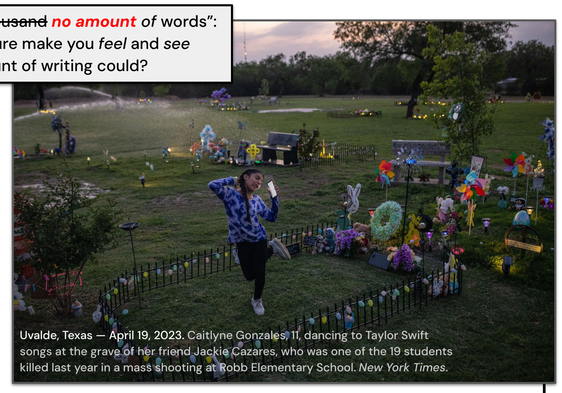
Representing the Discursive Construction 📣 of Visual Morality

Andre Ye
CSE 582: AI Ethics

1 ○ Problem

"A picture is worth ~~a thousand~~ ***no amount*** *of* words": What does this picture make you *feel* and *see* which no amount of writing could?



Uvalde, Texas — April 19, 2023. Caitlyne Gonzales, 11, dancing to Taylor Swift songs at the grave of her friend Jackie Cazares, who was one of the 19 students killed last year in a mass shooting at Robb Elementary School. *New York Times*.

[A] Morality is understudied in CV. We think of morality mainly in language, but psychological and philosophical work shows the importance of *vision* to moral perception, feeling, and thinking.

[B] Morality is treated as a "labeling problem". Major works in NLP treat morality as learning `f(context) -> judgment`. But philosophical and sociological work emphasizes the *discursive construction* of morality. Our moral beliefs *form* through *discourses* at personal and societal levels.

e.g. Overton window ○ "common sense" ○ censorship ○ taboo ○ praise

"Each society has its régime of truth, its 'general politics' of truth: that is, the types of discourses which it accepts and makes function as true."

– Michel Foucault, *Truth and Power*

4 ○ Ethics Discussion

Questioning "common sense" morality. A discursive approach to morality does not expect it to be "objective": it is dynamic, divergent, contextual.

Morality shouldn't be easy... it's hard. We should struggle over moral questions instead of reducing them to label assignment. But we also need to set limits – many things, at many times, should not be discourse objects.

From discourse to action? Action is not all of morality, but it's a big part. It remains unclear how moral discourse translates into moral action.

2 ○ Discursive Visual Representations

Can visual representations encode discursive information? What could we do with such representations?

Finetuned CLIP on image-discourse pairs: ~10 epochs

```
Text format:
"[comment] | [reply A] | [reply B]"
```

Discursive representations:

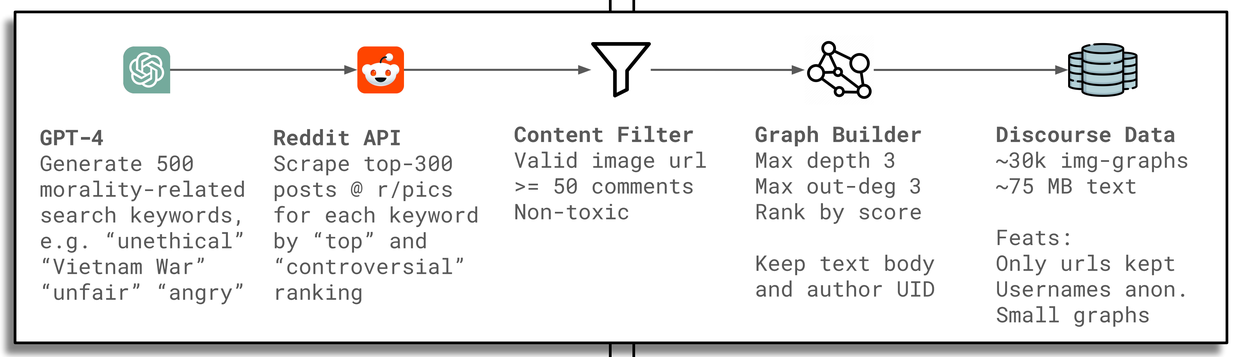
- improve MLP performance by 3% accuracy on Facebook Hateful Memes classification
- reduce MLP loss by 9% MSE on SMID moral response regression
- constant MLP loss on classic scene understanding tasks

Example (SMID)

No-FT CLIP: 4.25
DD-FT CLIP: 1.43
True Mean: 1.02

1 = immoral / blameworthy
5 = moral / praiseworthy

3 ○ Producing Discursive Graphs from Images



GPT-4
Generate 500 morality-related search keywords, e.g. "unethical" "Vietnam War" "unfair" "angry"

Reddit API
Scrape top-300 posts @ r/pics for each keyword by "top" and "controversial" ranking

Content Filter
Valid image url
>= 50 comments
Non-toxic

Graph Builder
Max depth 3
Max out-deg 3
Rank by score

Keep text body and author UID

Discourse Data
~30k img-graphs
~75 MB text

Feats:
Only urls kept
Usernames anon.
Small graphs

Can a model produce discourse from images? What kinds of HAI interactions might emerge?

Trained BLIP VQA (358M params); predict graph from image + title

```
Discourse graph format:
[{id: 0, reply_to: None, body: "abc"},
 {id: 1, reply_to: 0, body: "def"},
 {id: 2, reply_to: None, body: "ghi"}, …]
```

Image similarity search may align more with human moral intuition.

Discursive image search may be a more nuanced way to find images.

Prompt: `that's not fair! | he had it coming`

Sample model generations

I like how this is taken. Before the poor kid knows about it…
Yeah. the poor baby…
This poor baby…
This is one of the cutest photos I've seen

Those girls are so young…
It's a cute photo!
I wonder who the photographer is
I can't imagine what it would be like to be born like that little girl

Mommy needs to pick her kid up
Mommy isn't here
That's right!
At least they're behind a cage
What are they doing there?
Probably resting

*Minor edits for grammar and spelling – model still makes many mistakes and needs fine-tuning on a larger and higher-quality dataset.

Discursive graphs as image annotations (vs. captions).

User study with 5 participants: interaction with model. 15 min. Qualitative analysis.

Pros: [a] clearly represent moral issues and stakes, [b] feels more natural / less artificial [c] prompts critical responses (possible application to visual misinfo.)

Cons: [a] more mental labor to mentally process, [b] information can sometimes be redundant, [c] some comments are upsetting