Computational Methods for Multimodal Analysis: Case Studies on Hate Speech

Savvas Zannettou

SICSS-Saarbrücken

TU Delft

WARNING

CONTENT AND IMAGERY IN THIS TALK IS UNCENSORED AND MIGHT BE ***OFFENSIVE***

Hateful content is an everlasting problem with real-world impact

Anonymous ID:k6Gkl8Qt Thu 17 Nov 2022 21:49:42

>>404883670
Hahaha all joking aside , kill all Jews

Pittsburgh shooting: suspect railed against Jews and Muslims on site used by 'alt-right'

Robert Bowers appears to have used the platform Gab to accuse Jews of bringing 'evil Muslims' into US

▲ An ambulance arrives at the Tree of Life synagogue where a shooter opened fire. Photograph: Gene J Puskar/AP

TU Delft

Why is hateful content an everlasting problem?

- **Subjectivity & Disagreement:** No universal definition of “hate speech” and platforms, governments, and communities apply different standards.
- **Evolving Language & Symbols:** Users invent new slurs, coded language, memes, and emojis to evade detection.
- **Cultural & Contextual Nuances:** The same word/image may be offensive in one context but benign in another.
- **Scale & Speed:** Billions of daily posts across platforms.
- **Legal & Ethical Constraints:** Balancing free speech vs. harm prevention varies by jurisdiction.
- **Multimodality:** Detection is hard, especially across information formats like images and videos

There is a need for automated, generalizable, and multimodal solutions to detect hateful content!

