# Causal Intersectionality and Dual Form of Gradient Descent for Multimodal Analysis

: a Case Study on Hateful Memes

**Yosuke Miyanishi**, Nguyen Le Minh
Japan Advanced Institute of Science and Technology (JAIST)
LREC-COLING 2024

# Table of Contents

# 01

# Introduction

Our Motivation and Approach / Causality / Explainable AI (XAI) / Causality and XAI

# Our Motivation and Approach

| | |
|---|---|
| Key Question | • Is current machine learning framework sufficient for quantifying the model *understanding* of a task? |
| What We Do / What We Don't | • We *do not* create new models or datasets<br>• We propose the formal definition of a task<br>• We propose an analytical framework to quantify the relevance between the task's causal structure and inner workings of the SOTA models |

# Causality

- Causality describes the structure of the real world
- Two lines of causality science for graphical repersentation and quantification

Graphical Representation[*1]

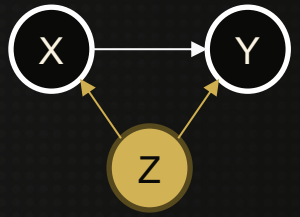

*1: Pearl, 2000 (10.1017/S0266466603004109)

*2: Rubin, 2008 (10.1214/08-AOAS187)

# Causality

- Causality describes the structure of the real world
- Two lines of causality science for graphical repersentation and quantification

Graphical Representation*1



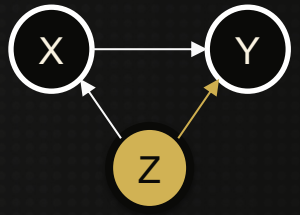

*1: Pearl, 2000 (10.1017/S0266466603004109)

*2: Rubin, 2008 (10.1214/08-AOAS187)

# Causality

- Causality describes the structure of the real world
- Two lines of causality science for graphical repersentation and quantification

Graphical Representation*1



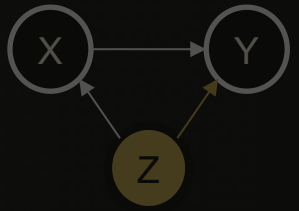

- Back Door Path Z->Y for X->Y Relationship

*1: Pearl, 2000 (10.1017/S0266466603004109)
*2: Rubin, 2008 (10.1214/08-AOAS187)

# Causality

- Causality describes the structure of the real world
- Two lines of causality science for graphical repersentation and quantification

## Graphical Representation[*1]




- Back Door Path Z->Y for X->Y Relationship

## Causality Metric[*2]

Hereafter
ATE

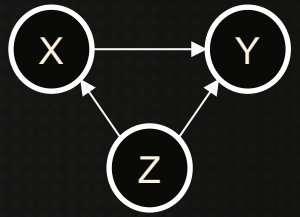
$$AverageTreatmentEffect = P(Y = 1|X = X_1) - P(Y = 0|X = X_1)$$

*1: Pearl, 2000 (10.1017/S0266466603004109)
*2: Rubin, 2008 (10.1214/08-AOAS187)

# Causality

- Causality describes the structure of the real world
- Two lines of causality science for graphical repersentation and quantification

## Graphical Representation[*1]




- Back Door Path Z->Y for X->Y Relationship

## Quantification[*2]

Hereafter ATE

$$AverageTreatmentEffect = P(Y = 1|X = X_1) - P(Y = 0|X = X_1)$$

How to relate the model inner working to the real-world causal structure?

*1: Pearl, 2000 (10.1017/S0266466603004109)
*2: Rubin, 2008 (10.1214/08-AOAS187)

# Explainable AI (XAI)

- XAI describes how the model reacts to the world
- Centric to XAI is gradient-based method

Example in Computer Vision*1

How can XAI method describe the understanding of the causality?

*1: Selvaraju et al., 2017 (10.1109/ICCV.2017.74)

# 02

# Related Work

Vision and Language (VL) / Causality in VL / Causal Intersectionality / Attention Attribution Score / Large Language Models (LLMs)

# Vision and Language (VL)

- Unique interplay of two modalities (text and image)

Visual Question Answering[*1]

Who is wearing glasses?

*1: Goyal et al., 2017 (10.1109/CVPR.2017.670)

*2: Kiela et al., 2020

# Vision and Language (VL)

- Unique interplay of two modalities (text and image)
- Strong interplay in hateful memes

## Visual Question Answering[*1]

Who is wearing glasses?

## Benign / Hateful Memes[*2]

Benign *Image*


Benign *Text*


*1: Goyal et al., 2017 (10.1109/CVPR.2017.670)
*2: Kiela et al., 2020 (10.48550/arXiv.2005.04790)

# Vision and Language (VL)

- Unique interplay of two modalities (text and image)
- Strong interplay in hateful memes

Visual Question Answering[*1]

Who is wearing glasses?

Benign / Hateful Memes[*2]

Benign *Image*


Benign *Text*



Hateful Text / Image

*1: Goyal et al., 2017 (10.1109/CVPR.2017.670)
*2: Kiela et al., 2020 (10.48550/arXiv.2005.04790)

# Vision and Language (VL)

- Unique interplay of two modalities (text and image)
- Strong interplay in hateful memes

## Visual Question Answering[*1]

Who is wearing glasses?

## Benign / Hateful Memes[*2]

Benign *Image*


Benign *Text*



Hateful Text / Image

We focus on the study of hateful memes

*1: Goyal et al., 2017 (10.1109/CVPR.2017.670)
*2: Kiela et al., 2020 (10.48550/arXiv.2005.04790)

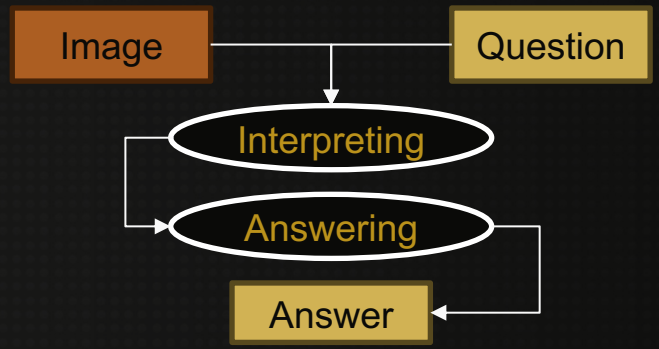
# Causality in VL

- Causality **of** / *to* the model has been tested for Visual Question Answering (VQA)

# Causality in VL

- Causality **of** / *to* the model has been tested for Visual Question Answering (VQA)

Bi-Phasic Cognitive VQA Model[*1]




- Causality *of* the model

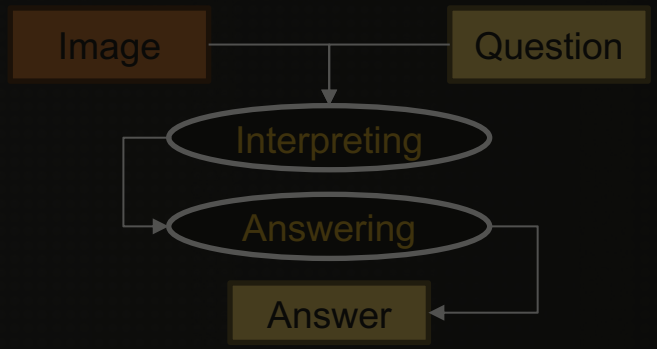
*1: Nguyen and Okazaki, 2023 (10.18653/v1/2023.emnlp-main.573)

