Understanding Impacts of High-Order Loss Approximations and Features in Deep Learning Interpretation

Sahil Singla

Joint work with

Eric Wallace, Shi Feng, Soheil Feizi

University of Maryland

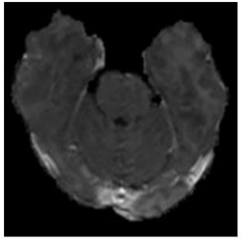
Pacific Ballroom #**69**, 6:30-9:00 PM

June 13th 2019

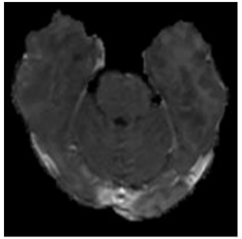
https://github.com/singlasahil14/CASO

Why Deep Learning Interpretation?

Why Deep Learning Interpretation?



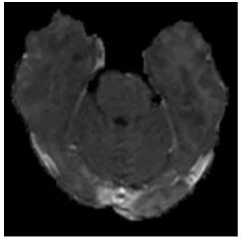
Why Deep Learning Interpretation?



Deep neural network

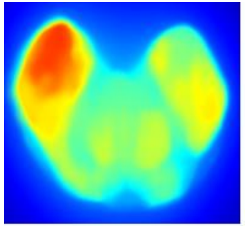
Classified as y=0 (low-grade glioma)

Why Deep Learning Interpretation?



Deep neural network

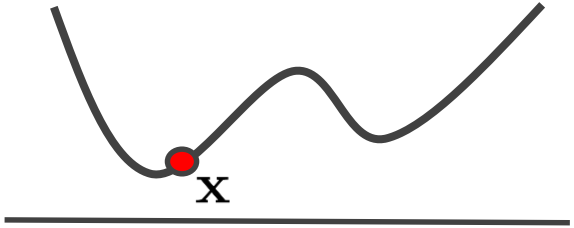
Classified as y=0 (low-grade glioma)



Saliency map to highlight salient features

We need to explain AI decisions to humans

Assumptions of Current Methods

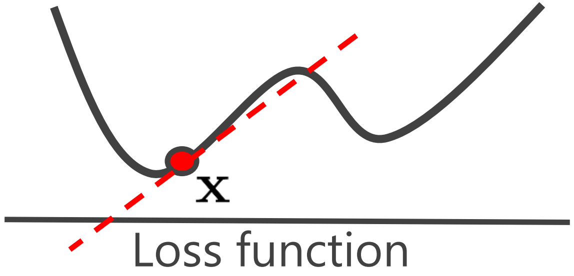


Loss function

$$\max_{\Delta} \ell(f_\theta(\mathbf{x} + \Delta), y)$$

$$\|\Delta\|_2 \leq \rho$$

Assumptions of Current Methods



$$\ell(f_\theta(\mathbf{x}+\Delta), y) \approx \ell(f_\theta(\mathbf{x}), y) + \mathbf{g}_{\mathbf{x}}^t \Delta$$

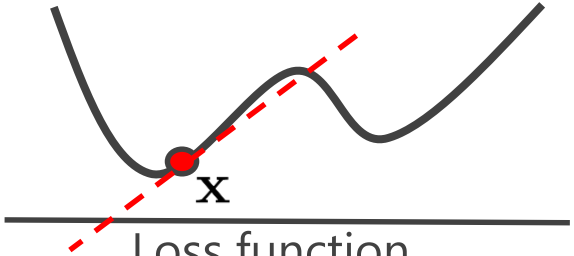
$$\max_{\Delta} \ell(f_\theta(\mathbf{x}+\Delta), y)$$

$$\|\Delta\|_2 \leq \rho$$

$$\max_{\Delta} \left[\overbrace{\mathbf{g}_{\mathbf{x}}^t \Delta}^{\text{First Order}} \right.$$