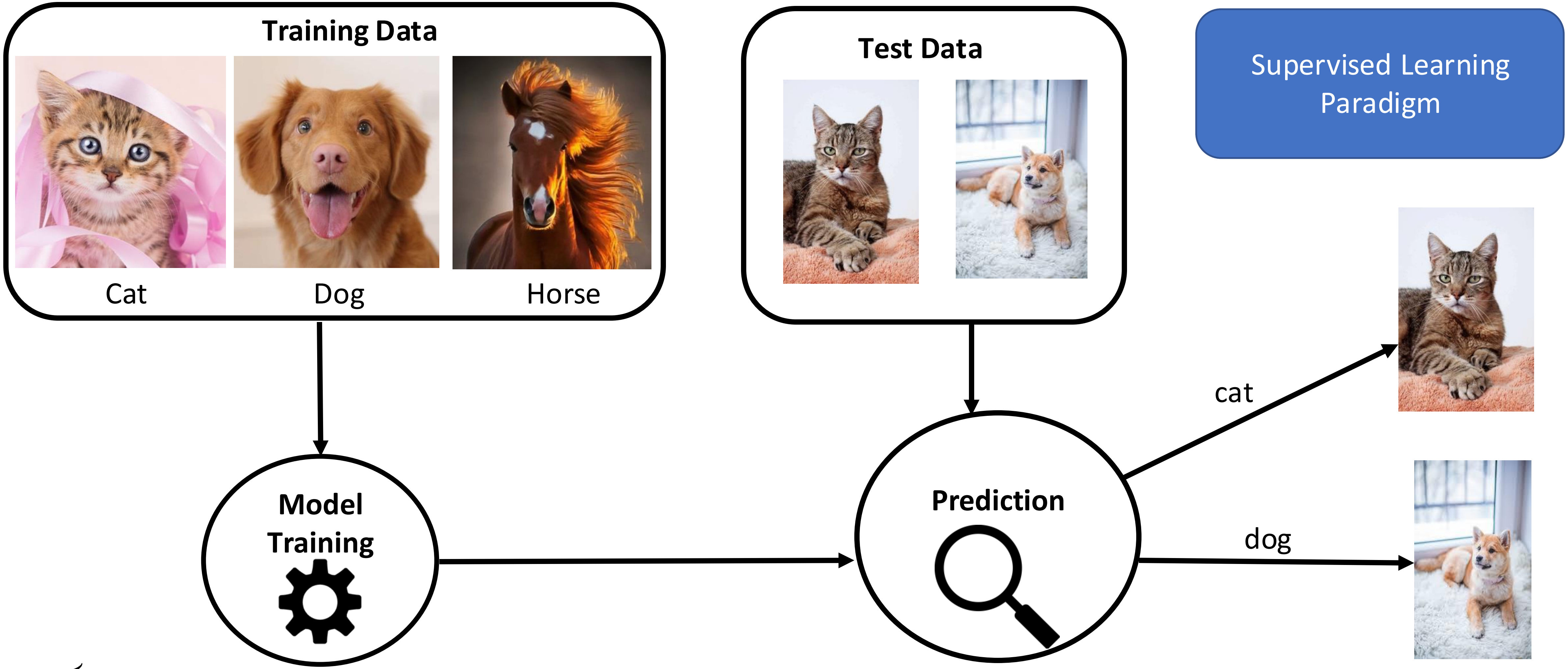
TU Delft

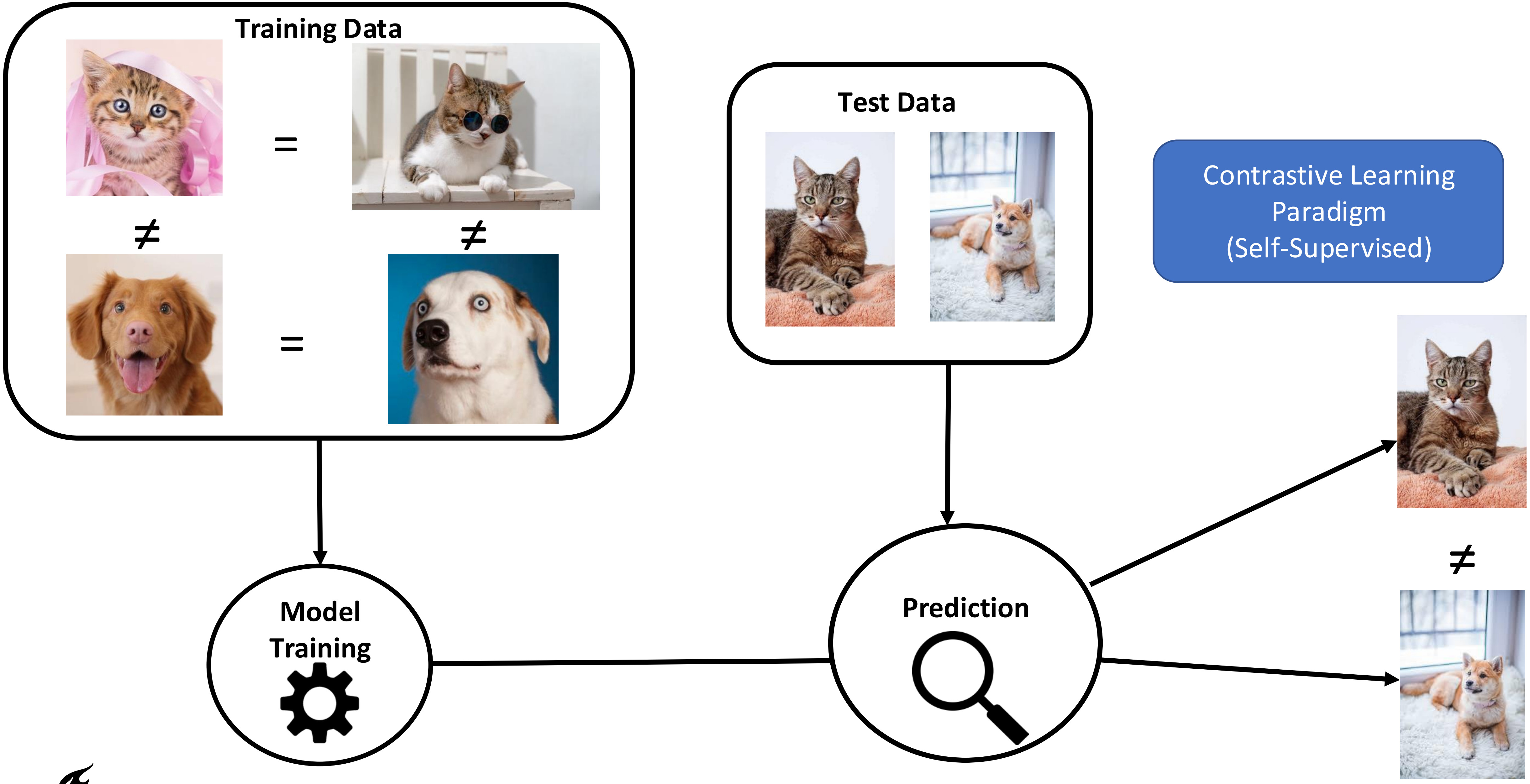
Agenda

- Introduction to Contrastive Learning and CLIP model

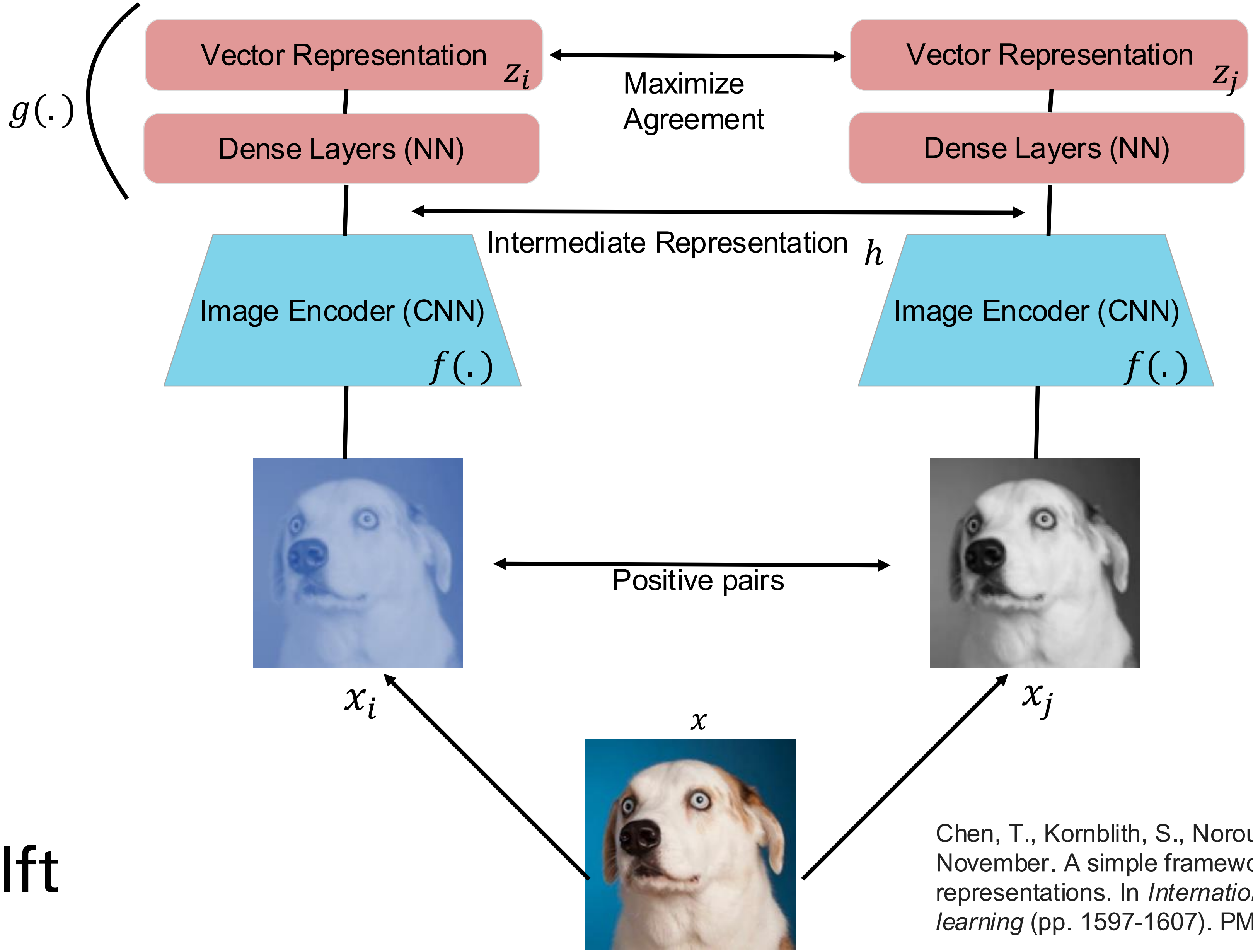
- Using CLIP model on the problem of hateful content online
  - Detecting Antisemitic and Islamophobic content (AAAI ICWSM 2023)
  - On the Evolution of Hateful Memes (IEEE S&P 2023)

TUDelft


Training Data
Cat
Dog
Horse
Test Data
Supervised Learning Paradigm
Model Training
Prediction
cat
dog


Training Data
=
≠
≠
=
Test Data
Contrastive Learning
Paradigm
(Self-Supervised)
≠
Model
Training
Prediction