*2: Niu et al., 2021 (10.1109/CVPR46437.2021.01254)

# Causality in VL

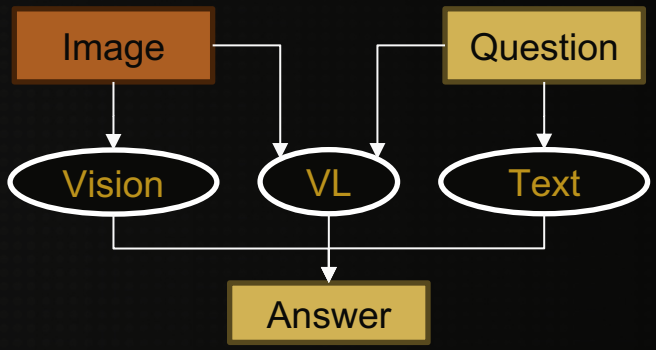
- Causality **of** / *to* the model has been tested for Visual Question Answering (VQA)

Bi-Phasic Cognitive VQA Model*1




- Causality *of* the model

Bias of Images in VQA*2

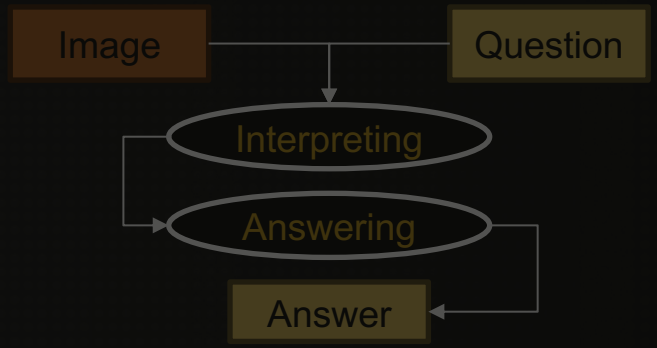



*1: Nguyen and Okazaki, 2023 (10.18653/v1/2023.emnlp-main.573)
*2: Niu et al., 2021 (10.1109/CVPR46437.2021.01251)

# Causality in VL

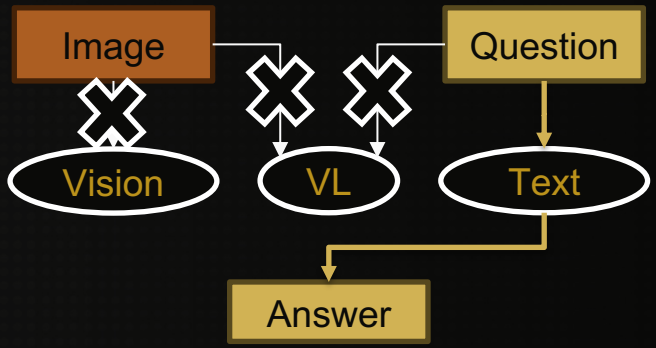
- Causality **of** / *to* the model has been tested for Visual Question Answering (VQA)

Causal Reasoning of VQA Model*1




- Causality *of* the model

Bias of Images in VQA*2

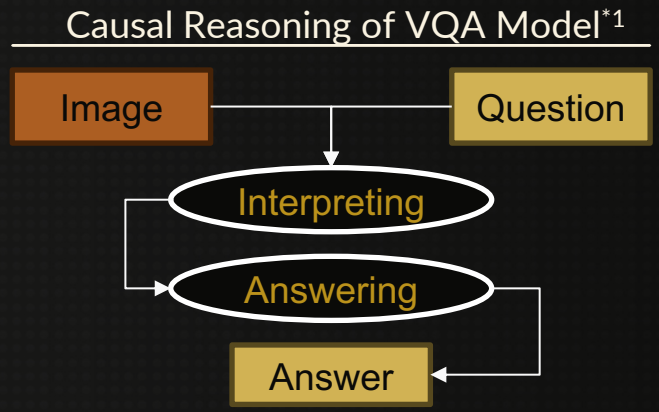


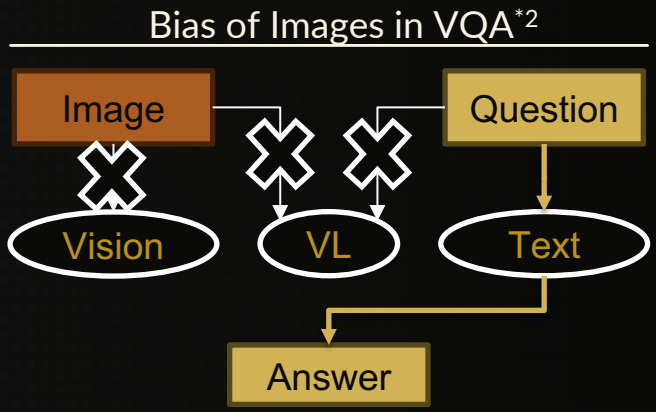

- Causality *to* the model

*1: Nguyen and Okazaki, 2023 (10.18653/v1/2023.emnlp-main.573)
*2: Niu et al., 2021 (10.1109/CVPR46437.2021.01251)

# Causality in VL

- Causality **of** / *to* the model has been tested for Visual Question Answering (VQA)




- Causality *of* the model

- Causality *to* the model

How can XAI method describe the understanding of the arbitrary model?

*1: Nguyen and Okazaki, 2023 (10.18653/v1/2023.emnlp-main.573)
*2: Niu et al., 2021 (10.1109/CVPR46437.2021.01251)

# Causal Intersectionality

- Interplay of two social categories (e.g. colour and gender)
- Interpreted as a form of indirect causal effect

## Definition*1

*the network of connections between social categories such as race, class, and gender, especially when this may result in additional disadvantage or discrimination*

## Intersectionality as Indirect Effect*2

$$\theta_{Z_1,Z_2} \neq \theta_{Z_1} + \theta_{Z_2}$$

*1: Definition in Oxford Dictionary. Image from demontattler.com
*2: Bright et al., 2016 (10.1086/684173)

# Causal Intersectionality

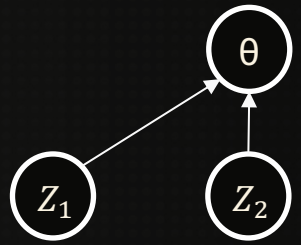
- Interplay of two social categories (e.g. colour and gender)
- Interpreted as a form of indirect causal effect

## Definition*1

*the network of connections between social categories such as race, class, and gender, especially when this may result in additional disadvantage or discrimination*

## Intersectionality as Indirect Effect*2

$$\theta_{Z_1,Z_2} \neq \theta_{Z_1} + \theta_{Z_2}$$




*1: Definition in Oxford Dictionary. Image from demontattler.com
*2: Bright et al., 2016 (10.1086/684173)

# Causal Intersectionality

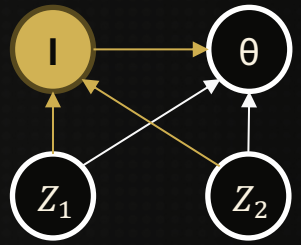
- Interplay of two social categories (e.g. colour and gender)
- Interpreted as a form of indirect causal effect

## Definition[*1]

*the network of connections between social categories such as race, class, and gender, especially when this may result in additional disadvantage or discrimination*

