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

\[
\max_{\Delta} \ell(f_\theta(x + \Delta), y)
\]

\[
\|\Delta\|_2 \leq \rho
\]
Assumptions of Current Methods

\[ \ell(f_\theta(x + \Delta), y) \approx \ell(f_\theta(x), y) + g_x^t \Delta \]

\[ \max_{\Delta} \ell(f_\theta(x + \Delta), y) \]

\[ \|\Delta\|_2 \leq \rho \]

1. **Linear approximation** of the loss
Assumptions of Current Methods

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

\[
\ell(f_\theta(x + \Delta), y) \approx \ell(f_\theta(x), y) + g^t_x \Delta
\]

\[
\max_\Delta \ell(f_\theta(x + \Delta), y) \leq \rho
\]

\[
\max_\Delta \left[ g^t_x \Delta - \lambda_2 \|\Delta\|_2^2 \right]
\]
Assumptions of Current Methods

1. **Linear approximation** of the loss

2. **Isolated features**: perturb $x(i)$ keeping all other features fixed

\[
\ell(f_\theta(x + \Delta), y) \approx \ell(f_\theta(x), y) + g_x^t \Delta
\]

\[
\max_{\Delta} \ell(f_\theta(x + \Delta), y)
\]

\[
\|\Delta\|_2 \leq \rho
\]

\[
\max_{\Delta} \left[ \underbrace{g_x^t \Delta} \right] - \lambda_2 \|\Delta\|_2^2
\]

\[
\Rightarrow \Delta^* = c \ g_x
\]
Desiderata of a New Interpretation Framework
Desiderata of a New Interpretation Framework

\[ \max_{\Delta} \ell(f_{\theta}(x + \Delta), y) \]
\[ \|\Delta\|_2 \leq \rho, \quad \|\Delta\|_0 \leq k \]

Group Features
**Desiderata of a New Interpretation Framework**

1. **Quadratic approximation** of the loss

\[
\ell(f_\theta(x + \Delta), y) \approx \ell(f_\theta(x), y) + g^t_x \Delta + \frac{1}{2} \Delta^t H_x \Delta
\]

\[
\max_{\Delta} \ell(f_\theta(x + \Delta), y) \quad \text{subject to} \quad \|\Delta\|_2 \leq \rho, \quad \|\Delta\|_0 \leq k
\]

**Group Features**

**Second Order**

\[
\max_{\Delta} \left[ g^t_x \Delta + \frac{1}{2} \Delta^t H_x \Delta \right]
\]
Desiderata of a New Interpretation Framework

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

\[
\ell(f_{\theta}(x + \Delta), y) \approx \ell(f_{\theta}(x), y) + g_x^T \Delta + \frac{1}{2} \Delta^T H_x \Delta
\]

Maximize

\[
\max_{\Delta} \left[ g_x^T \Delta + \frac{1}{2} \Delta^T H_x \Delta \right]
\]

Subject to

\[
\|\Delta\|_2 \leq \rho, \quad \|\Delta\|_0 \leq k
\]

Group Features
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:**
  \(~150k \times 150k\) for \(224 \times 224 \times 3\) 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** \( \lambda_2 > L/2 \) where \( L \) is the largest eigenvalue of \( H_x \)
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 \( \lambda_2 > L/2 \) where L is the largest eigenvalue of \( H_x \)

Can efficiently compute Hessian vector product
When Does Second-Order Matter?
When Does Second-Order Matter?

For a deep ReLU network:

- Theorem: \[ H_x = W (\text{diag}(p) - pp^T) W^T \]
When Does Second-Order Matter?

For a **deep ReLU** network:

- **Theorem:**
  \[ H_x = W(\text{diag}(p) - pp^T)W^T \]

- **Theorem:** If the probability of the predicted class is close to one and the number of classes is large:
  \[ \Delta^* \approx c \cdot g_x \quad \implies \quad \text{Second-Order} \approx \text{First-Order} \]
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)

Confidence = 0.213

First-order Interpretation

Second-order Interpretation
Second-Order vs First Order (qualitative)

Confidence = 0.213

Confidence = 0.868
Confronting the $L_1$ term
Confronting the $L_1$ term

$$\max_{\Delta} \left[ g_x^t \Delta + \frac{1}{2} \Delta^t H_x \Delta - \lambda_2 \| \Delta \|_2 - \lambda_1 \| \Delta \|_1 \right]$$

- **Smooth**
- **Non-Smooth**
Confronting the L₁ term

\[
\max_{\Delta} \left[ g_x^t \Delta + \frac{1}{2} \Delta^t H_x \Delta - \lambda_2 \|\Delta\|_2 - \lambda_1 \|\Delta\|_1 \right]
\]

- \(\|\Delta\|_1\) term is non-smooth
Confronting the L₁ term

\[
\max_{\Delta} \left[ g_x^t \Delta + \frac{1}{2} \Delta^t H_x \Delta - \lambda_2 \| \Delta \|^2_2 - \lambda_1 \| \Delta \|^1_1 \right]
\]

- \( \| \Delta \|^1_1 \) term is non-smooth
- How to select \( \lambda_1 \)?

Not smooth at 0

\( y = |x| \)
Confronting the $L_1$ term

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

- $\|\Delta\|_1$ term is non-smooth
- How to select $\lambda_1$?

Use **proximal gradient descent** to optimize the objective.
Confronting the $L_1$ term

$$\max_{\Delta} \left[ g_x^t \Delta + \frac{1}{2} \Delta^t H_x \Delta - \lambda_2 \|\Delta\|_2 - \begin{cases} \lambda_1 \|\Delta\|_1 & \text{Smooth} \\ \text{Non-Smooth} & \end{cases} \right]$$

- $\|\Delta\|_1$ term is non-smooth
- How to select $\lambda_1$?

Use proximal gradient descent to optimize the objective.

Select the $\lambda_1$ value that induces sparsity within a range $(0.75, 1)$. 

Not smooth at 0

$y = |x|$
Impact of Group Features
Impact of Group Features

$\lambda_1 = 0.0001$  $\lambda_1 = 0.00625$  $\lambda_1 = 0.0125$  $\lambda_1 = 0.025$

$\eta = 0.0117$  $\eta = 0.1185$  $\eta = 0.7136$  $\eta = 0.9591$

$\eta$ denotes sparsity
Impact of Group Features

First-Order

- $\lambda_1 = 0.0001$, $\eta = 0.0117$
- $\lambda_1 = 0.00625$, $\eta = 0.1185$
- $\lambda_1 = 0.0125$, $\eta = 0.7136$
- $\lambda_1 = 0.025$, $\eta = 0.9591$

Second-Order

- $\lambda_1 = 0$, $\eta = 0.0000$
- $\lambda_1 = 0.0001$, $\eta = 0.1418$
- $\lambda_1 = 0.00625$, $\eta = 0.9797$
- $\lambda_1 = 0.0125$, $\eta = 0.9986$

$\eta$ 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