Under the hood through the lens of SimCLR (Chen et al. 2020)



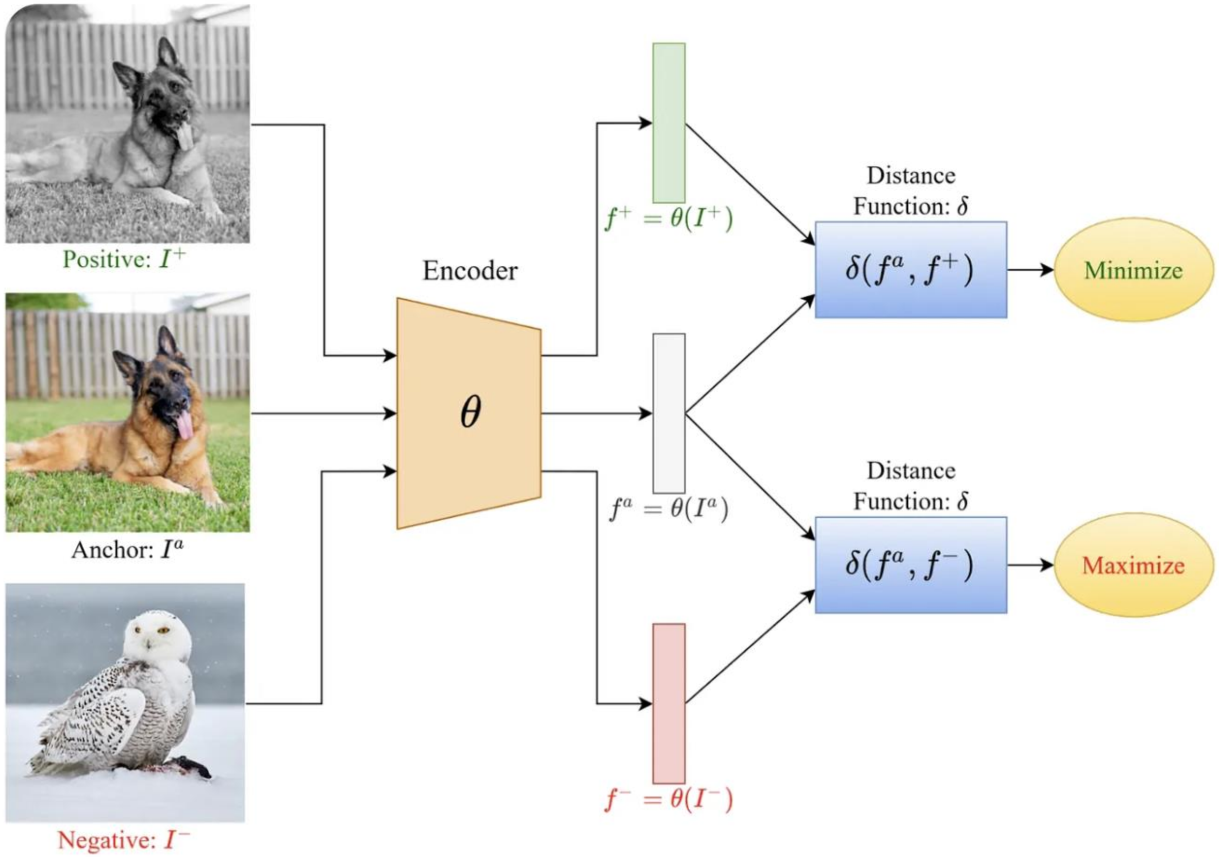
3. Projection and Contrastive Loss Minimization

2. Encoding

1. Data Augmentation

Chen, T., Kornblith, S., Norouzi, M. and Hinton, G., 2020, November. A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

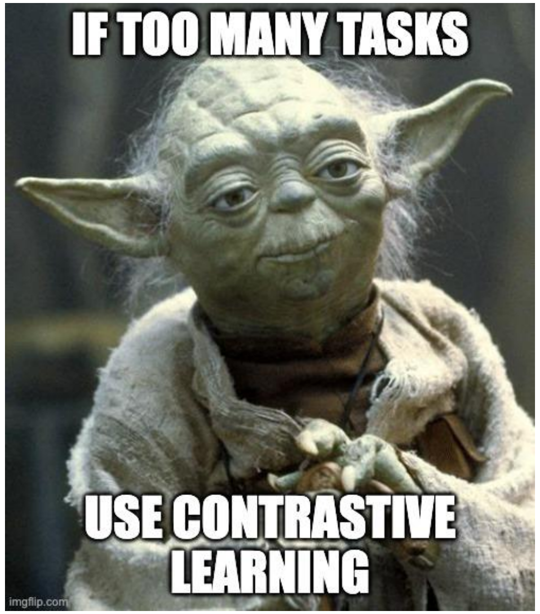
Contrastive Learning – Positive & Negative Samples



Source

Advantages of Contrastive Learning

- **Learn from unlabeled data**
  - Obtaining high-quality labeled datasets is a challenging and cumbersome task