## Intersectionality as **Indirect Effect**[*2]

$$\theta_{Z_1,Z_2} \neq \theta_{Z_1} + \theta_{Z_2}$$




*1: Definition in Oxford Dictionary. Image from demontattler.com
*2: Bright et al., 2016 (10.1086/684173)

# Causal Intersectionality

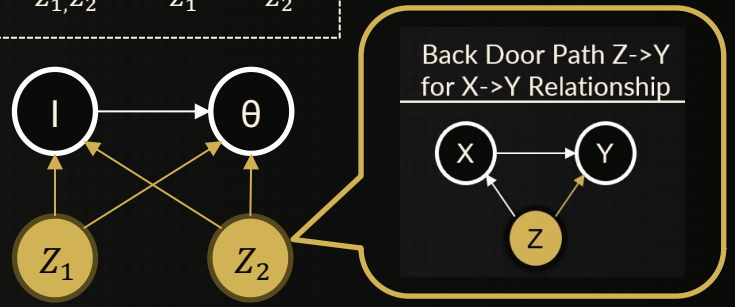
- Interplay of two social categories (e.g. colour and gender)
- Interpreted as a form of indirect causal effect

## Definition*1

*the network of connections between social categories such as race, class, and gender, especially when this may result in additional disadvantage or discrimination*

## Intersectionality as Indirect Effect*2

$$\theta_{Z_1,Z_2} \neq \theta_{Z_1} + \theta_{Z_2}$$




We expand to analyze multimodal indirect effect

*1: Definition in Oxford Dictionary. Image from demontattler.com
*2: Bright et al., 2016 (10.1086/684173)

# Gradient-Based Method in NLP

- Attention Attribution Score[*1]: attention importance to the model prediction

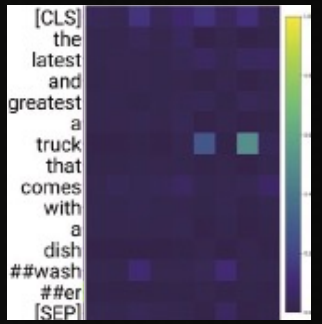
$$Attr(A) = A * \int_{\alpha=0}^{1} \frac{\partial \theta(\alpha A)}{\partial A} d\alpha$$

*1: Hao et al., 2021 (10.1609/aaai.v35i14.17533)
*2: Sundararajan et al., 2017 (10.5555/3305890.3306024)

# Gradient-Based Method in NLP

- Attention Attribution Score[*1]: attention importance to the model prediction
- Hee et al.[*3] showed modality attribution

$$Attr(A) = A * \int_{\alpha=0}^{1} \frac{\partial \theta(\alpha A)}{\partial A} d\alpha$$

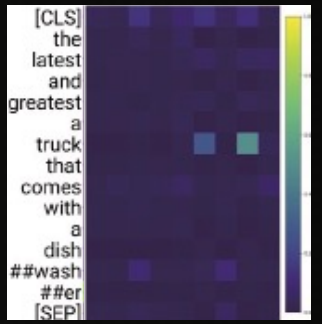
| | Image | Text |
|---|---|---|
| Attribution[*3] of VisualBERT[*4] |   |   |
| Averaged Gradient | 4.096±0.774 | 3.183±0.900 |

*1: Hao et al., 2021 (10.1609/aaai.v35i14.17533)
*2: Sundararajan et al., 2017 (10.5555/3305890.3306024)
*3: Hee et al., 2022 (10.1145/3485447.3512260)
*4: Li et al., 2020 (10.18653/v1/2020.acl-main.469)

# Gradient-Based Method in NLP

- Attention Attribution Score[*1]: attention importance to the model prediction
- Hee et al.[*3] showed modality attribution

$$Attr(A) = A * \int_{\alpha=0}^{1} \frac{\partial \theta(\alpha A)}{\partial A} d\alpha$$

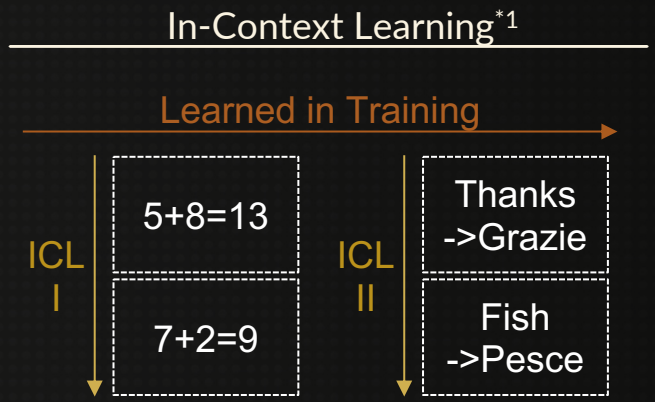
| | Image | Text |
|---|---|---|
| Attribution[*3] of VisualBERT[*4] |   |   |
| Averaged Gradient | 4.096±0.774 | 3.183±0.900 |

We show 1) cross-modal interaction 2) relationship to causal effect

*1: Hao et al., 2021 (10.1609/aaai.v35i14.17533)
*2: Sundararajan et al., 2017 (10.5555/3305890.3306024)
*3: Hee et al., 2022 (10.1145/3485447.3512260)
*4: Li et al., 2020 (10.18653/v1/2020.acl-main.469)

# Large Language Models (LLMs)

- Notable performance on gradient-free in-context learning (ICL) setting

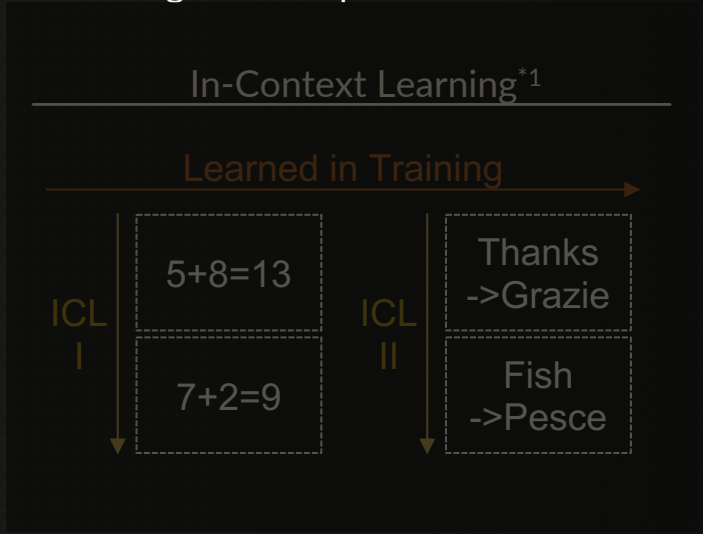



*1: Brown et al., 2020 (10.5555/3495724.3495883)

*2: Dai et al., 2023 (10.18653/v1/2023.findings-acl.247)

# Large Language Models (LLMs)

- Notable performance on gradient-free in-context learning (ICL) setting
- Meta-gradient/optimization: attention weights as the second form of gradient




