Analyzing the impact of Baselines in Integrated Gradients method for BERT-based text classification task

Chuhuan Shen

Leibniz Universität Hannover

Content

- Introduction
- Highlights
- Implement and Results
- Discussion

Introduction

- Neural Networks are differentiable, and the output can be written as a function of the parameters and input.
- The gradient can be used for *Sensitivity Analysis*: How sensitive is the output $f(\cdot)$ w.r.t to a small change in the input x ? $\frac{\partial f(x;\theta)}{\partial x}$
- The Vanilla Gradient method suffers from *saturation* problem: Gradients of input features may have small magnitudes around a sample even if the network depends heavily on those features.
- Improved Approach ---- Integrated Gradients

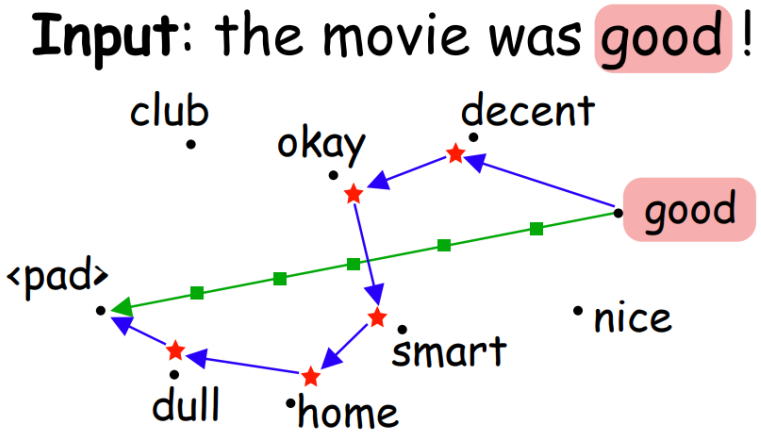
Introduction

$$IG_i(x, x') = (x_i - x'_i) \cdot \int_0^1 \frac{\partial F(x'_i + \alpha \cdot (x_i - x'_i))}{\partial x_i} d\alpha$$

- x is input
- x' is baseline
- $\frac{\partial F(x)}{\partial x_i}$ is the gradient of F along the i^{th} dimension at x.
- Path $\gamma(a) = x' + \alpha(x - x')$ for $\alpha \in [0, 1]$
- **Interpolation along linear path**

Introduction

Discretized Integrated Gradients (DIG)



- Linear interpolated points are not necessarily representative of the discrete word embedding distribution
- Nonlinear interpolation paths

Highlights

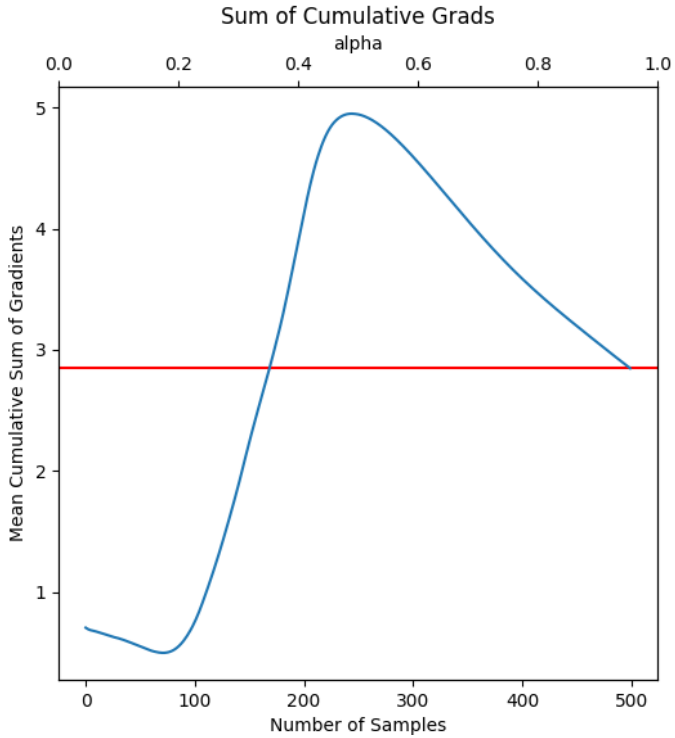
- Investigate the impact of different baselines and integration paths on the results
- Explore the distinction between text task and image task

