A Biologically Inspired Filter Significance Assessment Method for Model Explanation

Emirhan Böge, Yasemin Gunindi, Murat Bilgehan Ertan,
Erchan Aptoula, Nihan Alp, and Huseyin Ozkan

Presented by Yasemin Günindi, PhD student in Experimental Psychology
09.07.2025

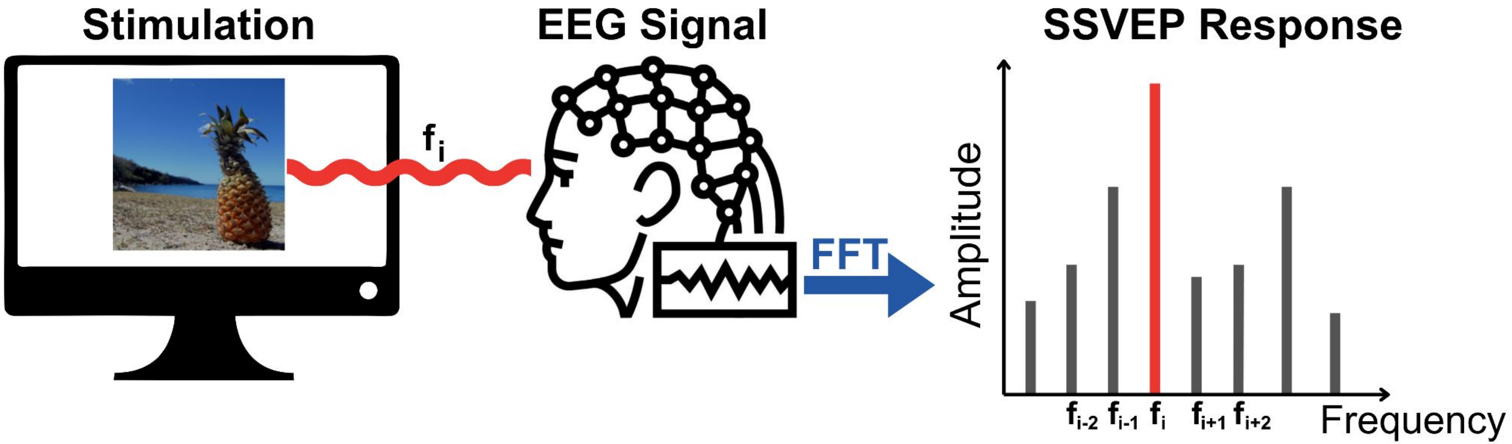
Motivation

- CNNs are said to be inspired by the brain’s visual processing network — but how biologically inspired are they, really?
- CNNS are powerful, but are hard to interpret
- Class Activation Maps (CAM) [1] visualize model decisions, but include noisy filter activations
- Inspiration from neuroscience to identify only the most functionally responsive filters
    - more efficient
    - biologically plausible
    - clearer

[1] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016)

Biological Inspiration: SSVEP

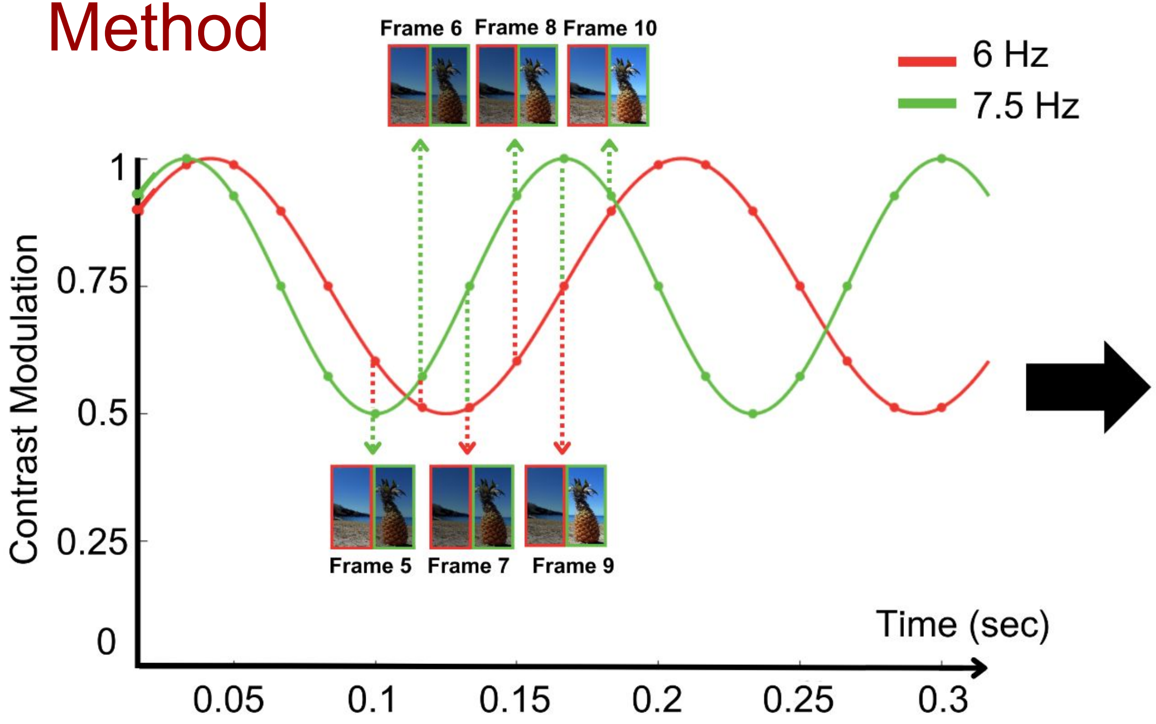
- ***SSVEP***: Steady-State Visually Evoked Potential[2]
- Used in neuroscience to probe frequency-locked neural responses[3]
- Can CNN filters show analogous tuning?