Equivalence of Fine-Tuning (FT) and ICL in Linear Attention (LA)*2

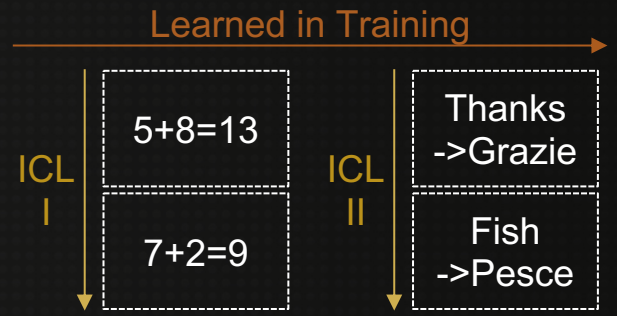
| | |
|---|---|
| LA | $\theta = A * Q$ |
| FT | $\theta = (A + \Delta A_{FT}) * Q$ |
| ICL | $\theta = (A + \Delta A_{ICL}) * Q$ |

*1: Brown et al., 2020 (10.5555/3495724.3495883)
*2: Dai et al., 2023 (10.18653/v1/2023.findings-acl.247)

# Large Language Models (LLMs)

- Notable performance on gradient-free in-context learning (ICL) setting
- Meta-gradient/optimization: attention weights as the second form of gradient

In-Context Learning[*1]




Equivalence of Fine-Tuning (FT) and ICL in Linear Attention (LA)[*2]

| | |
|---|---|
| LA | $\theta = A * Q$ |
| FT | $\theta = (A + \Delta A_{FT}) * Q$ |
| ICL | $\theta = (A + \Delta A_{ICL}) * Q$ |

We show how ICL and meta-gradient contributes to causal effect

*1: Brown et al., 2020 (10.5555/3495724.3495883)
*2: Dai et al., 2023 (10.18653/v1/2023.findings-acl.247)

# 03

# Experiment I

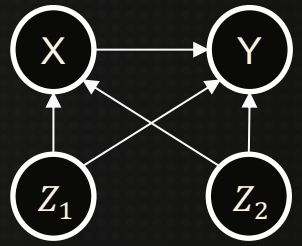
Metric for Causal Effect / Metric for Model Inner Workings / Experimental Settings / Results (miATE)

# Metric for Causal Effect

- Extended causal intersectionality for multimodal analysis

Causal Intersectionality / ATE

$$\theta_{Z_1,Z_2} \neq \theta_{Z_1} + \theta_{Z_2}$$



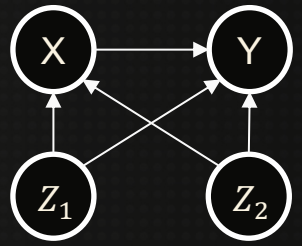

$$ATE = P(Y = 1|X = X_1) - P(Y = 0|X = X_1)$$

# Metric for Causal Effect

- Extended causal intersectionality for multimodal analysis

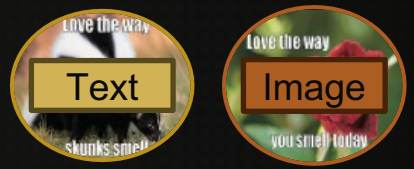
Causal Intersectionality / ATE

$$\theta_{Z_1,Z_2} \neq \theta_{Z_1} + \theta_{Z_2}$$




$$ATE = P(Y = 1|X = X_1) - P(Y = 0|X = X_1)$$

Multimodal Intersectional Average Treatment Effect (miATE)

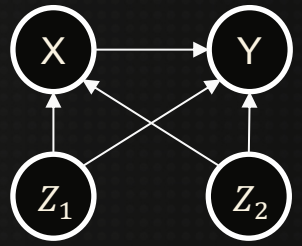


# Metric for Causal Effect

- Extended causal intersectionality for multimodal analysis

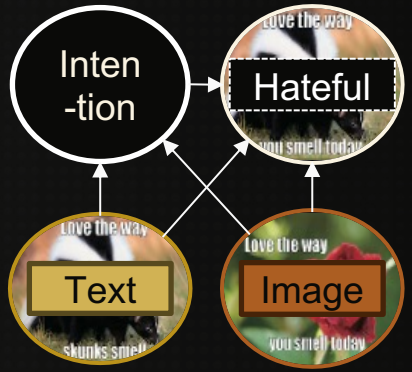
Causal Intersectionality / ATE

$$\theta_{Z_1,Z_2} \neq \theta_{Z_1} + \theta_{Z_2}$$




$$ATE = P(Y = 1|X = X_1) - P(Y = 0|X = X_1)$$

Multimodal Intersectional Average Treatment Effect (miATE)

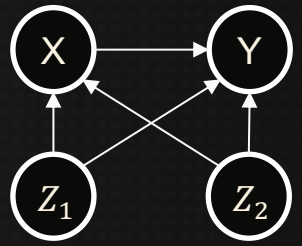


# Metric for Causal Effect

- Extended causal intersectionality for multimodal analysis

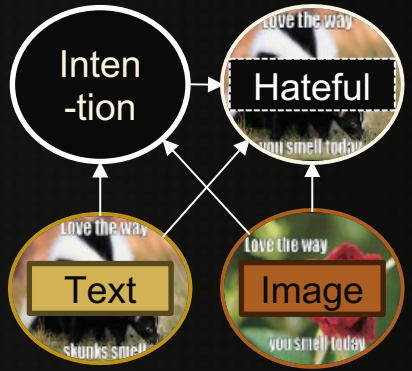
## Causal Intersectionality / ATE

$$\theta_{Z_1,Z_2} \neq \theta_{Z_1} + \theta_{Z_2}$$




$$ATE = P(Y = 1|X = X_1) - P(Y = 0|X = X_1)$$

## Multimodal Intersectional Average Treatment Effect (miATE)




$$\theta_{T,I} \neq \theta s_T + \theta_I$$

$$miATE = \theta_{T,I} - (\theta s_T + \theta_I)$$

# Metric for Model Inner Workings

- Modality Interaction Disentangled Attribution Score (MIDAS)

Attr and MIDAS

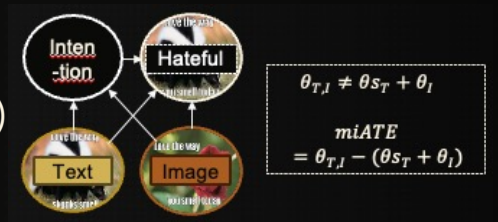
$$Attr(A) = A * \int_{\alpha=0}^{1} \frac{\partial \theta(\alpha A)}{\partial A} d\alpha$$

$$Interaction\ Type\ it \in \{text2text, image2image, cross\ modal\}$$
$$MIDAS^{it} = Attr^{it}(A_{T,I}) - (Attr^{it}(A_T) + Attr^{it}(A_I))$$

# Experimental Settings

**Data**

- Hateful memes challenge dataset (dev subset)
- Manually extracted hateful-benign pairs
  (78 hateful-image benign, 84 hateful-text benign)
- Added 3* manually crafted pairs
  (78*3=234 text benign, 84*3=252 image benign)




Analysis separated by original samples

**Model**

<Hateful Memes Detection>

- 3 differently pretrained BERT-based models in Vilio[*1]
  - Oscar[*2], VisualBERT , and UNITER[*3]

Similar trend to Oscar. Ommited from the presentation.