1. **Linear approximation** of the loss

Assumptions of Current Methods



Loss function

$$\ell(f_\theta(\mathbf{x}+\Delta), y) \approx \ell(f_\theta(\mathbf{x}), y) + \mathbf{g}_{\mathbf{x}}^t \Delta$$

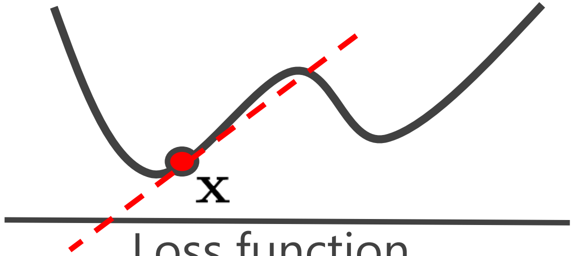
$$\max_{\Delta} \ell(f_\theta(\mathbf{x}+\Delta), y)$$

$$\|\Delta\|_2 \leq \rho$$

$$\max_{\Delta} \left[\overbrace{\mathbf{g}_{\mathbf{x}}^t \Delta}^{\text{First Order}} - \lambda_2 \|\Delta\|_2^2 \right]$$

1. **Linear approximation** of the loss
2. **Isolated features**: perturb $\mathbf{x}$(i) keeping all other features fixed

Assumptions of Current Methods



Loss function

$$\ell(f_\theta(\mathbf{x}+\Delta),y) \approx \ell(f_\theta(\mathbf{x}),y) + \mathbf{g}_\mathbf{x}^t\Delta$$

$$\max_{\Delta} \ell(f_\theta(\mathbf{x}+\Delta),y)$$

$$\|\Delta\|_2 \leq \rho$$

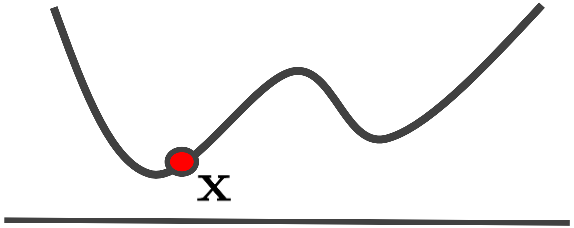
$$\max_{\Delta} \left[\overbrace{\mathbf{g}_\mathbf{x}^t\Delta}^{\text{First Order}} - \lambda_2\|\Delta\|_2^2 \right]$$

$$\boxed{\Longrightarrow \Delta^* = c\ \mathbf{g}_\mathbf{x}}$$

1. **Linear approximation** of the loss
2. **Isolated features**: perturb $\mathbf{x}$(i) keeping all other features fixed

Desiderata of a New Interpretation Framework

Desiderata of a New Interpretation Framework

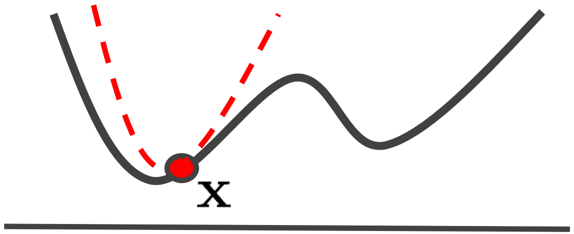


Loss function

$$\max_{\Delta} \ \ell(f_\theta(\mathbf{x} + \Delta), y)$$

$$\|\Delta\|_2 \leq \rho, \quad \underbrace{\|\Delta\|_0 \leq k}_{\text{Group Features}}$$