Inplements and Results

- The Zero Baseline
- The Constant Baseline
- The Maximum Distance Baseline
- The Blurred Baseline
- The Uniform Baseline
- Results on DIG

The Zero Baseline

- Zero represents black in image
- [PAD] is just an empty tokens



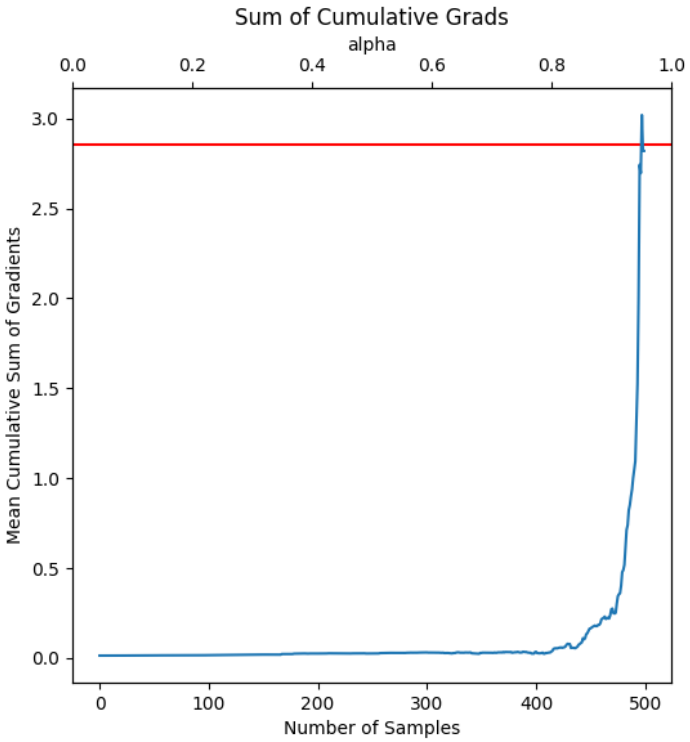
Legend: Negative Neutral Positive

Baseline	[CLS]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[SEP]
α=0.2	[CLS]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[SEP]
α=0.4	[CLS]	[unused814]	[PAD]	[PAD]	[PAD]	[unused814]	[PAD]	[unused155]	[PAD]	[SEP]
α=0.6	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]
α=0.8	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]

The Zero Baseline

DIG(greedy)

- Original baseline in the paper
- Also, the only convergent baseline in DIG

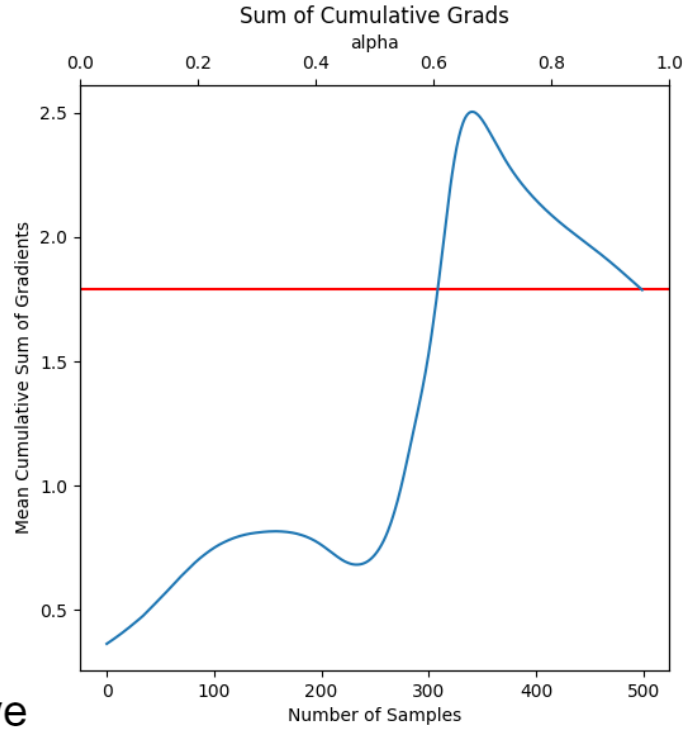


Legend: Negative Neutral Positive

Baseline	[CLS]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[SEP]
α=0.2	[CLS]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[SEP]
α=0.4	[CLS]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[SEP]
α=0.6	[CLS]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[SEP]
α=0.8	[CLS]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[SEP]
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]