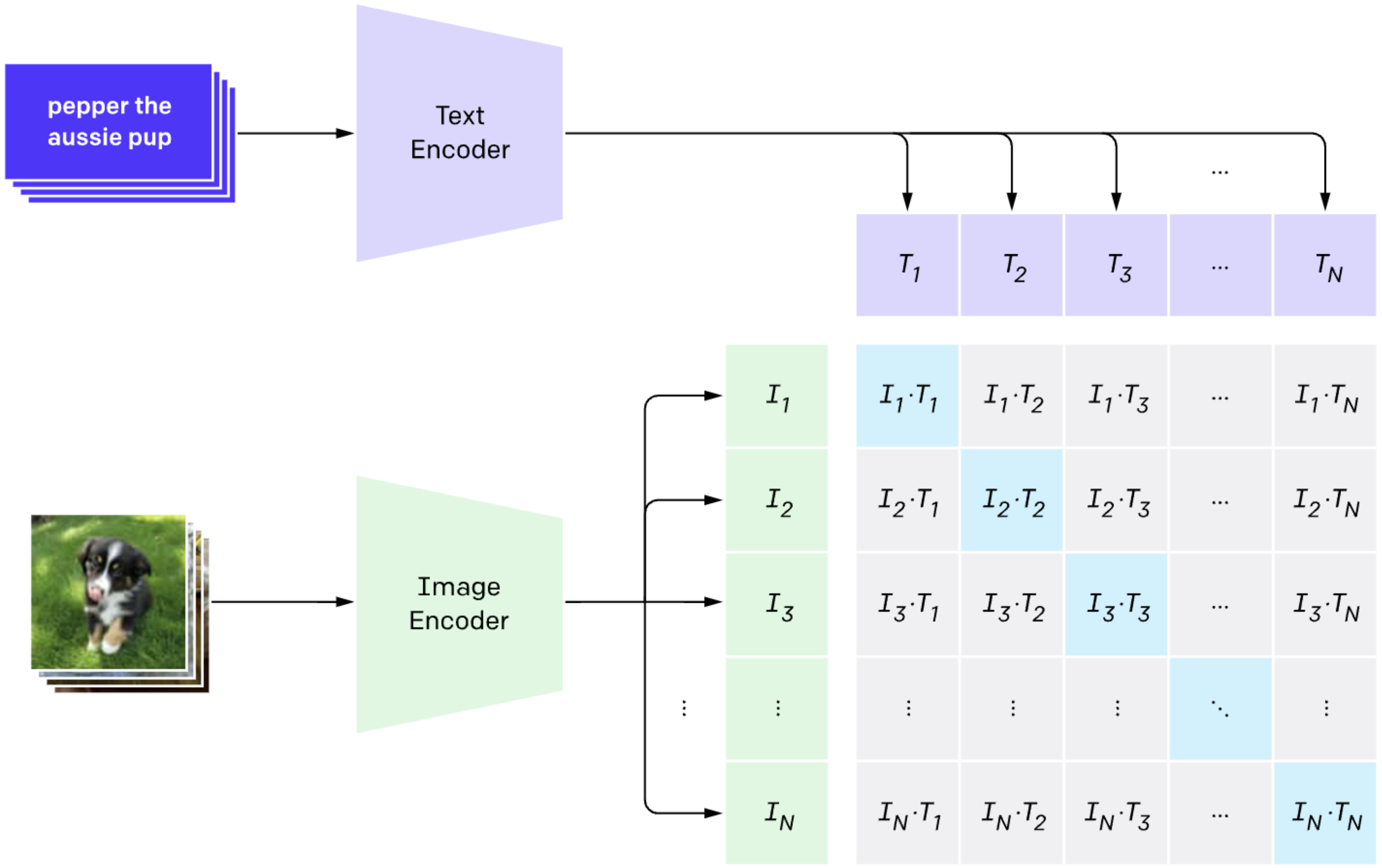
- **It can be applied to various downstream tasks**
  - The model is not trained on a specific task (e.g., classifying animals)
- **Mimics how humans learn by contrasting similar/dissimilar samples**
  - The trained model obtains more general knowledge compared to models trained on specific tasks

TU Delft

Connecting Text and Images using Contrastive Learning

TU Delft

OpenAI’s CLIP



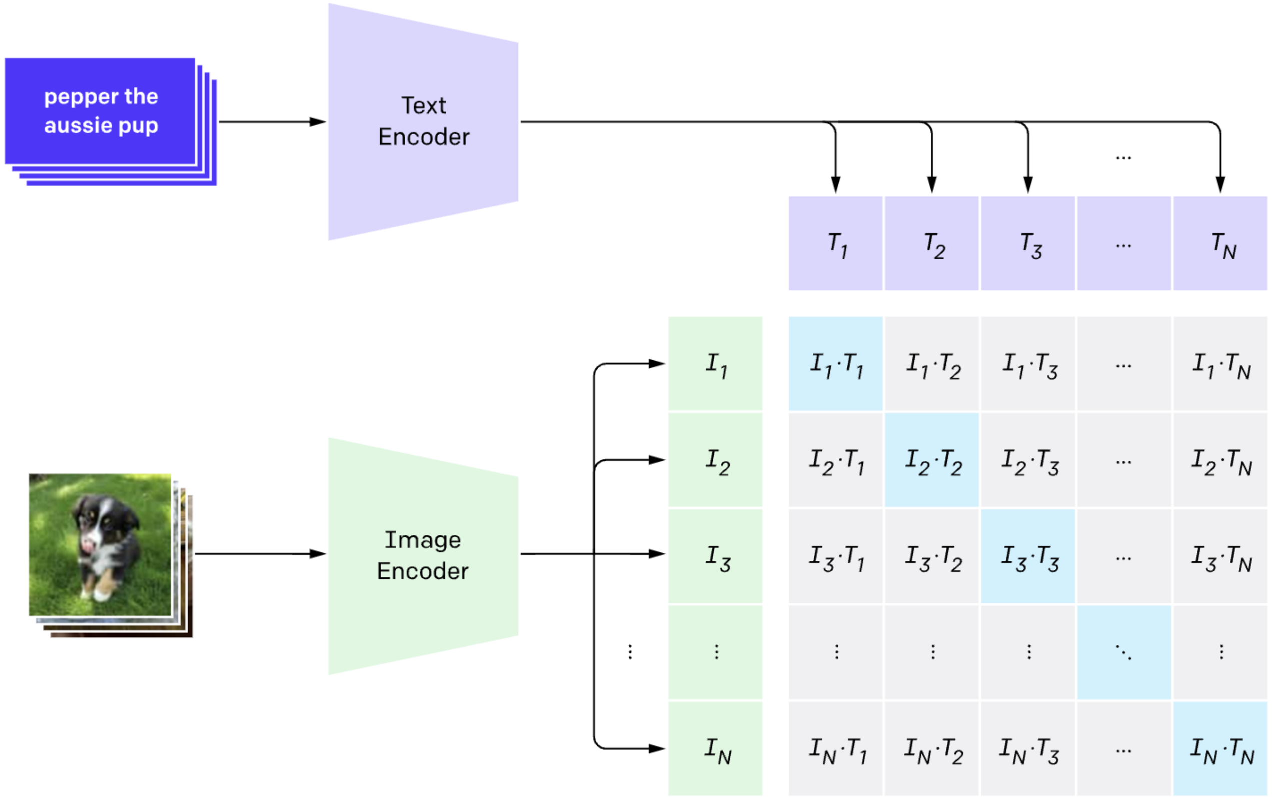
Contrastive pre-training

- 400M (image, text) pairs collected from the Internet
- Text Encoder and Image Encoder trained together
- Contrastive loss function

TU Delft

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.