Desiderata of a New Interpretation Framework



Loss function

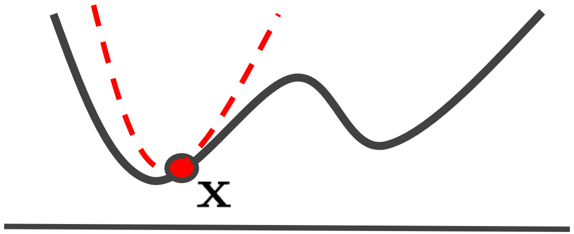
$$\ell(f_\theta(\mathbf{x}+\Delta), y) \approx \ell(f_\theta(\mathbf{x}), y) + \mathbf{g}_\mathbf{x}^t \Delta + \frac{1}{2}\Delta^t \mathbf{H}_\mathbf{x} \Delta$$

$$\max_{\Delta} \ \ell(f_\theta(\mathbf{x}+\Delta), y) \longrightarrow \max_{\Delta} \left[\overbrace{\mathbf{g}_\mathbf{x}^t \Delta + \frac{1}{2}\Delta^t \mathbf{H}_\mathbf{x} \Delta}^{\text{Second Order}} \right.$$

$$\|\Delta\|_2 \leq \rho, \quad \underbrace{\|\Delta\|_0 \leq k}_{\text{Group Features}}$$

1. **Quadratic approximation** of the loss

Desiderata of a New Interpretation Framework



Loss function

$$\ell(f_\theta(\mathbf{x}+\Delta),y) \approx \ell(f_\theta(\mathbf{x}),y) + \mathbf{g}_\mathbf{x}^t\Delta + \frac{1}{2}\Delta^t\mathbf{H}_\mathbf{x}\Delta$$

$$\max_{\Delta} \ \ell(f_\theta(\mathbf{x}+\Delta),y)$$

$$\|\Delta\|_2 \le \rho, \quad \underbrace{\|\Delta\|_0 \le k}_{\text{Group Features}}$$

$$\max_{\Delta} \ \Bigg[\overbrace{\mathbf{g}_\mathbf{x}^t\Delta + \frac{1}{2}\Delta^t\mathbf{H}_\mathbf{x}\Delta}^{\text{Second Order}} - \underbrace{\lambda_1\|\Delta\|_1}_{L_1 \ \text{relaxation}} - \lambda_2\|\Delta\|_2^2 \Bigg]$$

1. **Quadratic approximation** of the loss
2. **Group features:** find group of k pixels that maximizes the loss

Confronting the Second-Order term

Confronting the Second-Order term

- Optimization can be **non-concave maximization**

Confronting the Second-Order term

- Optimization can be **non-concave maximization**
- Hessian can be **VERY LARGE:** ~150k x 150k for 224 x 224 x 3 input

Confronting the Second-Order term

Concave for λ_2 > L/2 where L is the largest eigenvalue of $\mathbf{H_x}$

- Optimization can be **non-concave maximization**
- Hessian can be **VERY LARGE:** ~150k x 150k for 224 x 224 x 3 input

Confronting the Second-Order term

- Optimization can be **non-concave maximization**
- Hessian can be **VERY LARGE:** ~150k x 150k for 224 x 224 x 3 input

Concave for λ_2 > L/2 where L is the largest eigenvalue of $\mathbf{H_x}$

Can efficiently compute Hessian vector product

When Does Second-Order Matter?

When Does Second-Order Matter?

For a **deep ReLU** network:

- **Theorem:** $$\boxed{\mathbf{H}_{\mathbf{x}} = \mathbf{W}(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T)\mathbf{W}^T}$$

When Does Second-Order Matter?

For a **deep ReLU** network:

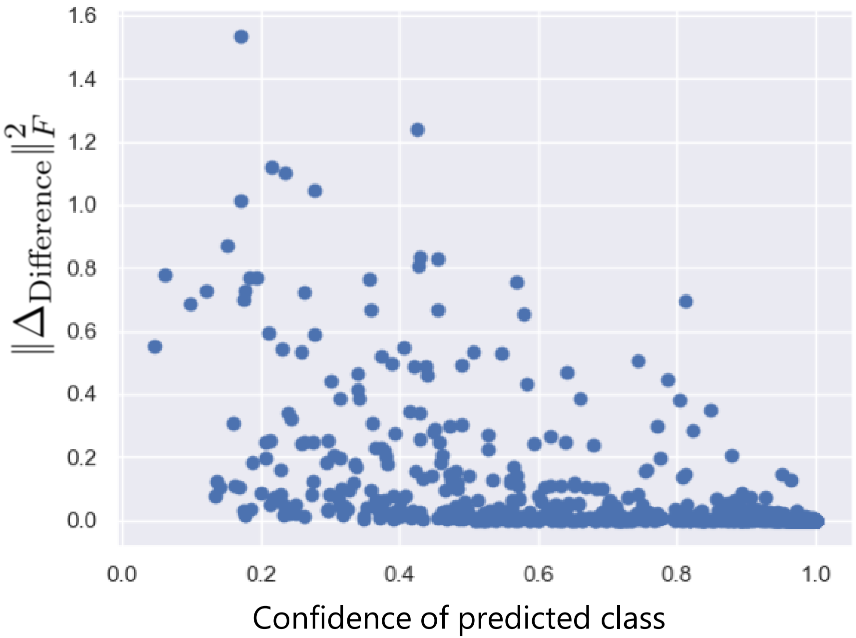
- **Theorem:** $$\boxed{\mathbf{H_x} = \mathbf{W}(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T)\mathbf{W}^T}$$

- **Theorem:** If the probability of the predicted class is close to one and the number of classes is large:

$$\Delta^* \approx c\,\mathbf{g_x} \quad \Longrightarrow \quad \textbf{Second-Order} \approx \textbf{First-Order}$$

Empirical results on the impact of Hessian

Empirical results on the impact of Hessian



RESNET-50 (uses only **ReLU**)

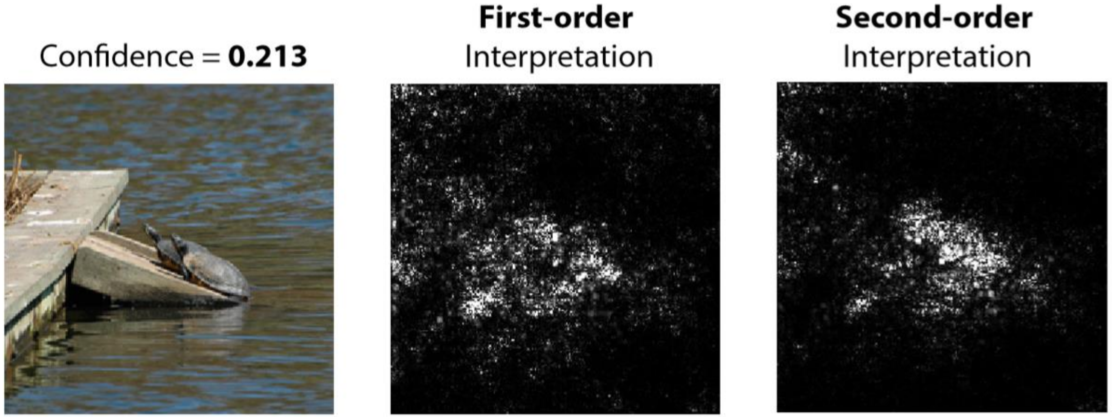
Empirical results on the impact of Hessian

RESNET-50 (uses only **ReLU**)