The Constant Baseline

- Randomly select a token from the input sentence
- Like constant color in the image
- Blind to the baseline token



Legend: ■ Negative □ Neutral ■ Positive

Baseline	[CLS]	##fl	##fl	##fl	##fl	##fl	##fl	##fl	##fl	[SEP]
α=0.2	[CLS]	##fl	##fl	##fl	##fl	##fl	##fl	##fl	##fl	[SEP]
α=0.4	[CLS]	##fl	##fl	##fl	##fl	##fl	##fl	##fl	##fl	[SEP]
α=0.6	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]
α=0.8	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]

The Constant Baseline

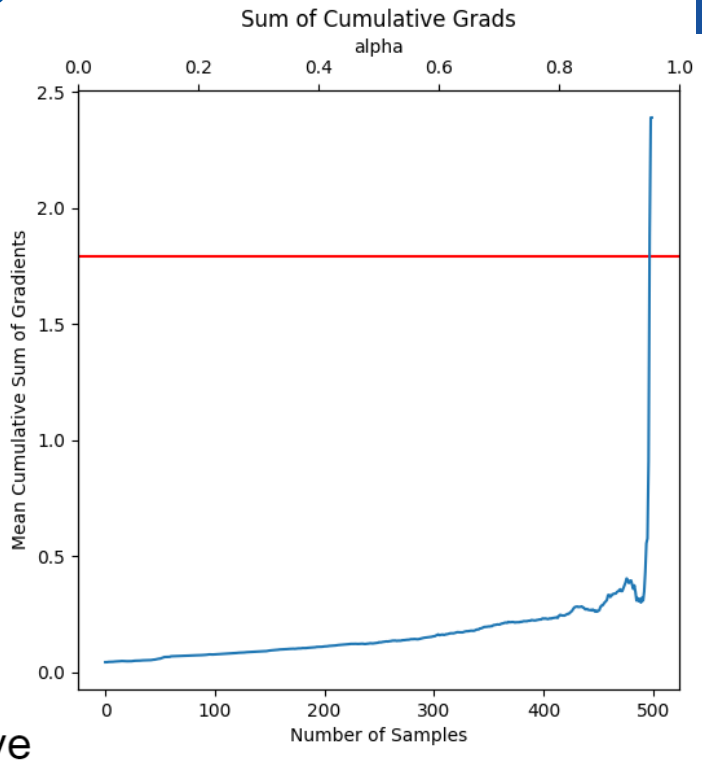
DIG(greedy)

