Gradient Frequency Modulation for Visually Explaining Video Understanding Models: Supplement Materials

Xinmiao Lin
xl3439@rit.edu
Wentao Bao
wb6219@rit.edu
Matthew Wright
Matthew.Wright@rit.edu
Yu Kong
Yu.Kong@rit.edu

Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY

1 Optimality of GFM

Below, we demonstrate that applying frequency modulation on ∇M is equivalent to the frequency modulation on the optimal mask M^*. We modulate the gradient map in the frequency domain, $G = \mathcal{H}(\nabla M)$, as follows:

$$\nabla \tilde{M} = \tilde{\mathcal{H}}(\mathcal{O}_A(G)) + \tilde{\mathcal{H}}(\mathcal{O}_B(G)), \tag{1}$$

Where $\mathcal{O}_A(G) = A \odot G$ and $\mathcal{O}_B(G) = B \odot G$ are linear. We now show that the Inverse Discrete Cosine Transform, $\tilde{\mathcal{H}}$, is also linear. $\tilde{\mathcal{H}}$ is defined as follows:

$$\begin{aligned} \tilde{\mathcal{H}}(G)_{x,y,z} &= a \sum_{k=0}^{T-1} \sum_{j=0}^{W-1} \sum_{i=0}^{H-1} c_i c_j c_k G_{i,j,k} d_{i,x}^{(H)} d_{j,y}^{(W)} d_{k,z}^{(T)} \\ &= a h_z^T \left(h_y^T \left(h_x^T G \right) \right) \end{aligned} \tag{2}$$

The equation (2) is a matrix multiplication and thus linear. Note that the functions $d(\cdot,\cdot)$ and c are defined the same as in the paper. We also know that DCT is invertible and linear, thus IDCT is also linear [1].

Therefore, the gradient frequency modulation can be summarized as:

$$\nabla \tilde{M} = \tilde{\mathcal{H}} \circ \mathcal{O} \circ \mathcal{H}(\nabla M) \tag{3}$$

Because $\tilde{\mathcal{H}}$, $\mathcal{H}$ and $\mathcal{O}$ are linear transformations, we have the following gradient ascent rule:

$$\begin{aligned}\mathcal{F}(M^{f+1}) &= \mathcal{F}(M^f + \varepsilon \nabla M) \\ &= \mathcal{F}(M^f) + \varepsilon \mathcal{F}(\nabla M)\end{aligned} \tag{4}$$

The equation (4) shows that applying frequency modulation on the gradient map ∇M is equivalent to frequency modulation on the optimal mask M^*.

2 Additional Ablation Studies Results

r_l	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
STC	55.5	58.8	61.9	61.9	62.9	62.7	**67.8**	63.5

Table 1: **Without high frequencies results with model R(2+1)d**. The results are the quantitative results are the purple circles in the figure 3 of the paper.

r_l	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.4
r_h	0.1	0.2	0.3	0.4	0.5	0.6	0.1	0.2	0.3	0.4	0.5
STC	61.9	61.9	61.9	61.9	61.9	62.7	61.9	61.8	62.0	62.8	62.8

Table 2: **Low and High Frequencies where** $r_l = \{0.3, 0.4\}$. This table presents the results of the using a combination of low and high frequencies on the dataset Epic-Kitchens-Object and the model R(2+1)D.

r_l	0.5	0.5	0.5	0.5	0.6	0.6	0.6
r_h	0.1	0.2	0.3	0.4	0.1	0.2	0.3
STC	62.9	62.5	62.1	61.7	62.7	**63.0**	62.7

Table 3: **Low and High Frequencies where** $r_l = \{0.5, 0.6\}$. This table presents the results of the using a combination of low and high frequencies on the dataset Epic-Kitchens-Object and the model R(2+1)D.

In table 1, we present the results of the STC performance using only low frequencies. Note that the table corresponds to the horizontal purple circles in the figure 3 of the paper. The table 2 reports the performance using a combination of low and high frequency signals which correspond to the vertical pink and yellow circles in the figure 3. The table 3 reports the results of $r_l = \{0.5, 0.6\}$ where $r_l = \{0.5\}$ are the green circles in the figure 3.