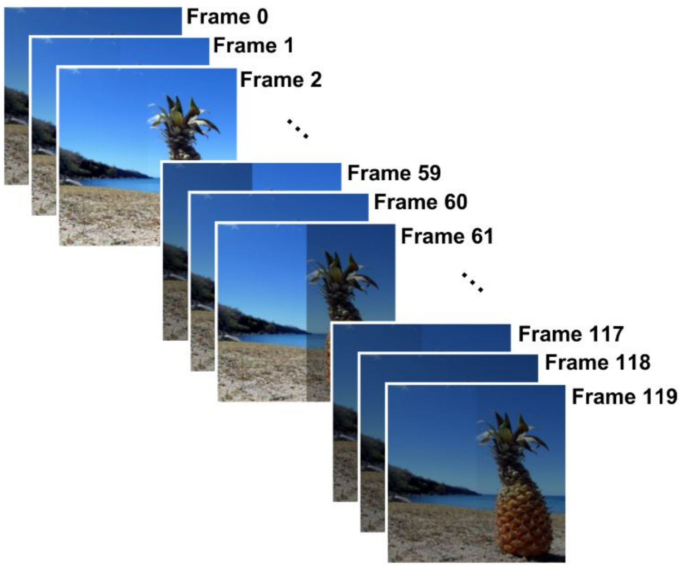


[2] Regan, D.: An effect of stimulus colour on average steady-state potentials evoked in man. Nature 210(5040), 1056–1057 (1966)

[3] Norcia, A.M., Appelbaum, L.G., Ales, J.M., Cottereau, B.R., Rossion, B.: The steady-state visual evoked potential in vision research: A review. Journal of vision 15(6) (2015)

Method





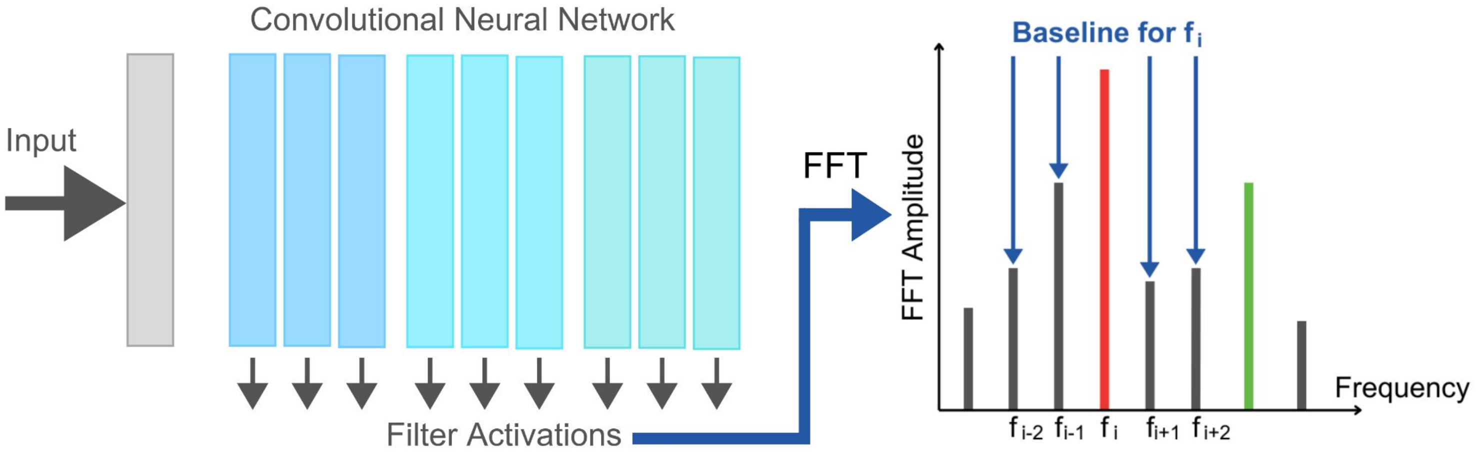
$$\omega_i = 2\pi f \frac{i}{\mathrm{FPS}} + \phi$$