- Not satisfy the axiom

  completeness

$$\Sigma_{i=1}^{n} \text{IntegratedGrads}_i(x) = F(x) - F(x')$$

- Not converged
- Maybe not enough samples

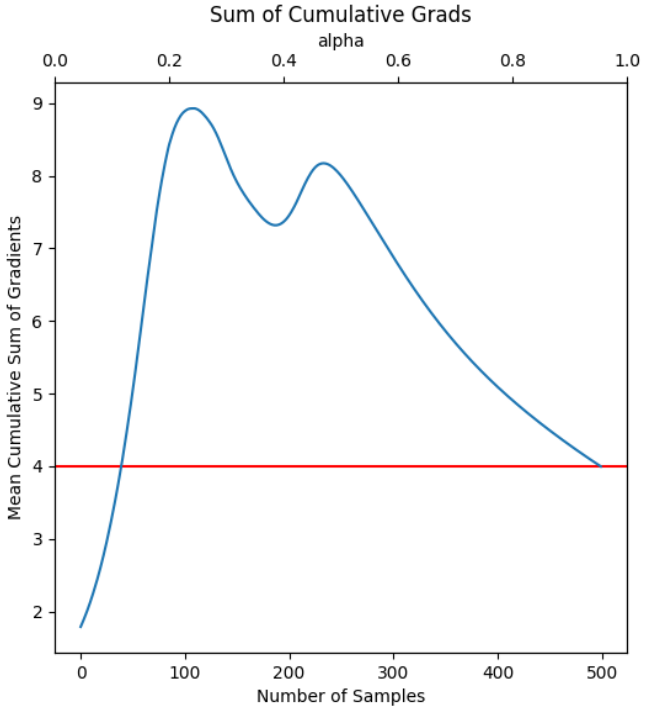


Legend: Negative Neutral Positive

Baseline	[CLS]	##fl	##fl	##fl	##fl	##fl	##fl	##fl	[SEP]
α=0.2	[CLS]	##fl	##fl	##fl	##fl	##fl	##fl	##fl	[SEP]
α=0.4	[CLS]	##fl	##fl	##fl	##fl	##fl	##fl	##fl	[SEP]
α=0.6	[CLS]	##fl	##fl	##fl	##fl	##fl	##fl	##fl	[SEP]
α=0.8	[CLS]	##fl	##fl	##fl	##fl	##fl	##fl	##fl	[SEP]
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]

The Maximum Distance Baseline

- k-nearest neighbors
- Euclidean distance
- 500 nearest neighbors for each token



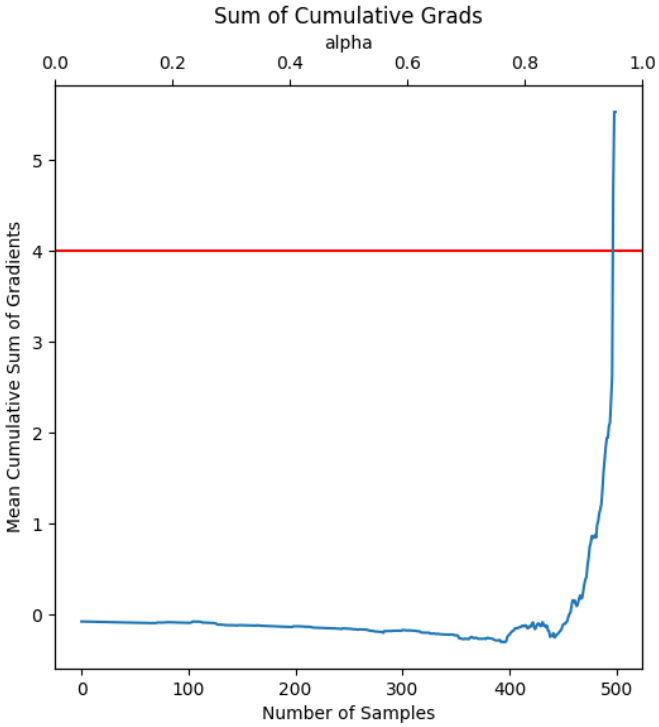
Legend: Negative Neutral Positive

Baseline	[CLS]	helen	[unused558]	##ma	[unused651]	seventh	उ	1998	[unused668]	[SEP]
α=0.2	[CLS]	helen	[unused558]	##ma	[unused651]	seventh	उ	1998	[unused668]	[SEP]
α=0.4	[CLS]	helen	[unused558]	##ma	[unused651]	seventh	उ	1998	[unused814]	[SEP]
α=0.6	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]
α=0.8	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]

The Maximum Distance Baseline

DIG(greedy)