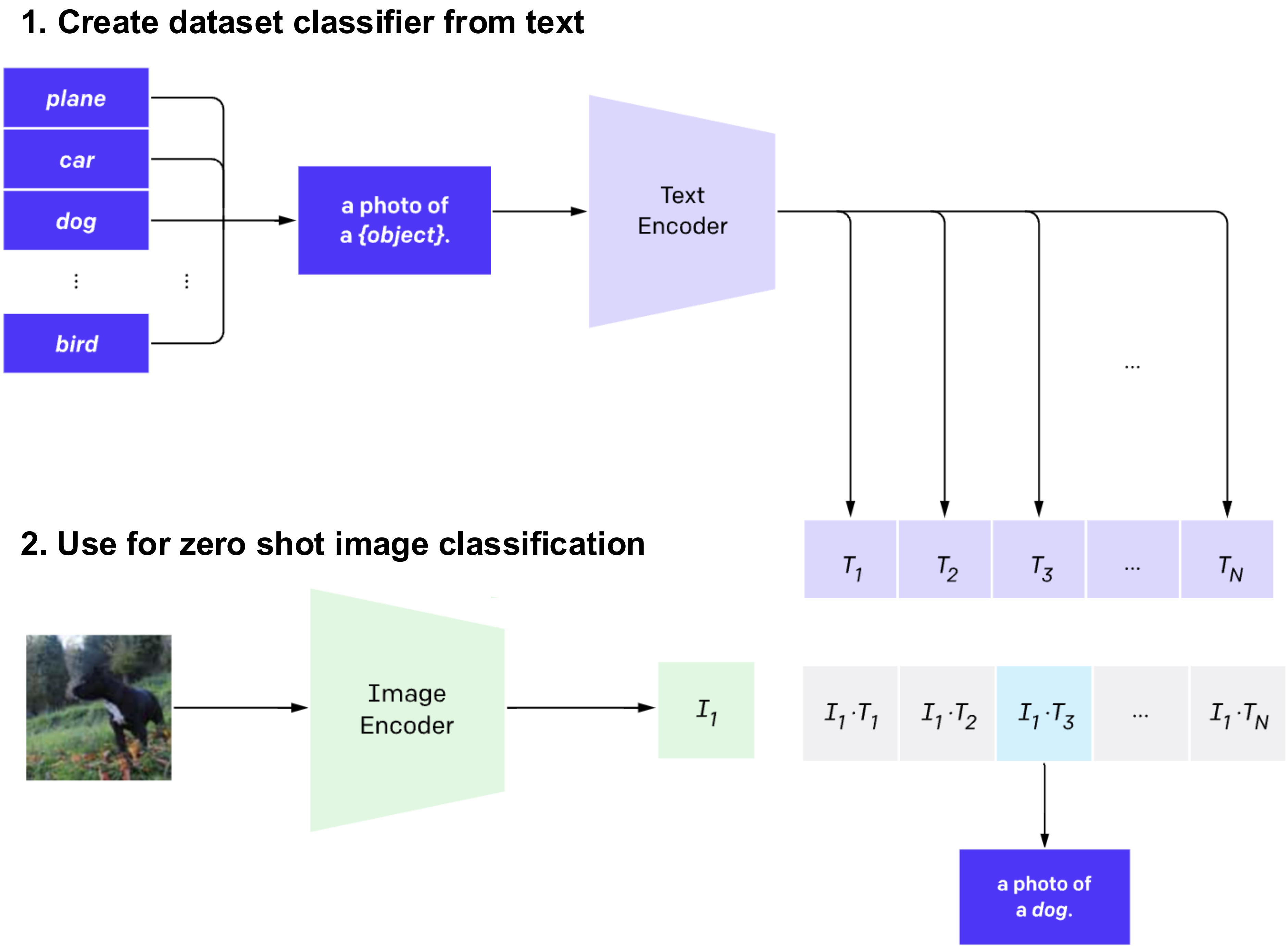
CLIP Contrastive Pre-Training




```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

TU Delft

CLIP for Zero-Shot Classification



CLIP for Cross-Modal Retrieval

- Use pretrained Image Encoder to convert all images to embeddings
- Use pretrained Text Encoder to convert all text to embeddings

- Then you can perform queries like:
  - Given a specific text, which is the closest image?
  - Given a specific image, which is the closest text?

- Fast indexing with Faiss library →

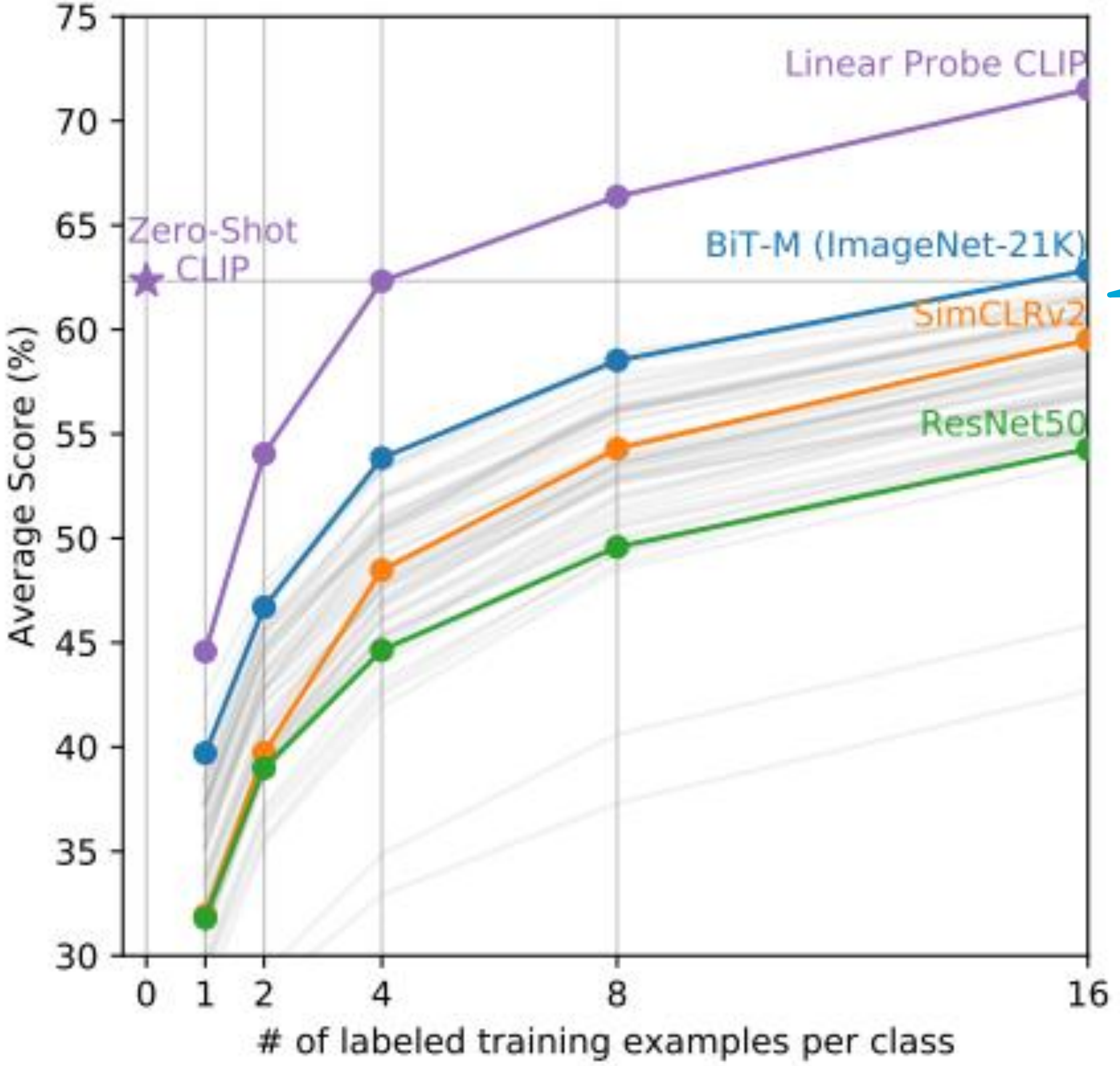
Billion-scale similarity search with GPUs

Jeff Johnson
Facebook AI Research
New York

Matthijs Douze
Facebook AI Research
Paris

Hervé Jégou
Facebook AI Research
Paris

Performance



Zero-Shot CLIP outperforms few-shot alternatives like ResNet

Variants of CLIP

- **OpenAI's CLIP (2021):** English-centric model, training data is not known

- **OpenCLIP from LAION:** Open re-implementation with a known dataset.
  - Various sizes: 400M, 2B, 5B
  - Available here

- **Multilingual CLIP:** Supports multiple languages of text using Knowledge Distillation on the original CLIP model
  - Available here

Using OpenAI’s CLIP for Studying Hate Speech

Property 1: Assessing Similarity Irrespectively of Modality

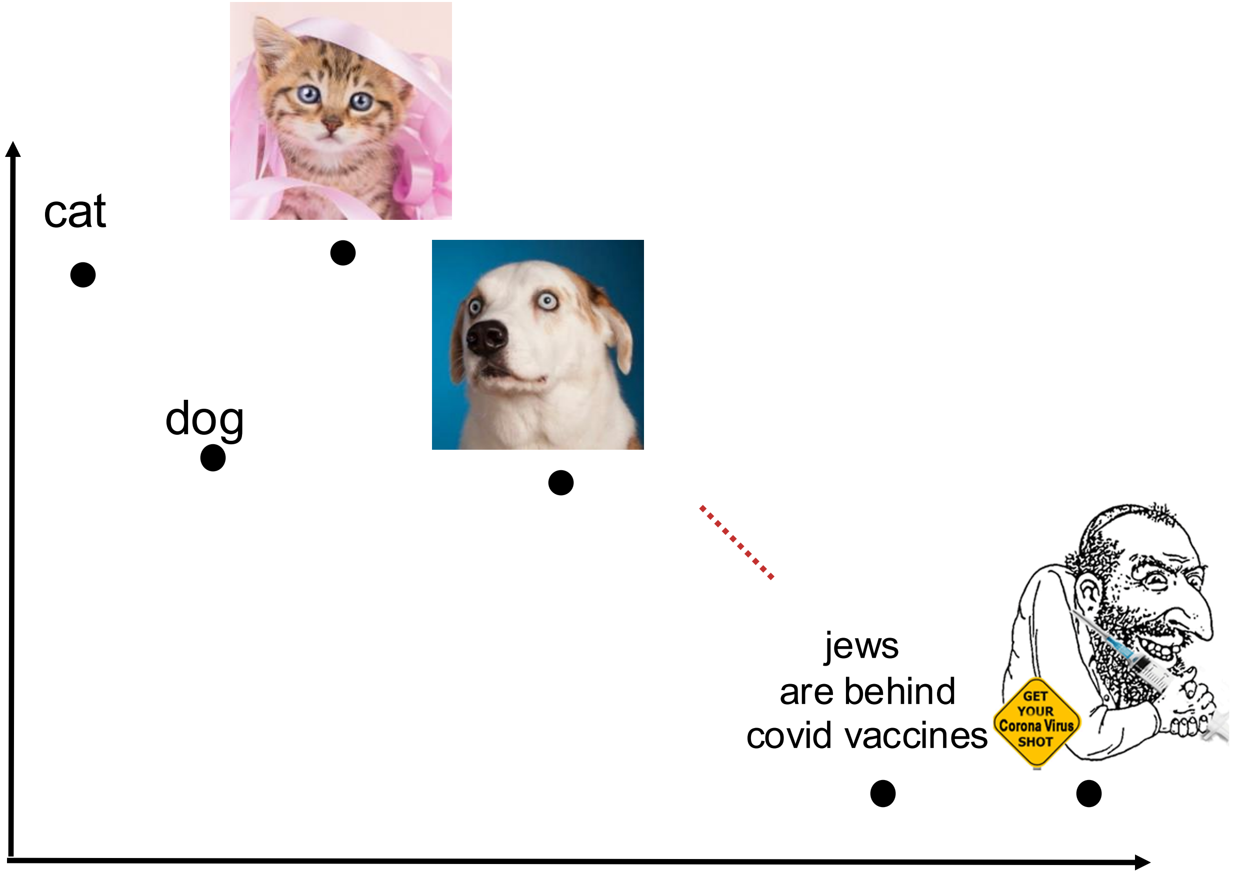
cat — 0.62 0.74 0.04 … 0.42

dog — 0.54 0.82 0.03 … 0.48

0.75 0.88 0.12 … 0.35

0.67 0.79 0.09 … 0.39

Vector representations