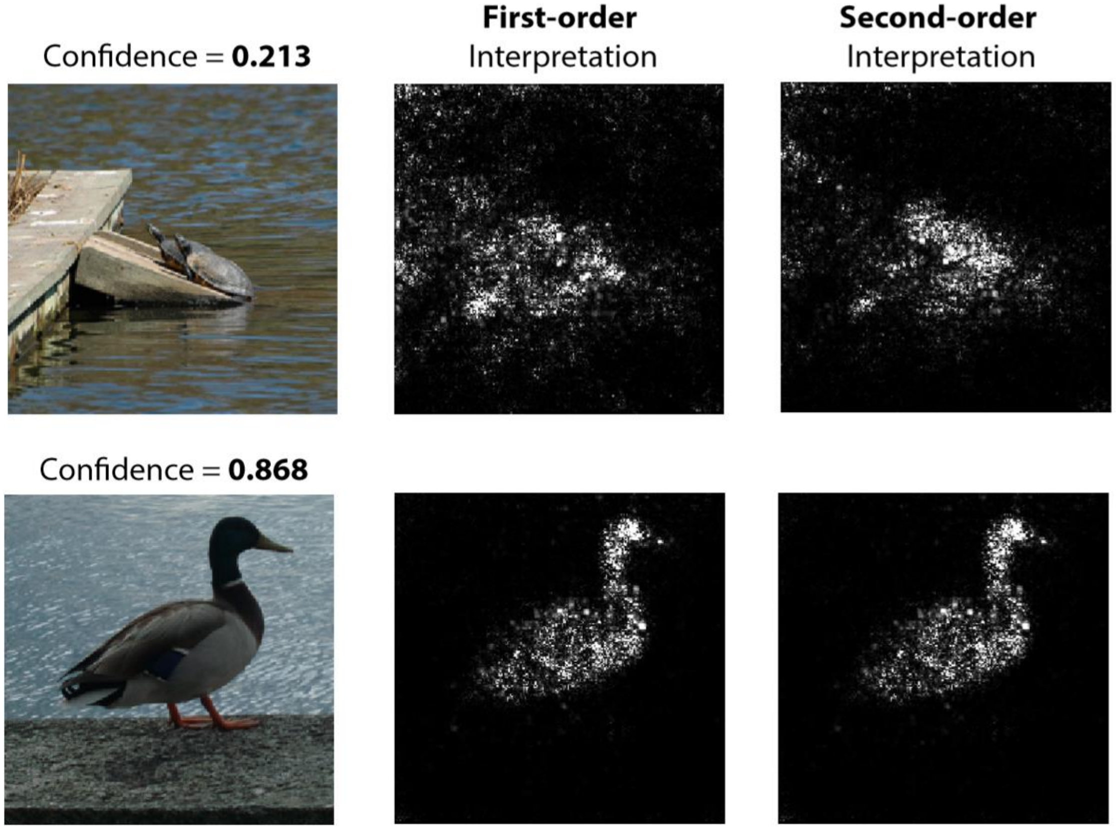
SE-RESNET-50 (uses **Sigmoid**)

Second-Order vs First Order (qualitative)

Second-Order vs First Order (qualitative)



Second-Order vs First Order (qualitative)



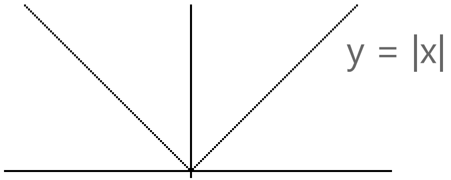
Confronting the L_1 term

Confronting the L_1 term

$$\max_{\Delta}\left[\underbrace{\mathbf{g}_{\mathbf{x}}^{t}\Delta + \frac{1}{2}\Delta^{t}\mathbf{H}_{\mathbf{x}}\Delta - \lambda_2\|\Delta\|_2}_{Smooth} - \underbrace{\lambda_1\|\Delta\|_1}_{Non\text{-}Smooth}\right]$$

Confronting the L_1 term

$$\max_{\Delta}\left[\underbrace{\mathbf{g}_{\mathbf{x}}^{t}\Delta + \frac{1}{2}\Delta^{t}\mathbf{H}_{\mathbf{x}}\Delta - \lambda_2\|\Delta\|_2}_{Smooth} - \underbrace{\lambda_1\|\Delta\|_1}_{Non\text{-}Smooth}\right]$$