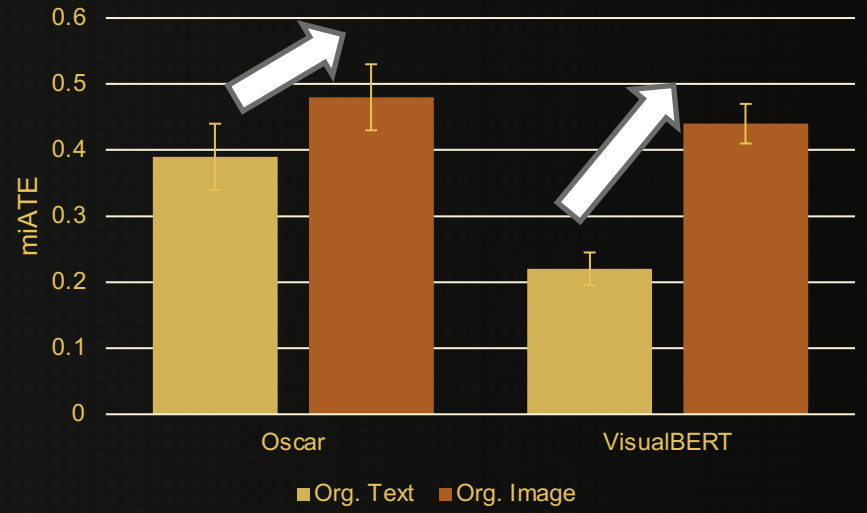
<Probing for MIDAS-miATE analysis>

- LightGBM[*4] with Optuna[*5] hyperparameter search

*1: Muennighoff, 2020 (arXiv:2012.07788)
*2: Sundararajan et al., 2017 (10.5555/3305890.3306024)
*3: Li et al., 2020 (10.1007/978-3-030-58577-8_8)
*4: Guolin et al., 2017 (10.5555/3294996.3295074)
*5: Akiba et al., 2019 (10.1145/3292500.3330701)

# Results (miATE)

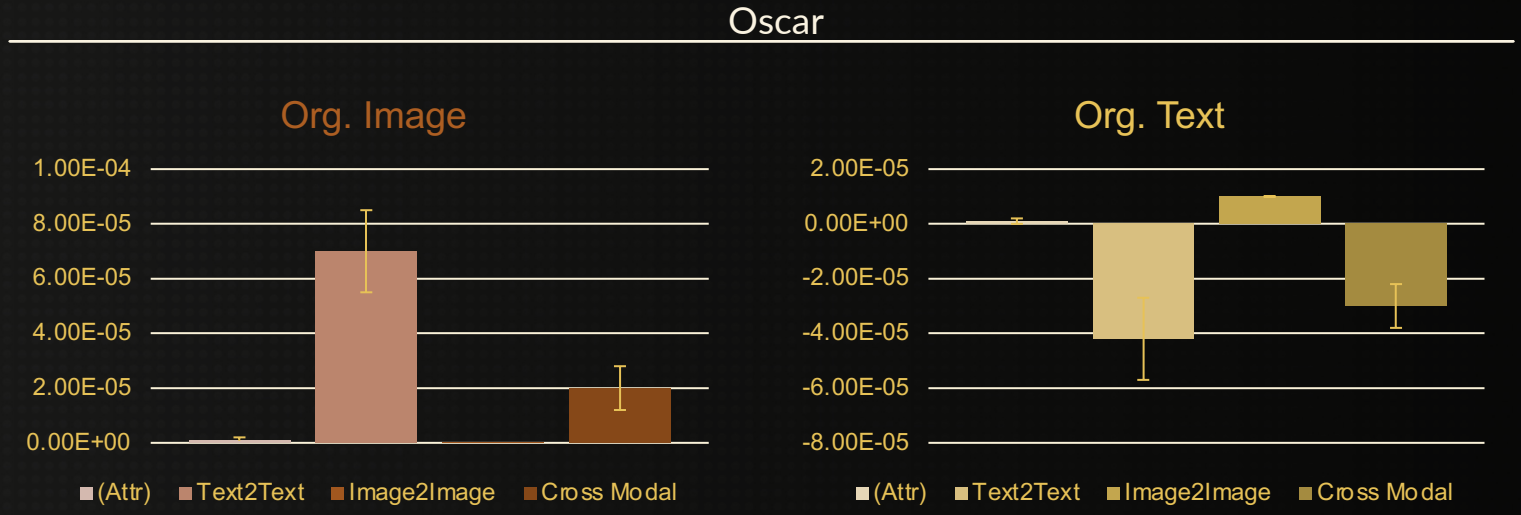
- Higher miATE in text-oriented task
- Most remarkable in VisualBERT




More sensitive to the difference of text?

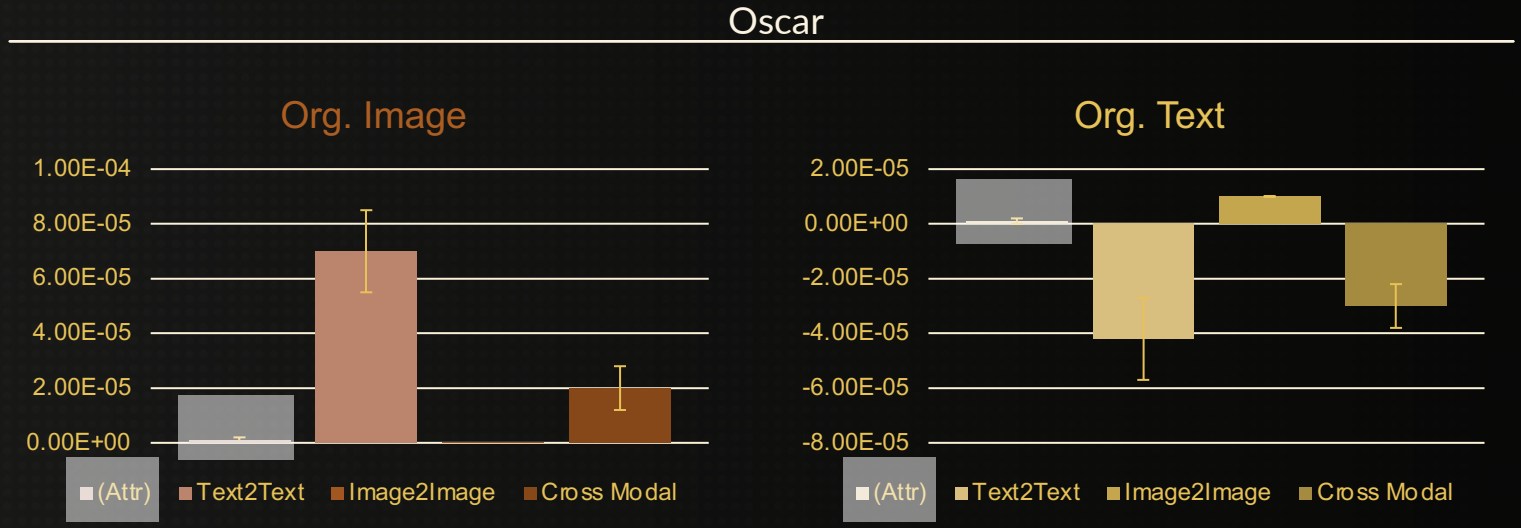
# Results (MIDAS)