Projection of vector representations to a 2D space

TU Delft

Property 2: Relationships between representations

Visual Semantic Regularities

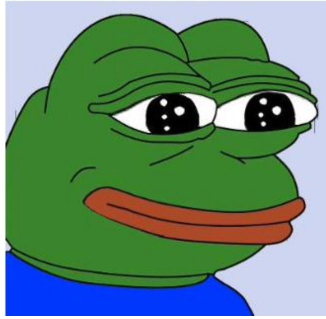
+

≈

−

≈

Visual-Linguistic Semantic Regularities



+ **nazi** ≈

− **CNN News** ≈

Understanding and Detecting Hateful Content using Contrastive Learning

AAAI ICWSM 2023
Joint work with Felipe González-Pizarro

TU Delft

OpenAI’s CLIP for hateful content detection

All images
posted on 4chan

Text Query:
“cute cat sleeping”

CLIP
model

Similar images
to the text query


0.334


0.322


0.321


0.316


0.315


0.304

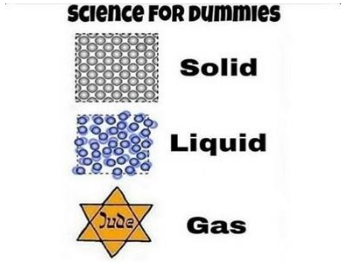
All images
posted on 4chan

Text Query:
“gas the jews”

CLIP
model

Similar images
to the text query

0.376



0.357

0.350

0.340

0.332

0.329

Dataset

- We focus on 4chan's /pol/:
  - Main Web community on 4chan that discusses world events and politics
  - Known for the dissemination of hateful content and conspiracy theories

- Collected all textual posts and images shared on 4chan's /pol/ between June 2016 and the end of 2017.
  - 66M posts (Papasavva et al. 2020)
  - 5.8M images (Zannettou et al. 2020)

Papasavva, A., Zannettou, S., De Cristofaro, E., Stringhini, G. and Blackburn, J., 2020, May. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *Proceedings of the international AAAI conference on web and social media* (Vol. 14, pp. 885-894).