- Not converged
- Interpolation changes significantly only at big α

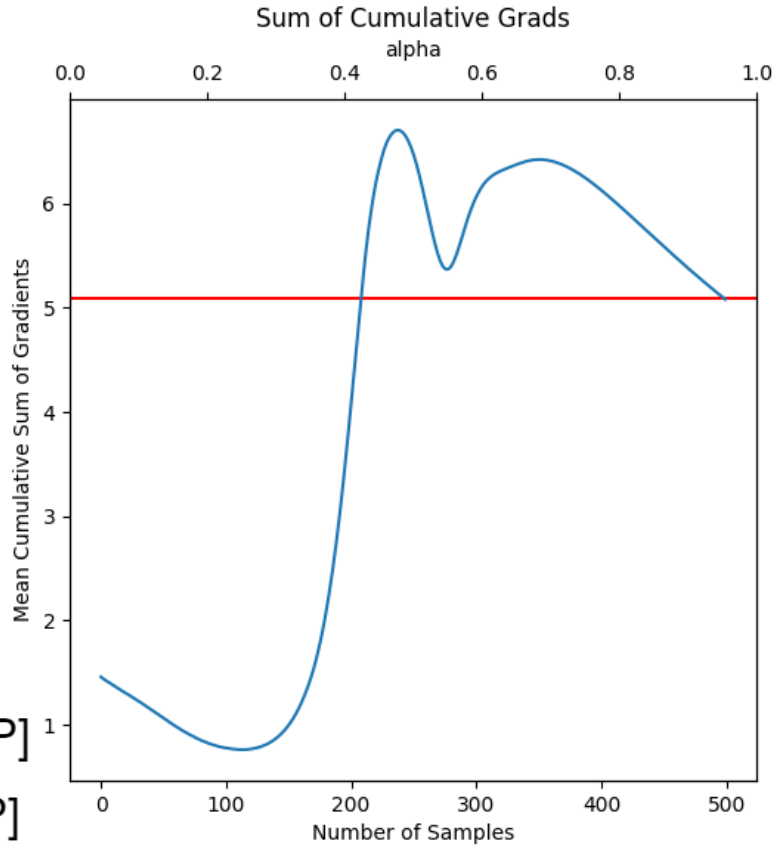


Legend: Negative Neutral Positive

Baseline	[CLS]	helen	[unused558]	##ma	[unused651]	seventh	ड	1998	[unused668]	[SEP]	
α=0.2	[CLS]	helen	[unused558]	##ma	[unused651]	seventh	ड	1998	[unused668]	[SEP]	
α=0.4	[CLS]	helen	[unused558]	##ma	[unused651]	seventh	ड	1998	[unused668]	[SEP]	
α=0.6	[CLS]	helen	[unused558]	##ma	[unused651]	seventh	ड	1998	[unused668]	[SEP]	
α=0.8	[CLS]	helen	[unused558]	##ma	[unused651]	seventh	ड	1998	[unused668]	[SEP]	
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]	

The Blurred Baseline

- How to blur text ?
  - Use DistilBertForMaskedLM to predict synonym
  - Replace the mask token with the predict token



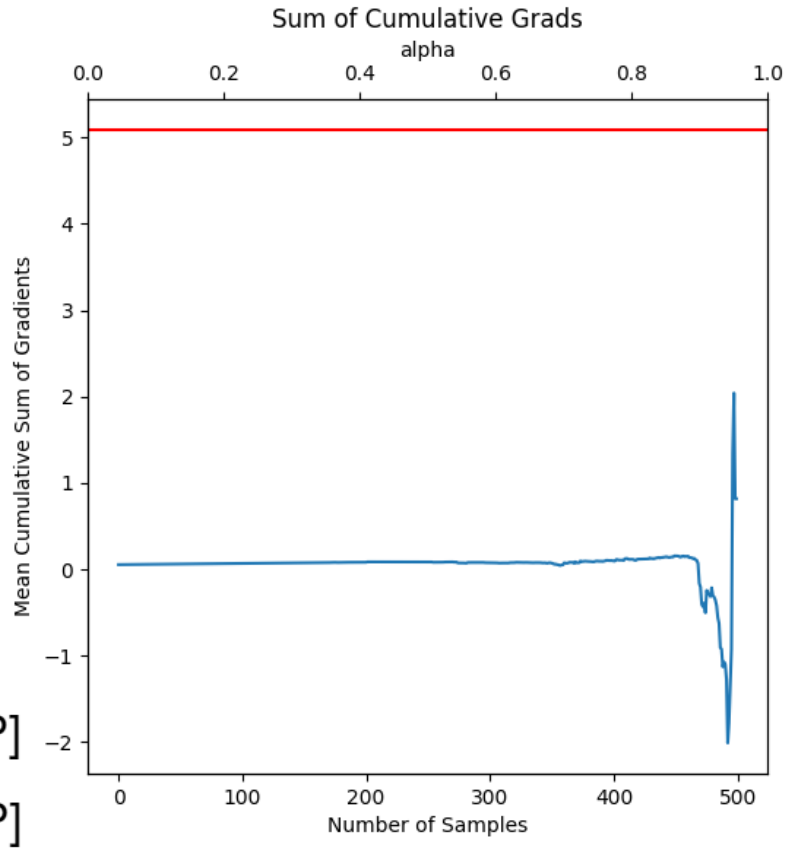
Legend: Negative Neutral Positive

Baseline	[CLS]	.	.	.	ing	.	.	-	.	[SEP]
α=0.2	[CLS]	.	.	.	ing	.	.	-	.	[SEP]
α=0.4	[CLS]	.	.	.	ing	.	.	-	.	[SEP]
α=0.6	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]
α=0.8	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]

