Interpreting Optical Diffractive Neural Networks (ODNN) via Gradient-based Physical Activation Mapping (GPAM)

Hong-Yueh Huang[1], Cheng-Jun Huang[1], Ting-Siang Ou Yang[1], Jui-Chien Lin[1], Ming-Hua Lin[1] and Li-Bang Wang[1]

[1] Department of Physics, National Tsing Hua University, Hsinchu, Taiwan 30013, Republic of China

Optical diffractive neural networks (ODNNs) compute through light propagation and diffraction, but their internal physical decision mechanism is difficult to interpret. We propose Gradient-based Physical Activation Mapping (GPAM), which estimates the contribution of each diffractive neuron by back-propagating gradients through physical wave propagation. We further establish a quantitative evaluation framework including localization, energy concentration, weakly-supervised segmentation, and faithfulness analysis using mask ratio. Experiments on MNIST and Fashion-MNIST demonstrate that GPAM reveals physically meaningful regions and that masking high-relevance regions significantly reduces confidence and accuracy. The proposed framework provides a tool for interpreting optical neural networks and analyzing optical system behavior.

Keywords: Optical diffractive neural networks, Interpretability, Optical computing, Saliency map, Optical neural networks

ODNN Setup & Datasets

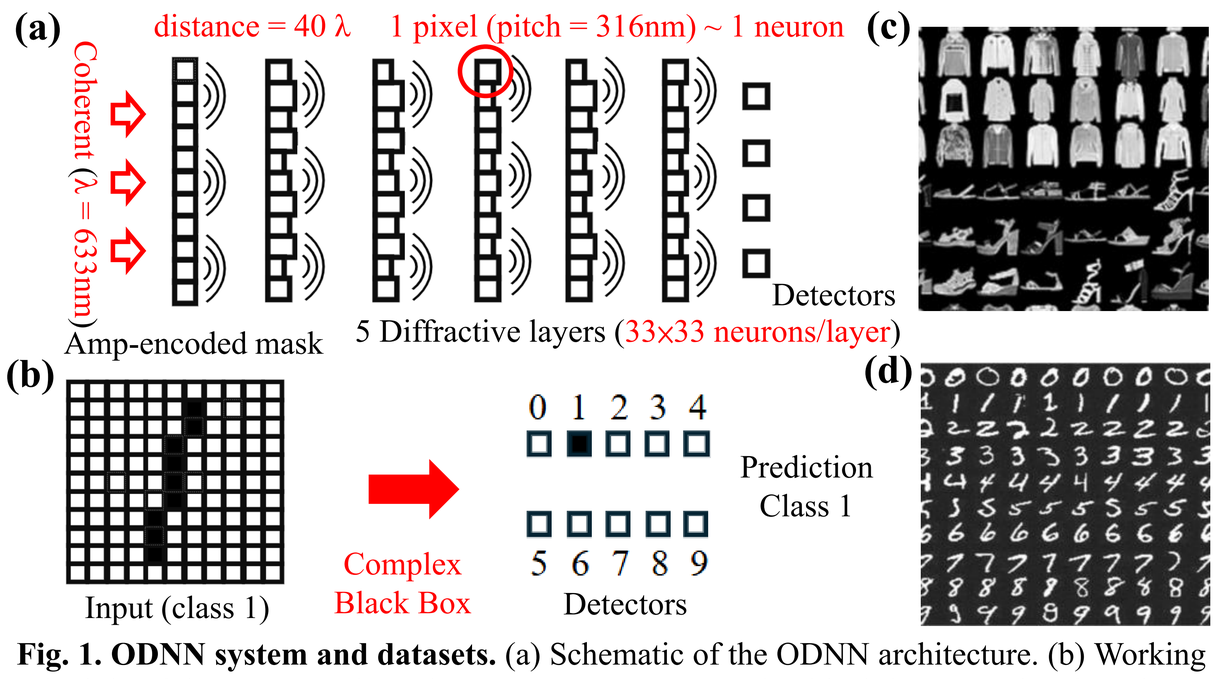


Fig. 1. ODNN system and datasets. (a) Schematic of the ODNN architecture. (b) Working principle of the system, where the ODNN acts as a complex black-box mapping from input to detector outputs. (c) Fashion-MNIST dataset. (d) MNIST dataset. (**Train/Test = 6w/1w**)

GPAM Method

Optical Forward Model

$$\mathbf{U}_{l+1} = \mathbf{P}_l \mathbf{D}_l \mathbf{U}_l$$
$$\mathbf{U}_L = \mathbf{P}_{L-1}\mathbf{D}_{L-1}\cdots\mathbf{P}_1\mathbf{D}_1\mathbf{U}_0$$
$$\mathbf{I}(x,y) = |\mathbf{U}_L(x,y)|^2$$
$$\mathbf{y}_k = \sum_{(x,y)\in \mathbf{D}_k} \mathbf{I}(x,y)$$