Zannettou, S., Finkelstein, J., Bradlyn, B. and Blackburn, J., 2020, May. A quantitative approach to understanding online antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 786-797).

Research Questions

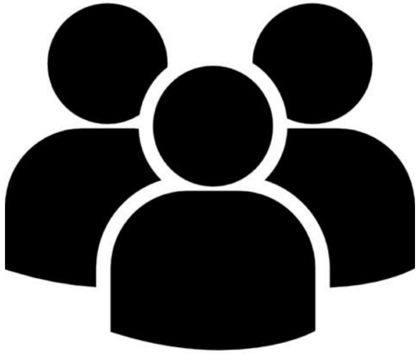
- **RQ1:** Can large pre-trained models that leverage the Contrastive Learning paradigm, like OpenAI's CLIP, identify hateful content with acceptable performance? How does CLIP's performance compare to state-of-the-art classifiers for detecting hateful imagery?

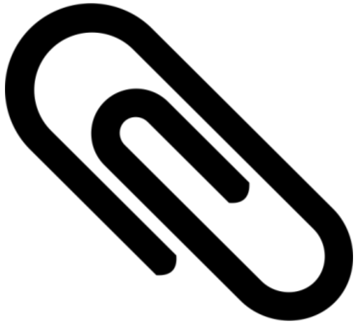
- **RQ2:** How prevalent are hateful imagery and textual hate speech on 4chan's /pol/?

- To limit the scope of the work, we focus on online Antisemitism and Islamophobia

TU Delft

Methods





1. Identify Antisemitic/Islamophobic phrases

Semi-automatic process that uses:
- Google's Perspective API
- Human annotations

2. Identify Antisemitic/Islamophobic imagery based on the phrases

Automatically using:
- OpenAI's CLIP model
- Cosine similarities between the phrases and unlabeled images

Identifying Antisemitic/Islamophobic phrases

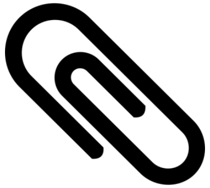
- Select toxic 4chan posts as detected by Google's Perspective API
  - 4.5M (out of 66M) posts with SEVERE_TOXICITY score >=0.8
- Select posts that have mentions to Jews or Muslims/Islam
  - 336K posts out of 4.5M
- Split posts into sentences and select common sentences (appearing at least 5 times)
  - 4.5K common phrases that are toxic and mention Jews/Muslims/Islam