The Blurred Baseline

DIG(greedy)

- Not converged



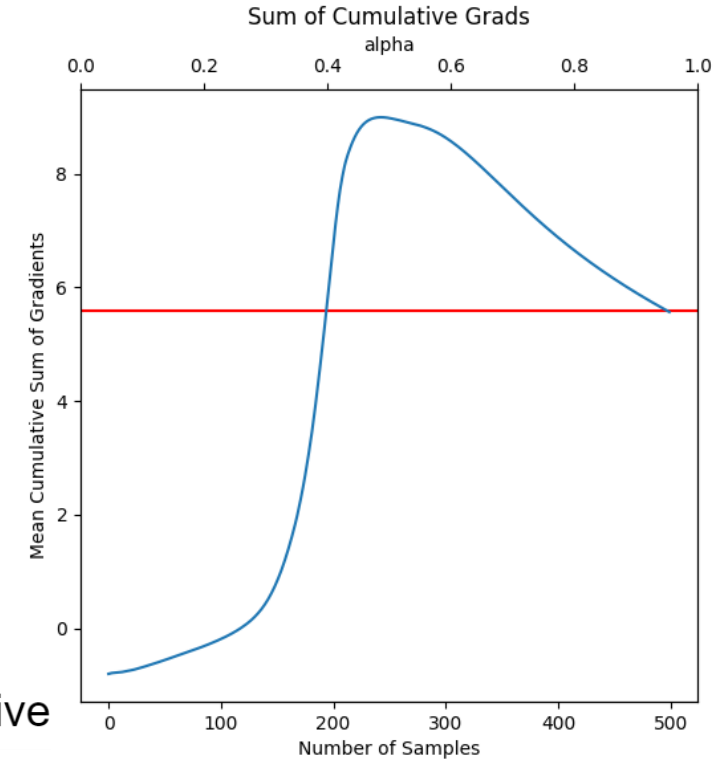
Legend: Negative Neutral Positive

Baseline	[CLS]	.	.	.	ing	.	.	-	.	[SEP]
α=0.2	[CLS]	.	.	.	ing	.	.	-	.	[SEP]
α=0.4	[CLS]	.	.	.	ing	.	.	-	.	[SEP]
α=0.6	[CLS]	.	.	.	ing	.	.	-	.	[SEP]
α=0.8	[CLS]	.	.	.	ing	.	.	-	.	[SEP]
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]