- In contrast to Attr, captured the attention to the difference of modalities



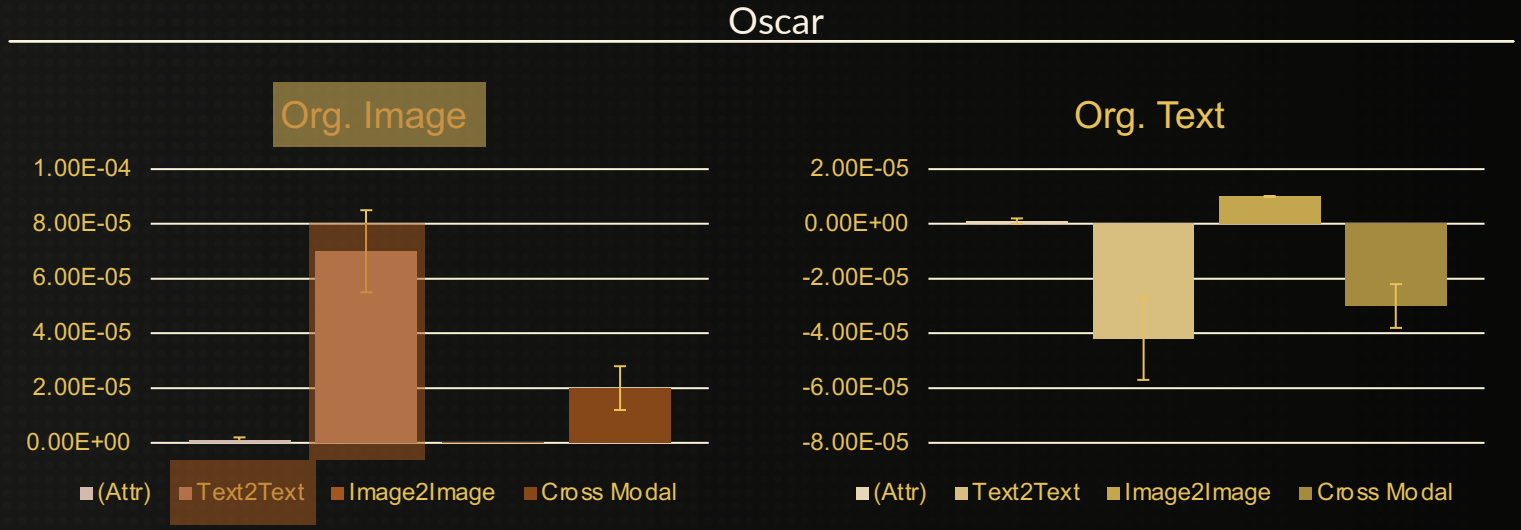
# Results (MIDAS)

- In contrast to Attr, captured the attention to the difference of modalities



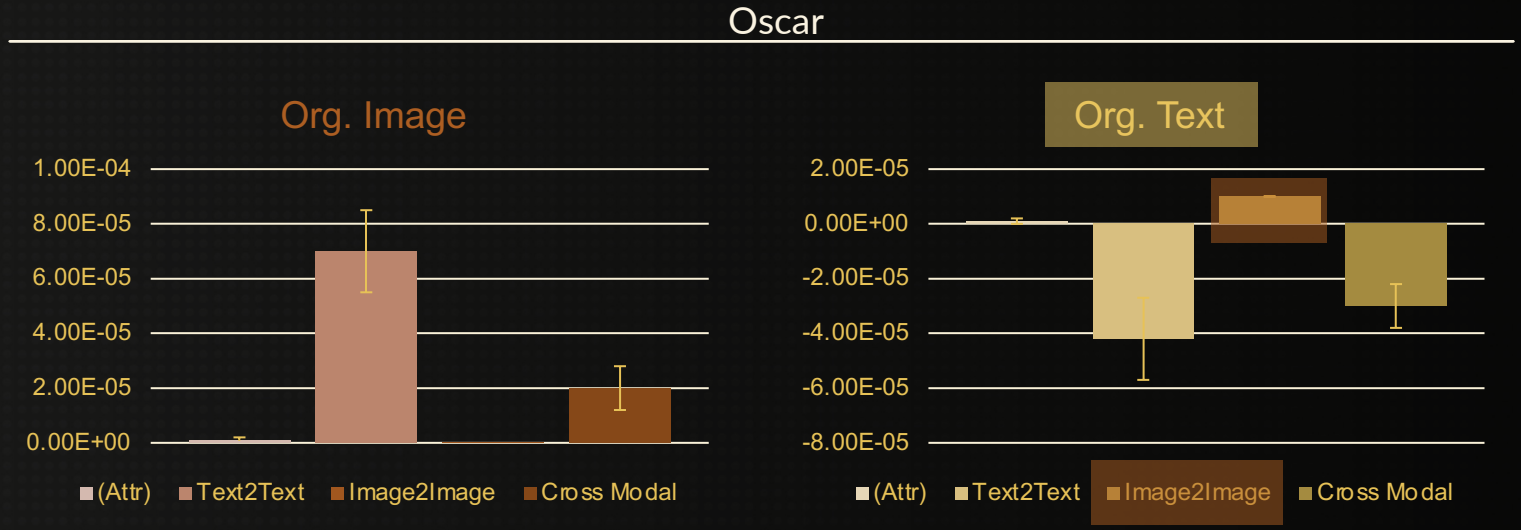
# Results (MIDAS)

- In contrast to Attr, captured the attention to the difference of modalities



# Results (MIDAS)

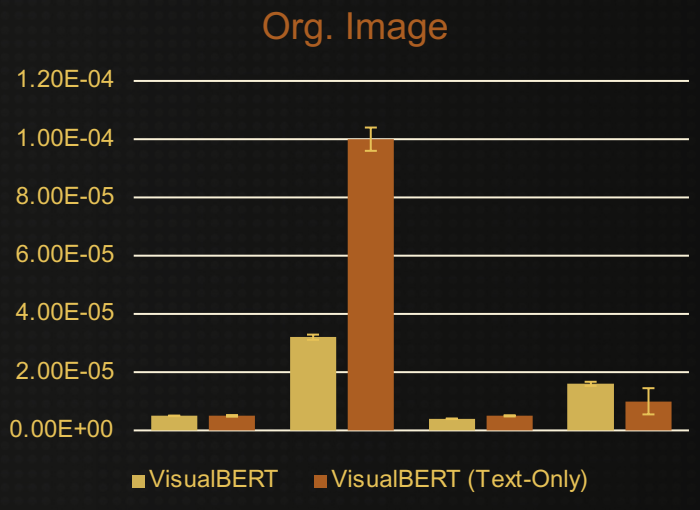
- In contrast to Attr, captured the attention to the difference of modalities

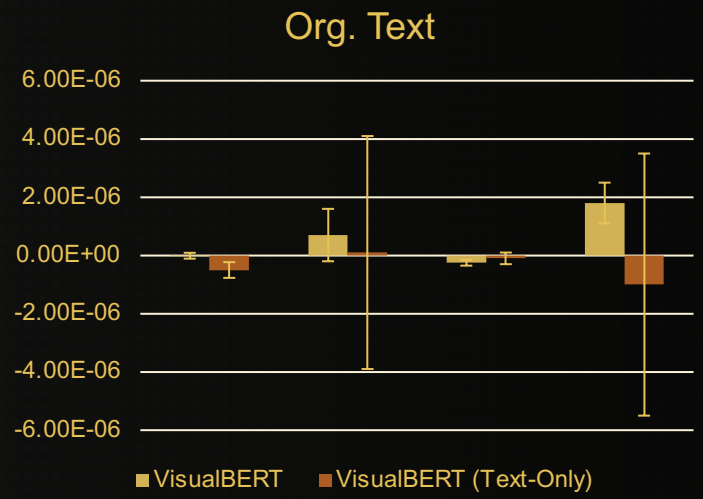


# Results (MIDAS)

- In contrast to Attr, captured the attention to the difference of modalities
- VisualBERT is biased toward textual information, enhanced by encoder