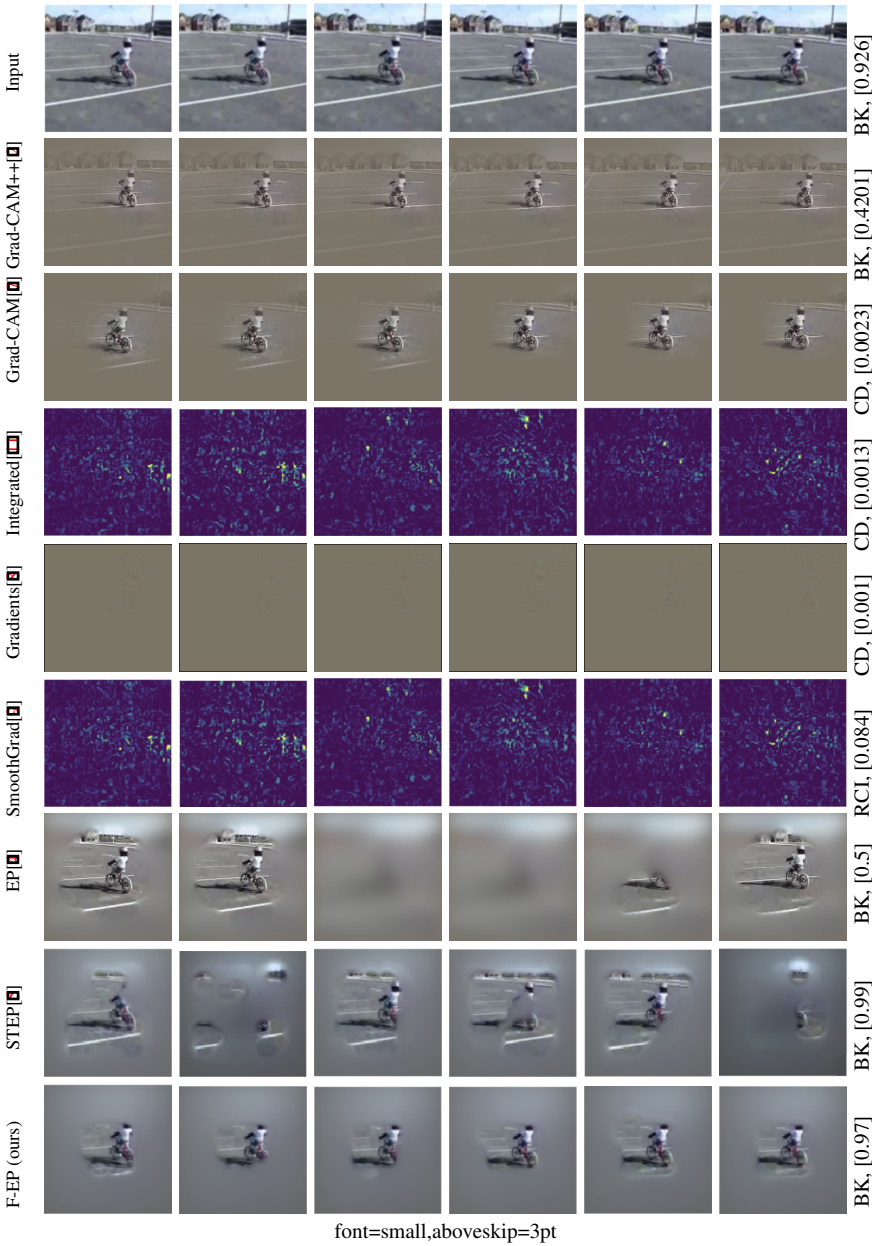
We see that when the performance of F-EP using a combination of low and high frequencies is superior than the baselines (3D_EP has 58.0, STEP has 61.0), but lower than the best performance of using low frequency signals only which is 67.8 at $r_l = 0.7$. For $r_l = \{0.3, 0.4\}$, a larger amount of high frequency signals improve the STC performance, while when r_l increases, $r_l = \{0.5, 0.6\}$, an increasing amount of high frequency signals do not necessarily help for better spatiotemporal consistency.

3 Qualitative Results

Figure 1 compares the baseline methods with F-EP on the dataset UCF101-24 and the model TSM. F-EP is able to localize the activity which is biking accurately both spatially and temporally. Although Grad-CAM and Grad-CAM++ also localize the biking girl, a substantial

amount of background information is also included. Backpropagation-based methods Gradients, Integrated Grad and SmoothGrad produce noisy explanations that are hard to interpret.

Figure 2 compares the methods on the dataset Epic-Kitchens-Object and the model R(2+1)D. We see that F-EP is able to localize the object cupboard for most of the scene, but it also highlights the closet which is not the target object (4th frame). One future work direction is to optimize the masks to attend only to salient frames. CAM- and backpropagation-based methods are noisy and not spatiotemporal consistent, e.g., Grad-CAM++ and Grad-CAM fail to localize the cupboard. The baselines 3D_EP and STEP fails to localize cupboard in the last two frames and the cupboard at the second object is masked by the explanations. Integrated Grad and SmoothGrad do not capture the cupboard object at the second frame.



font=small,aboveskip=3pt

Figure 1: **Visual comparison of explanation methods on the UCF101-24 dataset [10] with model TSM [8].** The input (first row) contains consecutive frames from the activity Biking. On the right of each method, the first word is the predicted label on the explanation where BK = Biking, CD = CliffDiving, RCI = RockClimbingIndoor, and the number denotes the predicted probability.

font=small,aboveskip=3pt

Figure 2: **Visual comparison of explanation methods on the Epic-Kitchens-Object dataset [3]**. The input (first row) contains frames from the object cupboard. On the right of each method, the first word is the predicted label on the explanation where CB = cupboard and the number denotes the corresponding predicted probability.

References

[1] J.F. Blinn. What's that deal with the dct? *IEEE Computer Graphics and Applications*, 13, 1993.

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, 2018.

[3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[4] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, 2019.

[5] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *WACV*, 2021.

[6] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019.

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[8] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[9] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. In *ICML Workshop*, 2017.

[10] Khurram Soomro, Amir Roshan Zamir, Mubarak Shah, Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[11] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.