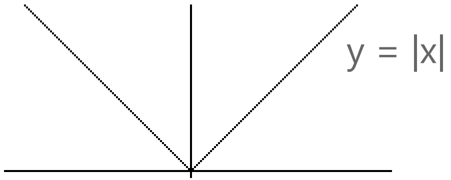


Not smooth at 0

- $\|\Delta\|_1$ term is non-smooth

Confronting the L_1 term

$$\max_{\Delta}\left[\underbrace{\mathbf{g}_{\mathbf{x}}^{t}\Delta+\frac{1}{2}\Delta^{t}\mathbf{H}_{\mathbf{x}}\Delta-\lambda_2\|\Delta\|_2}_{Smooth}-\underbrace{\lambda_1\|\Delta\|_1}_{Non-Smooth}\right]$$

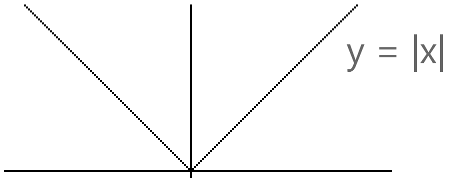


Not smooth at 0

- $\|\Delta\|_1$ term is non-smooth
- How to select λ_1 ?

Confronting the L_1 term

$$\max_{\Delta}\left[\underbrace{\mathbf{g}_{\mathbf{x}}^t\Delta + \frac{1}{2}\Delta^t\mathbf{H}_{\mathbf{x}}\Delta - \lambda_2\|\Delta\|_2}_{Smooth} - \underbrace{\lambda_1\|\Delta\|_1}_{Non\text{-}Smooth}\right]$$



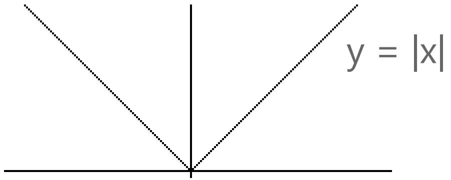
Not smooth at 0

- $\|\Delta\|_1$ term is non-smooth
- How to select λ_1 ?

Use proximal gradient descent to optimize the objective.

Confronting the L_1 term

$$\max_{\Delta}\left[\underbrace{\mathbf{g}_{\mathbf{x}}^{t}\Delta+\frac{1}{2}\Delta^{t}\mathbf{H}_{\mathbf{x}}\Delta-\lambda_2\|\Delta\|_2}_{Smooth}-\underbrace{\lambda_1\|\Delta\|_1}_{Non-Smooth}\right]$$



Not smooth at 0

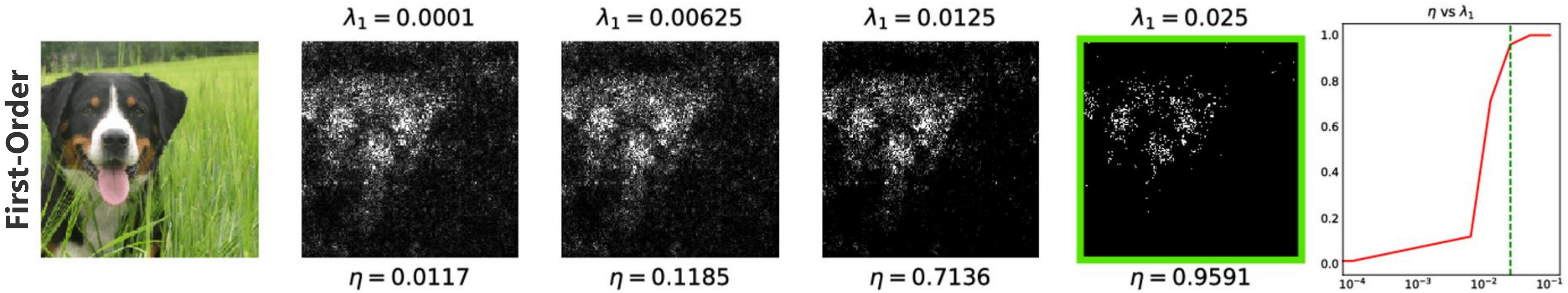
- $\|\Delta\|_1$ term is non-smooth
- How to select λ_1 ?