The Uniform Baseline

- Random sampling among

  500 neighbor tokens



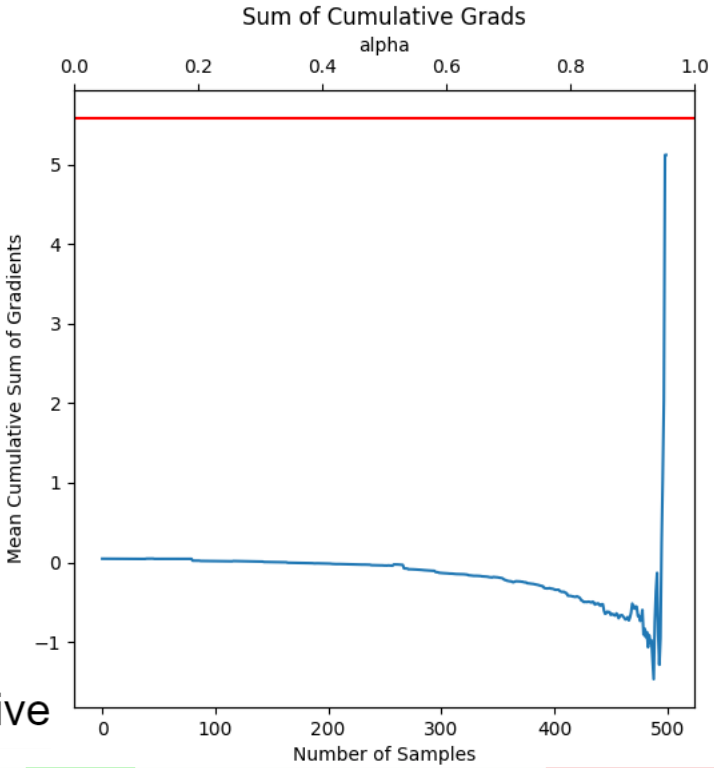
Legend: Negative Neutral Positive

Baseline	[CLS]	london	1947	21	##∂	in	1872	became	facilitate	[SEP]	
α=0.2	[CLS]	london	1947	21	##∂	in	1872	became	facilitate	[SEP]	
α=0.4	[CLS]	london	1947	21	##∂	in	1872	became	facilitate	[SEP]	
α=0.6	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]	
α=0.8	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]	
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]	

The Uniform Baseline

DIG(greedy)

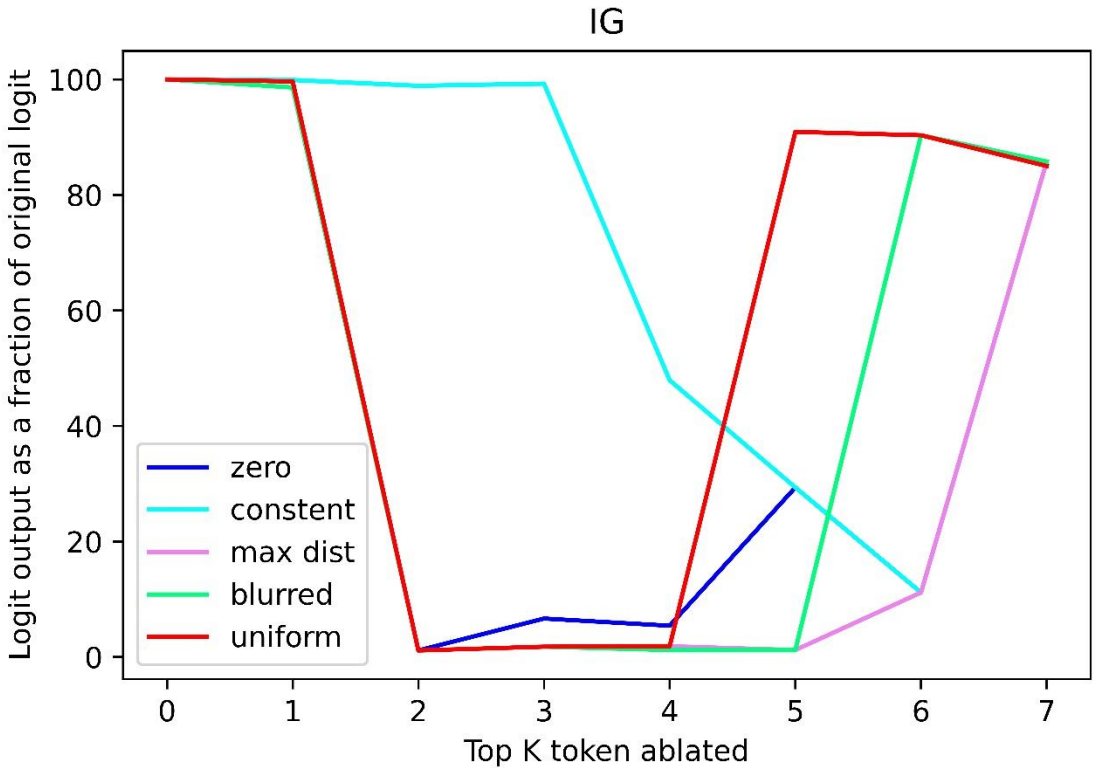
- Not converged
- High magnitude gradients accumulate at big values of α



Legend: Negative Neutral Positive

Baseline	[CLS]	london	1947	21	##∂	in	1872	became	facilitate	[SEP]	
α=0.2	[CLS]	london	1947	21	##∂	in	1872	became	facilitate	[SEP]	
α=0.4	[CLS]	london	1947	21	##∂	in	1872	became	facilitate	[SEP]	
α=0.6	[CLS]	london	1947	21	##∂	in	1872	became	facilitate	[SEP]	
α=0.8	[CLS]	london	1947	21	##∂	in	1872	became	facilitate	[SEP]	
Input	[CLS]	un	##fl	##in	##ching	##ly	bleak	and	desperate	[SEP]	