Backward Optical Field

$$\mathbf{G}_L = \mathbf{M}_k \cdot \mathbf{U}_L^*$$
$$\mathbf{G}_l = \mathbf{P}_l^\dagger \mathbf{G}_{l+1}$$

Pixel Perturbation Concept

$$\mathbf{U}_l(x,y) \to \mathbf{U}_l(x,y) + \delta\mathbf{U}$$
$$\delta \mathbf{y}_k = \frac{\partial \mathbf{y}_k}{\partial \mathbf{U}_l(x,y)}\,\delta\mathbf{U}$$
$$\frac{\partial \mathbf{y}_k}{\partial \mathbf{U}_l(x,y)}$$

GPAM Relevance Definition

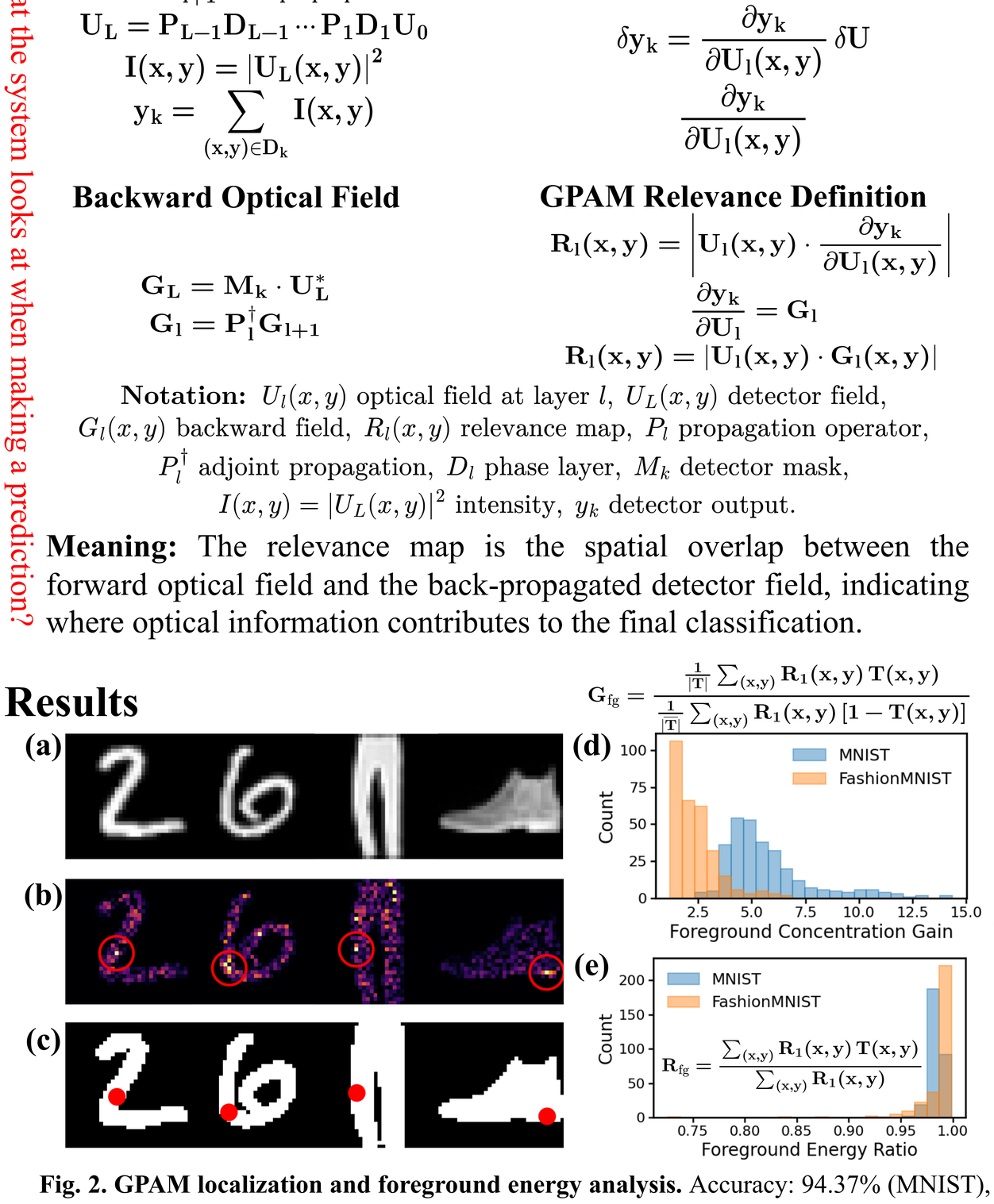
$$\mathbf{R}_l(x,y) = \left|\mathbf{U}_l(x,y)\cdot\frac{\partial \mathbf{y}_k}{\partial \mathbf{U}_l(x,y)}\right|$$
$$\frac{\partial \mathbf{y}_k}{\partial \mathbf{U}_l} = \mathbf{G}_l$$
$$\mathbf{R}_l(x,y) = |\mathbf{U}_l(x,y)\cdot\mathbf{G}_l(x,y)|$$