ω_i: modulation angle
f: modulation frequency
FPS: frame rate
i: sequential frame index
ϕ: phase shift (0)

$$I' = \left(\frac{\sin(\omega_i) + 1}{2} (s_{\max} - s_{\min}) + s_{\min} \right) \cdot I$$

$s_{\max}, s_{\min}$: maximum and minimum modulation factors
I: image intensity values

Quantifying Filter Importance



- Record filter activations over time.
- Apply FFT to get frequency response.
- Compute SNR(f) = Signal / Neighboring Noise
- Rank filters by mean SNR across frequencies.

$$\mathrm{SNR}(f) = \frac{F(f)}{\frac{1}{|\mathcal{N}(f)|} \sum_{k \in \mathcal{N}(f)} F(k)}$$

SSVEP-Guided CAM

- CAM: weighted sum of filter activations
- Replace all filters with top-K based on SSVEP
- Improves focus, reduces noise

$$M_{\text{SSVEP-CAM}}(x, y) = \sum_{k \in \mathcal{K}} w_k A_k(x, y)$$

$A_k(x, y)$: the activation map

w_k : importance weight

Experimental Setup

Models: **VGG-16**[4], **ResNet-50**[5], **ResNeXt-50**[6]

Dataset: 5000 **ImageNet**[7] validation images

Tools: **GradCAM**[8], **GradCAM++**[9], **EigenCAM**[10], **LayerCAM**[11]

[4] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR (2015)

[5] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR. pp. 770–778 (2016)

[6] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)

[7] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision 115, 211–252 (2015)

[8] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128(2), 336—-359 (Oct 2019)

[9] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Gradcam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (2018)

[10] Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1—-7 (Jul 2020)