VisualBERT

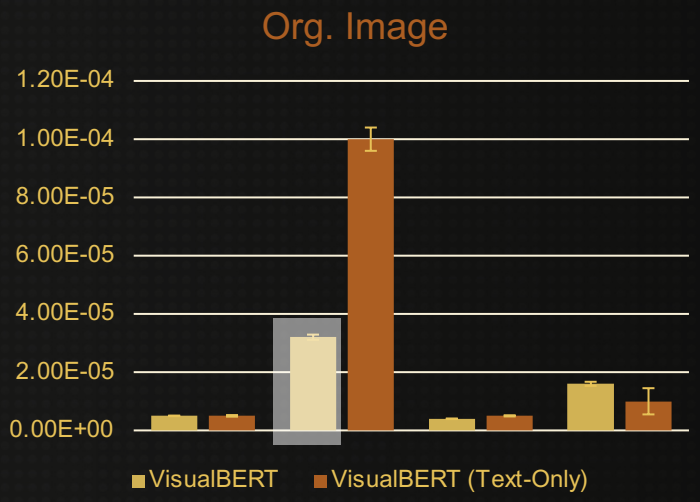


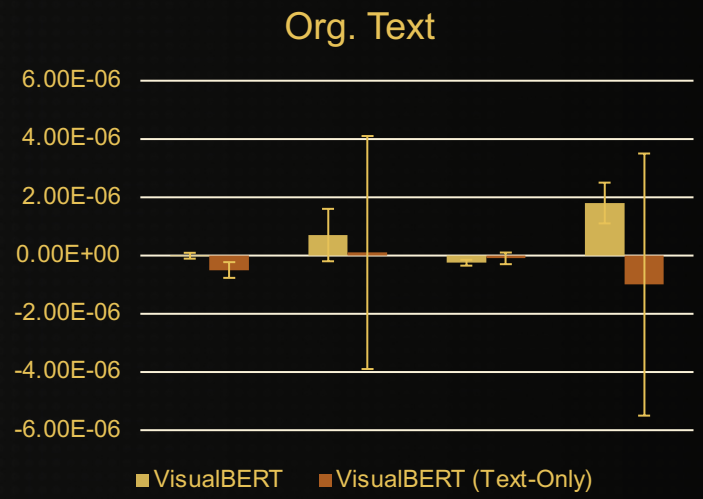



# Results (MIDAS)

- In contrast to Attr, captured the attention to the difference of modalities
- VisualBERT is biased toward textual information, enhanced by encoder

VisualBERT

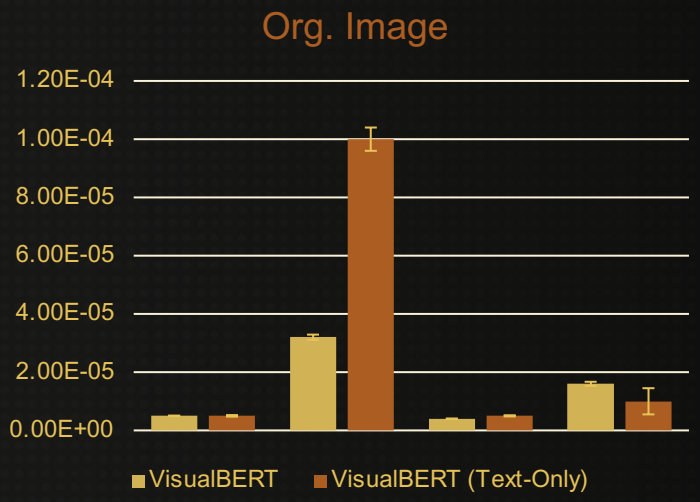


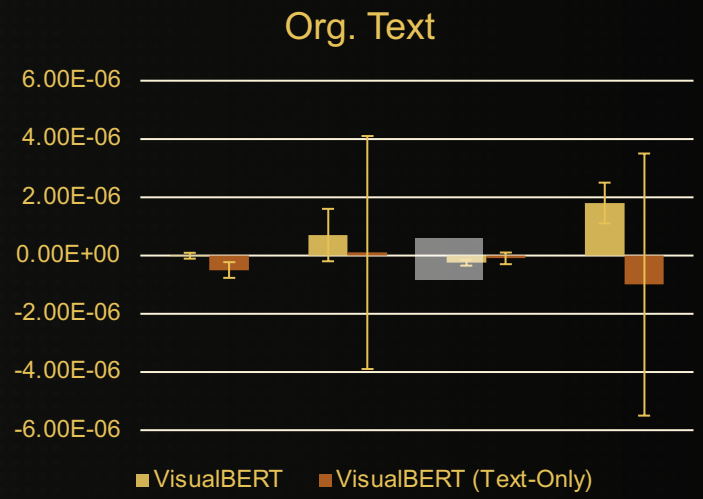



# Results (MIDAS)

- In contrast to Attr, captured the attention to the difference of modalities
- VisualBERT is biased toward textual information, enhanced by encoder

VisualBERT

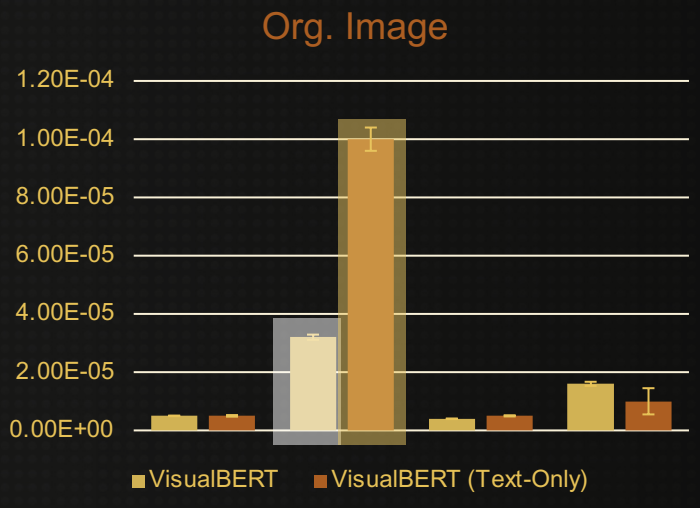


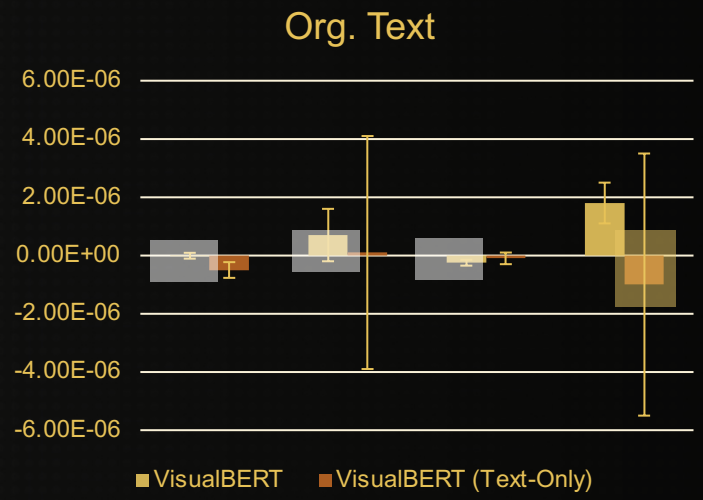



# Results (MIDAS)

- In contrast to Attr, captured the attention to the difference of modalities
- VisualBERT is biased toward textual information, enhanced by encoder