Results

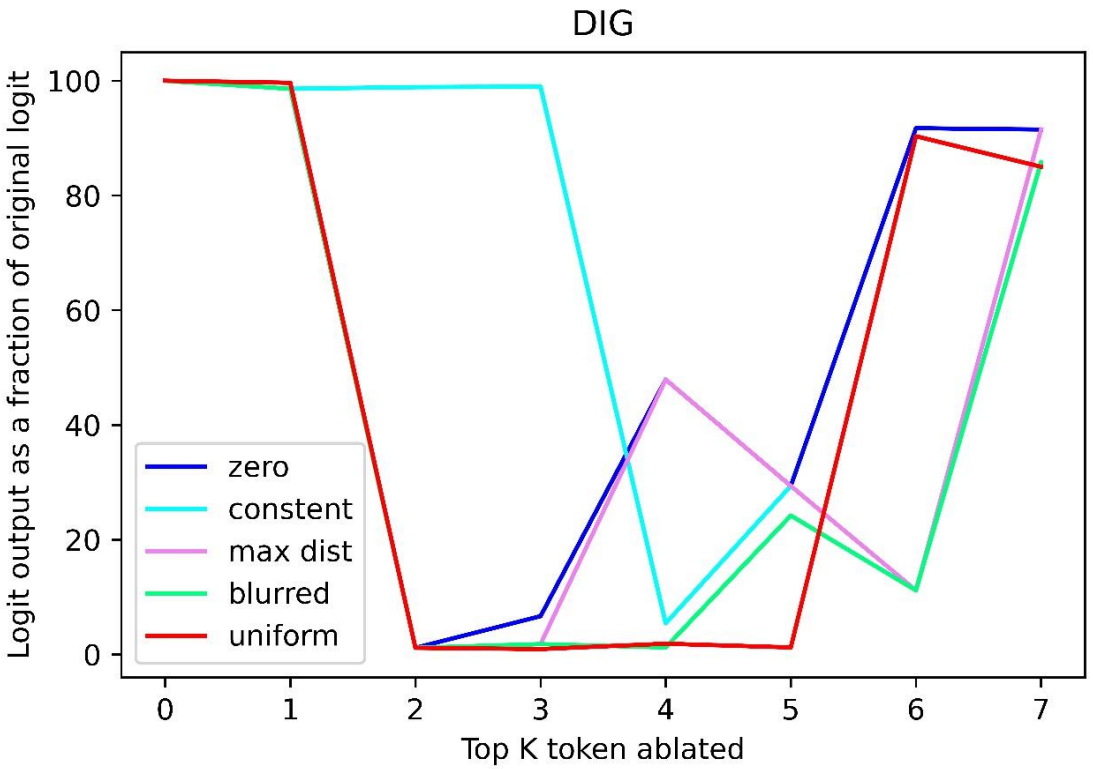


Order of token ablation

desperate	bleak	and	##ly	un	##fl	##ching	##in
and	bleak	un	desperate	##ly	##fl	##ching	##in
bleak	desperate	un	##ly	##fl	and	##ching	##in
bleak	desperate	un	##fl	##ly	##ching	and	##in
desperate	bleak	un	##ly	##ching	##fl	##in	and

'un', '##fl', '##in', '##ching', '##ly', 'bleak', 'and', 'desperate'

Results



Order of token ablation

desperate	bleak	and	un	##ly	##in	##fl	##ching
and	bleak	un	desperate	##ly	##fl	##ching	##in
bleak	desperate	un	##ly	##fl	and	##ching	##in
bleak	desperate	un	##fl	##ly	##ching	and	##in
desperate	bleak	un	##ly	##ching	##fl	##in	and

'un', '##fl', '##in', '##ching', '##ly', 'bleak', 'and', 'desperate'

Discussion

- The results of the integrated gradient depend heavily on the selection of path γ and baseline, which is *strongly artificial* and *not an exact verity result.*
- Identifying significant features alone may not be sufficient to understand model behavior. Interactions between features is also vital.

Leibniz Universität Hannover

Thanks for your attention