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

Evaluation Metrics

- **Energy Concentration (EC)**
  - *Are activations focused?*

$$EC = \frac{\sum_{i \in S} H_i}{\sum_{j \in \Omega} H_j}$$

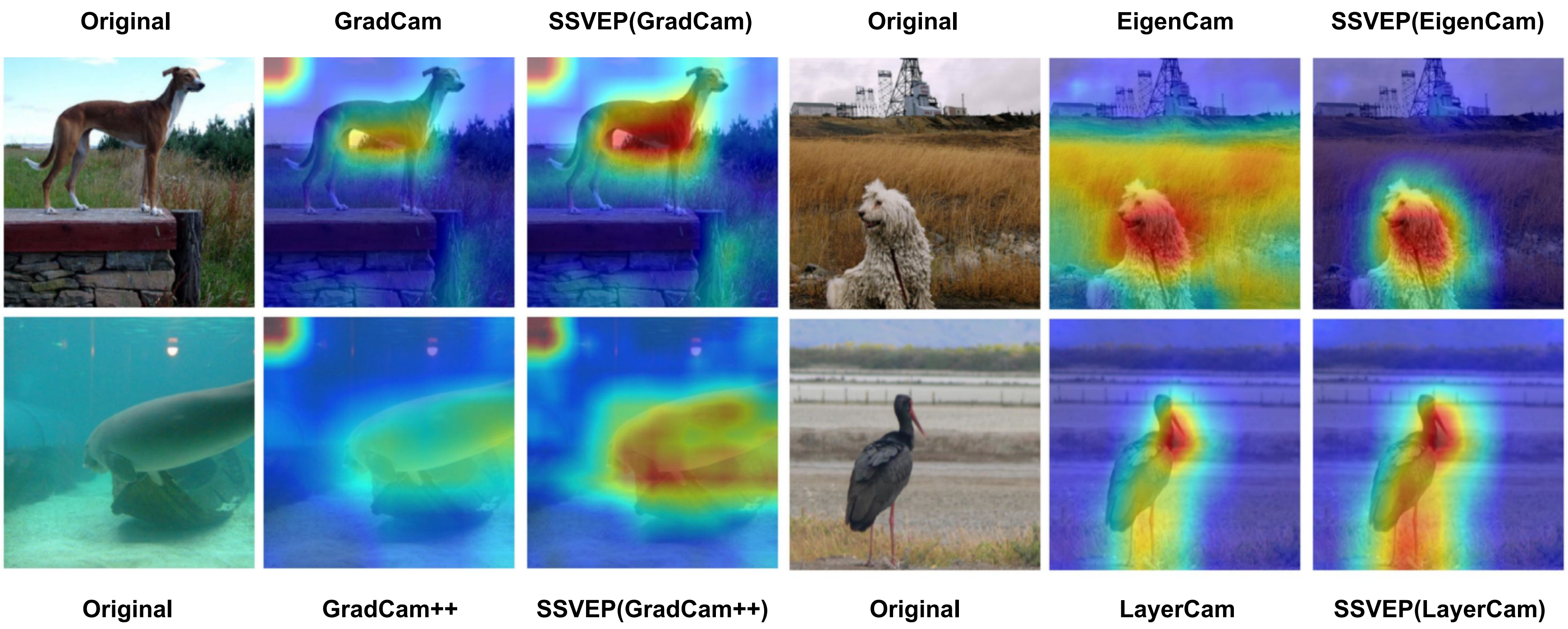
H_i : the heatmap intensity at pixel i

Ω : the set of all pixels in the heatmap

$S \subset \Omega$: the subset of pixels corresponding to the top 20% highest heatmap intensities

- **Loc-1** and **Loc-5** Accuracy (Zhou et al., 2016)
  - *Are highlighted regions aligned with class-relevant areas?*

Results


Original
GradCam
SSVEP(GradCam)
Original
EigenCam
SSVEP(EigenCam)
Original
GradCam++
SSVEP(GradCam++)
Original
LayerCam
SSVEP(LayerCam)

Energy Concentration (%) *(M ± SD)*

Algorithm	**VGG-16**		**ResNet-50**		**ResNeXt-50**	
	Baseline	SSVEP	Baseline	SSVEP	Baseline	SSVEP
GradCAM	47.50 ± 8.26	$\mathbf{53.91 \pm 11.05}$	$\mathbf{45.24 \pm 6.49}$	44.31 ± 7.12	$\mathbf{42.02 \pm 5.63}$	39.39 ± 5.34
GradCAM++	39.18 ± 4.92	$\mathbf{44.89 \pm 10.19}$	39.31 ± 4.87	$\mathbf{42.68 \pm 7.17}$	$\mathbf{39.21 \pm 4.06}$	38.56 ± 5.04
EigenCAM	62.79 ± 17.33	$\mathbf{69.50 \pm 13.52}$	49.90 ± 11.48	$\mathbf{54.37 \pm 11.68}$	41.81 ± 7.28	$\mathbf{44.72 \pm 7.60}$
LayerCAM	46.63 ± 7.95	$\mathbf{47.09 \pm 8.00}$	45.66 ± 8.55	$\mathbf{46.48 \pm 9.01}$	44.27 ± 7.46	$\mathbf{44.75 \pm 7.65}$

Localization Accuracy

Algorithm	**Prediction Level**	**VGG-16**		**ResNet-50**		**ResNeXt-50**	
		Baseline	SSVEP	Baseline	SSVEP	Baseline	SSVEP
GradCAM	loc1	**19.76%**	17.04%	18.52%	**18.92%**	**18.28%**	18.12%
	loc5	**24.50%**	21.16%	22.64%	**23.14%**	**22.18%**	21.86%
GradCAM++	loc1	**17.44%**	16.72%	18.72%	**19.90%**	18.12%	**18.52%**
	loc5	**21.54%**	20.66%	22.80%	**24.22%**	21.62%	**22.32%**
EigenCAM	loc1	**21.28%**	19.00%	25.06%	**26.12%**	21.72%	**22.78%**
	loc5	**25.78%**	22.78%	29.92%	**31.14%**	25.88%	**27.24%**
LayerCAM	loc1	21.84%	**22.00%**	**23.14%**	23.08%	**22.52%**	22.26%
	loc5	26.52%	**26.94%**	**28.16%**	28.06%	**27.04%**	26.82%
Average	loc1	**20.08%**	18.69%	21.36%	**22.01%**	20.16%	**20.42%**
	loc5	**24.59%**	22.89%	25.88%	**26.64%**	24.18%	**24.56%**

Discussion

- Consistent improvement in EC with SSVEP-enhanced versions
- Maintaining competitive localization accuracy
- Focus vs completeness
- Architecture sensitivity: VGG-16
- Top-K% filters instead of fixed number

Conclusion & Future Work

- SSVEP filter selection = plausible, effective, low-noise
- Future work: dynamic K, more datasets, task-generalization