VisualBERT



# Results (MIDAS vs miATE)

Formal Relationship of miATE and MIDAS

$$miATE = \theta_{T,I} - (\theta s_T + \theta_I)$$

$$Attr(A) \approx A * G(A) \; where \; G(A) = \frac{\partial \theta(A)}{\partial A}$$

$$MIDAS \approx A_{T,I} * G(A_{T,I}) - (A_T * G(A_T) + A_I * G(A_I))$$

$$\sum_n MIDAS \approx \mathbb{E}[A_{T,I}] - (\mathbb{E}[A_T] + \mathbb{E}[A_I])$$

# Results (MIDAS vs miATE)

Empirical Relationship of miATE and MIDAS

| | | Oscar | VisualBERT |
|---|---|---|---|
| AUC(%) | | $74.3 \pm 2.39$ | $94.1 \pm 4.20$ |
| Feature Importance | text2text | $35 \pm 25$ | $253 \pm 66$ |
| | image2image | $30 \pm 23$ | $169 \pm 74$ |
| | cross-modal | $19 \pm 14$ | $169 \pm 74$ |

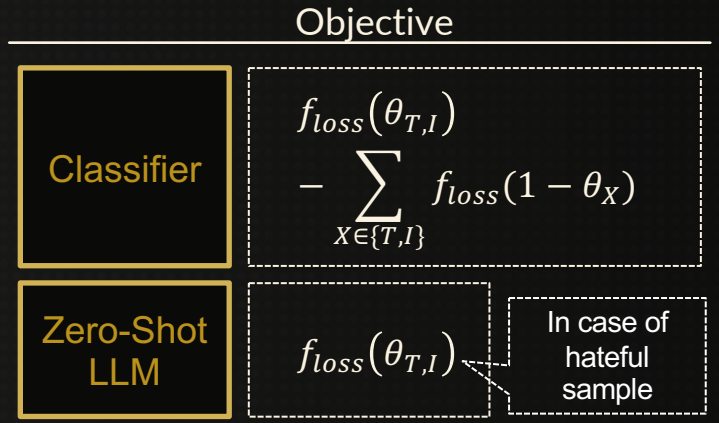
# 04

# Experiment II

Objective and Task Design / Meta Optimization /
Experimental Settings / Results

# Objective and Task Design

- Classifier's objective implicitly maximize miATE, while zero-shot LLM not



# Objective and Task Design

- Classifier's objective implicitly maximize miATE, while zero-shot LLM not
- Presenting positive / negative examples to align with the classifiers

Objective

Classifier

$$f_{loss}(\theta_{T,I}) - \sum_{X \in \{T,I\}} f_{loss}(1 - \theta_X)$$

Zero-Shot LLM

$$f_{loss}(\theta_{T,I})$$

In case of hateful sample

Task Design

*Which one is hateful?*

# Meta Optimization

- Multifaceted nature of the task for chat-bot style LLM

Task Type Classification (TTC) / Label Identification (LI)

| | | |
|---|---|---|
| TTC | - | *Sorry, I couldn't understand your instructions.* |
| | + | *This option could be pessimistic.*  |
| LI | - | *This option could be hateful.*  |
| | + | *This option could be hateful.*  |

# Meta Optimization

- Multifaceted nature of the task for chat-bot style LLM

Task Type Classification (TTC) / Label Identification (LI)

| | | |
|---|---|---|
| TTC | - | *Sorry, I couldn't understand your instructions.* |
| | + | *This option could be pessimistic.*  |
| LI | - | *This option could be hateful.*  |
| | + | *This option could be hateful.*  |

# Meta Optimization

- Multifaceted nature of the task for chat-bot style LLM
- Similar analysis for meta-gradient

Task Type Classification (TTC) / Label Identification (LI)

| | | |
|---|---|---|
| TTC | - | *Sorry, I couldn't understand your instructions.* |
| | + | *This option could be pessimistic.*  |
| LI | - | *This option could be hateful.*  |
| | + | *This option could be hateful.*  |

Meta-Gradient by Interaction Type

$$SubTask\ st = TTC$$
$$Interaction\ Type\ it \in \{t2t, i2i, cross\ modal\}$$
$$\theta = \sum_{st}\sum_{it}\left(A_{it}^{st} + \Delta A_{it}^{st}\right) * Q$$

# Experimental Settings

Model

<Hateful Memes Detection>

- Llama-2[*1] with BLIP-2[*2] caption extraction

<Probing for meta-gradient vs subtask output>

- LightGBM[*3] with Optuna[*4] hyperparameter search

*1: Touvron et al., 2023 (10.48550/arXiv.2307.09288)
*2: Li et al., 2020 (10.5555/3618408.3619222)
*3: Guolin et al., 2017 (10.5555/3294996.3295074)
*4: Akiba et al., 2019 (10.1145/3292500.3330701)

# Results

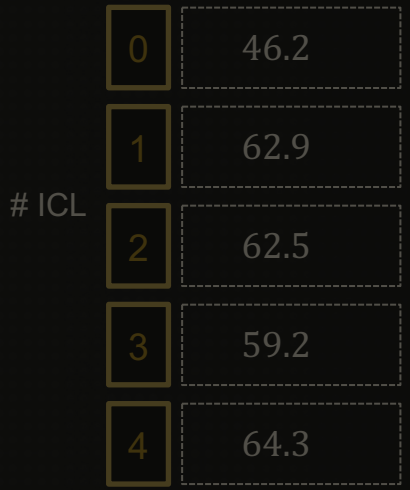
- ICL impacts comprehension of the challenging task

TTC Performance (Accuracy %)

# Results

- ICL impacts comprehension of the challenging task
- No big difference of contribution among interaction type

Performance (Accuracy %)




$$\theta = \sum_{st} \sum_{it} \left( A_{it}^{st} + \Delta A_{it}^{st} \right) * Q$$

Feature Importance

| | A | ΔA |
|---|---|---|
| text2text | $65 \pm 41$ | $24 \pm 17$ |
| image2image | $43 \pm 25$ | $14 \pm 8$ |
| cross-modal | $58 \pm 33$ | $31 \pm 21$ |

# 05

# Conclusion

# Conclusion

| | |
|---|---|
| Contribution | • Causal intersectionality for multimodal analysis<br>• Metrics for causal effect and model inner workings<br>• Relationship between causal effect and model inner workings<br>• Impact of meta-optimization on the model performance |
| Future Work | • Generalizability for other hateful memes datasets<br>• Generalizability for other problems (e.g. missing modality, medical diagnosis) |
| Limitations | • Resources for hateful memes dataset<br>• Language constraint |

# Thanks!

## Do you have any questions?

yosuke.miyanishi@jaist.ac.jp