient:
```
['is', 'now', 'useless', '.']
```

Input Sentence:
```
"Still the best by far and even better with online sync & backup. This is
the first todo list I've ever paid for and been happy to do it. I love the
granular approach to organization so you can get as complex as you like. I
use it every day for just about everything.
```

Ground Truth:
```
Positive.
```

Predict:
```
Positive.
```

Most important context given by backward gradient:
```
['about', 'everything', '.', '[SEP]']
```

Most important context given by forward activation x backward gradient:
```
['best', 'by', 'far', 'and']
```

Input Sentence:
```
There's no simple reminders or ability to add to do task, everything has
to be added via calendar which is irritating me...app is useless right now
```

Ground Truth:
```
Negative.
```

Predict:
```
Negative.
```

Most important context given by backward gradient:
```
["'", 's', 'no', 'simple']
```

Most important context given by forward activation x backward gradient:
```
['app', 'is', 'useless', 'right']
```

Figure 18. Four NLP explanatory examples