Notation: $U_l(x,y)$ optical field at layer l, $U_L(x,y)$ detector field, $G_l(x,y)$ backward field, $R_l(x,y)$ relevance map, P_l propagation operator, $P_l^\dagger$ adjoint propagation, D_l phase layer, M_k detector mask, $I(x,y) = |U_L(x,y)|^2$ intensity, y_k detector output.

Meaning: The relevance map is the spatial overlap between the forward optical field and the back-propagated detector field, indicating where optical information contributes to the final classification.

What the system looks at when making a prediction?

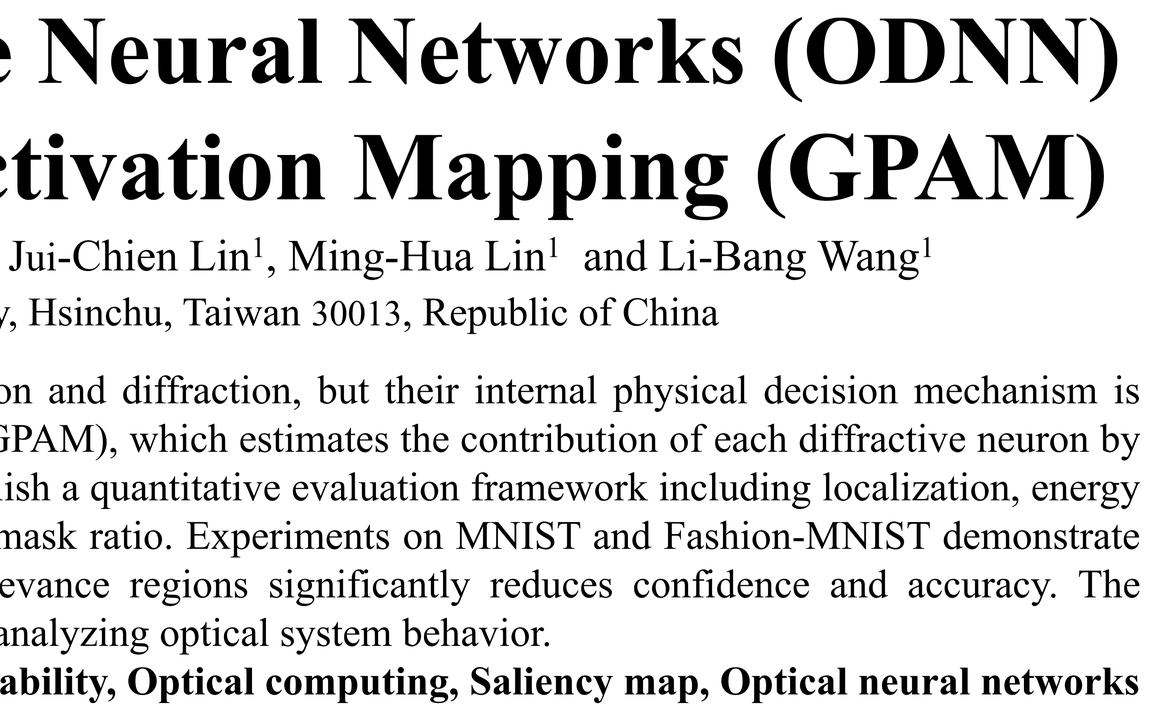
Results



$$\mathbf{G}_{\text{fg}} = \frac{\frac{1}{|\mathbf{T}|}\sum_{(x,y)} \mathbf{R}_1(x,y)\,\mathbf{T}(x,y)}{\frac{1}{|\mathbf{T}|}\sum_{(x,y)} \mathbf{R}_1(x,y)\,[1-\mathbf{T}(x,y)]}$$