Use proximal gradient descent to optimize the objective.

Select the λ_1 value that induces sparsity within a range (0.75, 1).

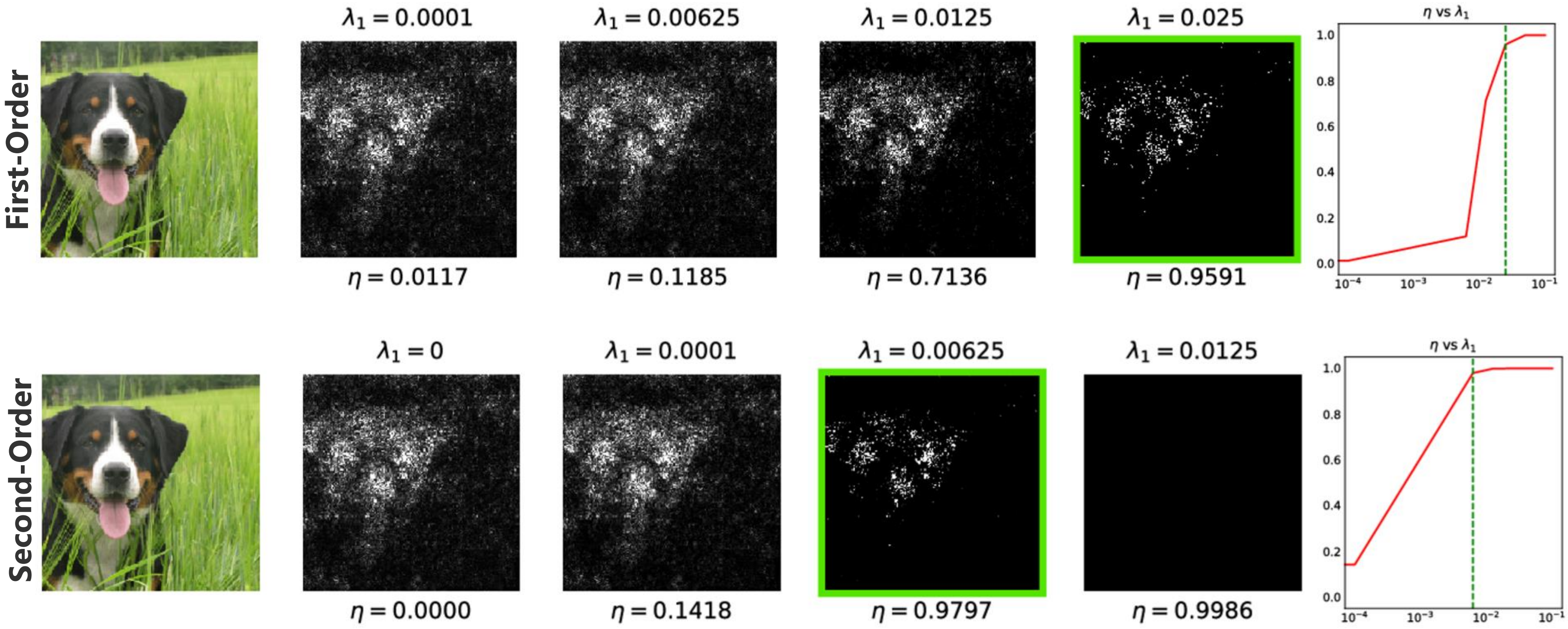
Impact of Group Features

Impact of Group Features



η denotes sparsity

Impact of Group Features



η denotes sparsity

Conclusions

- A new formulation for interpretation
  - Second-Order information
  - Group Features
- Efficient Computation

Pacific Ballroom **#69**, 6:30-9:00 PM
https://github.com/singlasahil14/CASO