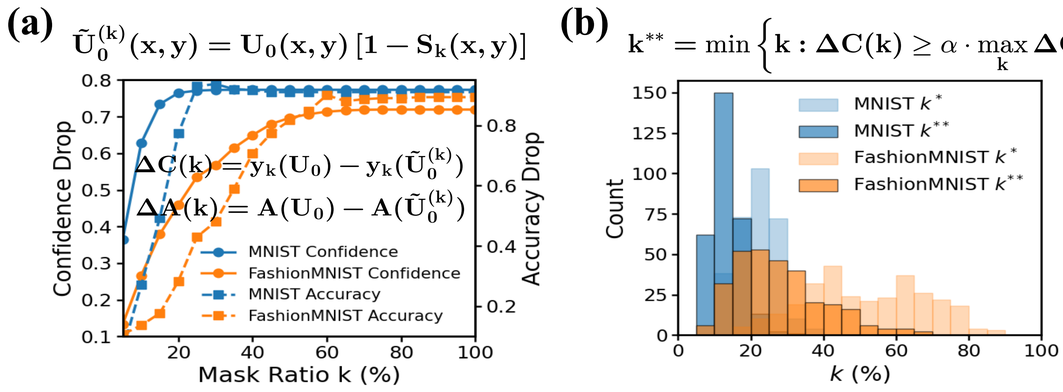
$$\mathbf{R}_{\text{fg}} = \frac{\sum_{(x,y)} \mathbf{R}_1(x,y)\,\mathbf{T}(x,y)}{\sum_{(x,y)} \mathbf{R}_1(x,y)}$$

Fig. 2. GPAM localization and foreground energy analysis. Accuracy: 94.37% (MNIST), 92.75% (FashionMNIST). (a) Input images. (b) GPAM relevance maps $\boldsymbol{R_1}$ with peak locations.(c) Ground truth masks $\boldsymbol{T}$; pointing game success = 100% for correct predictions.(d) Foreground concentration gain. (e) Foreground energy ratio.



$$\mathbf{S}_k(x,y) = \begin{cases} 1, & R_1(x,y) \text{ is in top-}k\% \\ 0, & \text{otherwise} \end{cases}$$
$$\mathbf{IoU(k)} = \frac{|\mathbf{S_k} \cap \mathbf{T}|}{|\mathbf{S_k} \cup \mathbf{T}|}$$
$$\mathbf{k}^* = \arg\max_{\mathbf{k}} \mathbf{IoU(k)}$$
$$\mathbf{IoU}_{\text{bbox}} = \frac{|\mathbf{B}(\mathbf{S_k}) \cap \mathbf{B}(\mathbf{T})|}{|\mathbf{B}(\mathbf{S_k}) \cup \mathbf{B}(\mathbf{T})|}$$
$$\mathbf{d} = \|\mathbf{c}_{\mathbf{S_k}} - \mathbf{c}_{\mathbf{T}}\|$$

Fig. 3. GPAM weakly-supervised localization evaluation. (a) Input. (b) Ground truth. (c) Segmentation from GPAM top-k regions. (d) Bounding box derived from (c). (e) IoU vs top-k. (f) Maximum IoU. (g) BBox IoU at k^*. (h) Center distance at k^*.



$$\tilde{\mathbf{U}}_0^{(k)}(x,y) = \mathbf{U}_0(x,y)\,[1-\mathbf{S}_k(x,y)]$$
$$\Delta\mathbf{C}(k) = \mathbf{y}_k(\mathbf{U}_0) - \mathbf{y}_k(\tilde{\mathbf{U}}_0^{(k)})$$
$$\Delta\mathbf{A}(k) = \mathbf{A}(\mathbf{U}_0) - \mathbf{A}(\tilde{\mathbf{U}}_0^{(k)})$$
$$\mathbf{k}^{**} = \min\left\{\mathbf{k} : \Delta\mathbf{C}(\mathbf{k}) \geq \alpha \cdot \max_{\mathbf{k}} \Delta\mathbf{C}(\mathbf{k})\right\}$$

Fig. 4. Faithfulness and critical mask ratio analysis. (a) Confidence drop and accuracy drop as a function of mask ratio k when masking the top-k% GPAM relevance pixels. (b) Distribution of the critical mask ratios k^* (segmentation optimum) and k^{**}(faithfulness threshold) for MNIST and Fashion-MNIST datasets.

Discussion

Datasets images contain sparse foreground strokes on a dark background, which may bias localization methods toward foreground regions. As a control experiment, future work will introduce background illumination bias while maintaining classification accuracy, to verify that GPAM localization reflects decision-relevant regions rather than input sparsity.

GPAM is not only a visualization tool, but also a design and analysis tool for optical systems. **Potential Applications** were shown below.

- Physical AI interpretability
- Optical neural network debugging
- Optical computing optimization
- Optical system/inverse design
- Robust optical neural networks
- Optical microscopy & sensing

Conclusion

We proposed GPAM to interpret optical diffractive neural networks and established a quantitative framework to evaluate localization, segmentation, and faithfulness. Results show that GPAM reveals physically meaningful regions and that masking high-relevance regions significantly reduces prediction confidence and accuracy. This work moves optical neural networks from black-box models toward physically interpretable computing systems.

GPAM enables explainable optical neural networks !