- Two authors independently annotate the 4.5K phrases
  - Identified 326 Antisemitic and 94 Islamophobic phrases

Identifying Antisemitic/Islamophobic images

5.8M images
shared on 4chan

326 Antisemitic phrases
94 Islamophobic phrases

OpenAI



Select images where

$$cosine(P, I) \geq \theta$$

To reduce #false positives
keep images with at least
N matching phrases

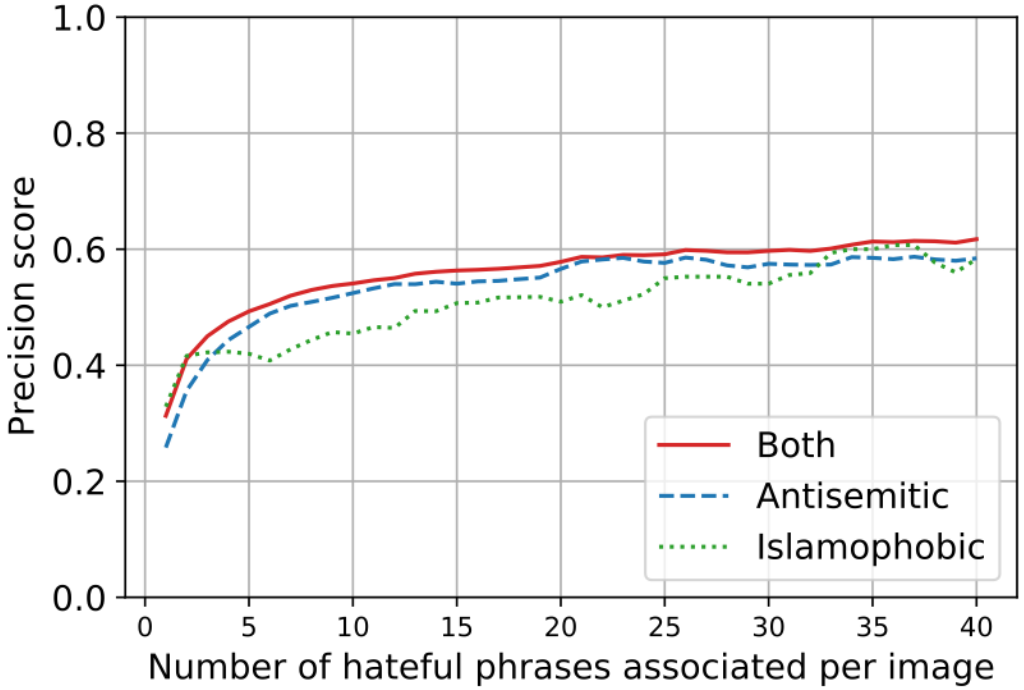
15K Antisemitic images
5.5K Islamophobic images

TU Delft

Selecting parameters for identifying hateful images

- Annotated another 2K images randomly selected from all images that have $cosine(P, I) \geq 0.3$
  - Identified that there are many false positives

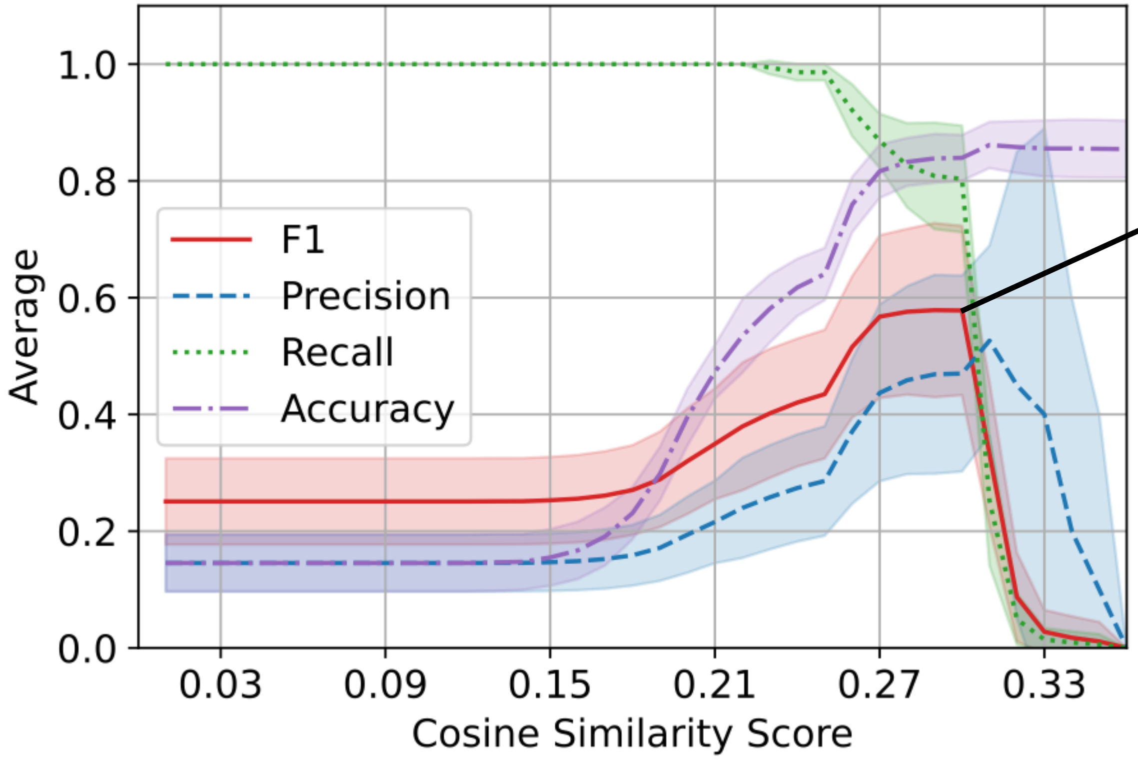
- Increase the performance by selecting images with at least 10 matching phrases



TU Delft

Selecting parameters for identifying hateful images

- To select θ we created a manually annotated ground truth dataset
  - 2000 images obtained from 10 randomly selected phrases (cover the entire cosine similarity range)



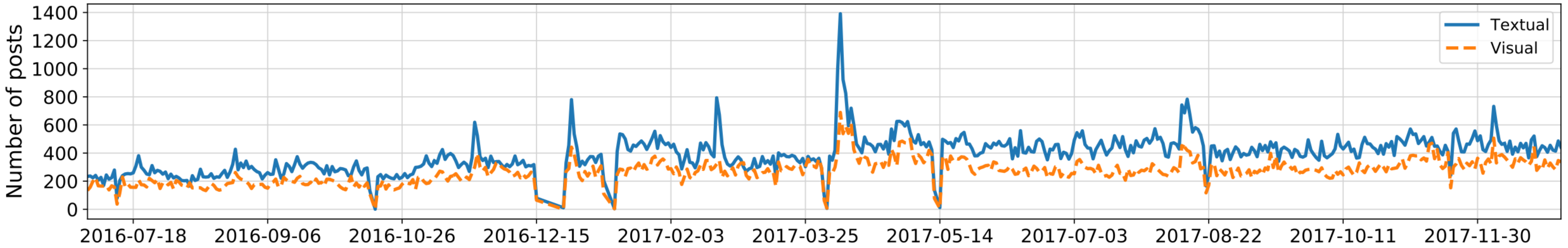
Best performance when
$\theta = 0.3$

Performance comparison with baselines

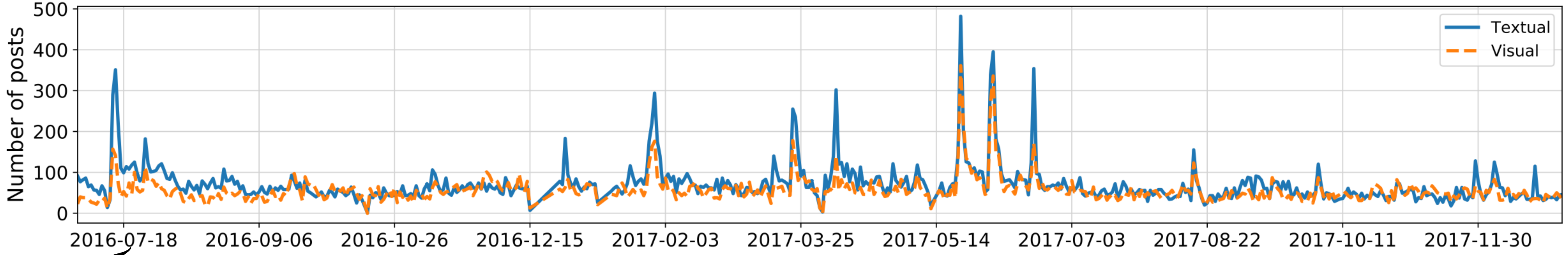
	Accuracy	Precision	Recall	F1
MMBT-Grid	0,70	0,37	0,61	0,46
MOMENT A-C	0,60	0,27	0,51	0,35
MOMENT A-P	0,57	0,29	**0,69**	0,40
CLIP Model	**0,81**	**0,54**	0,53	**0,54**

TUDelft

Posts with textual/visual hateful content over time



Antisemitism



Islamophobia

TUDelft

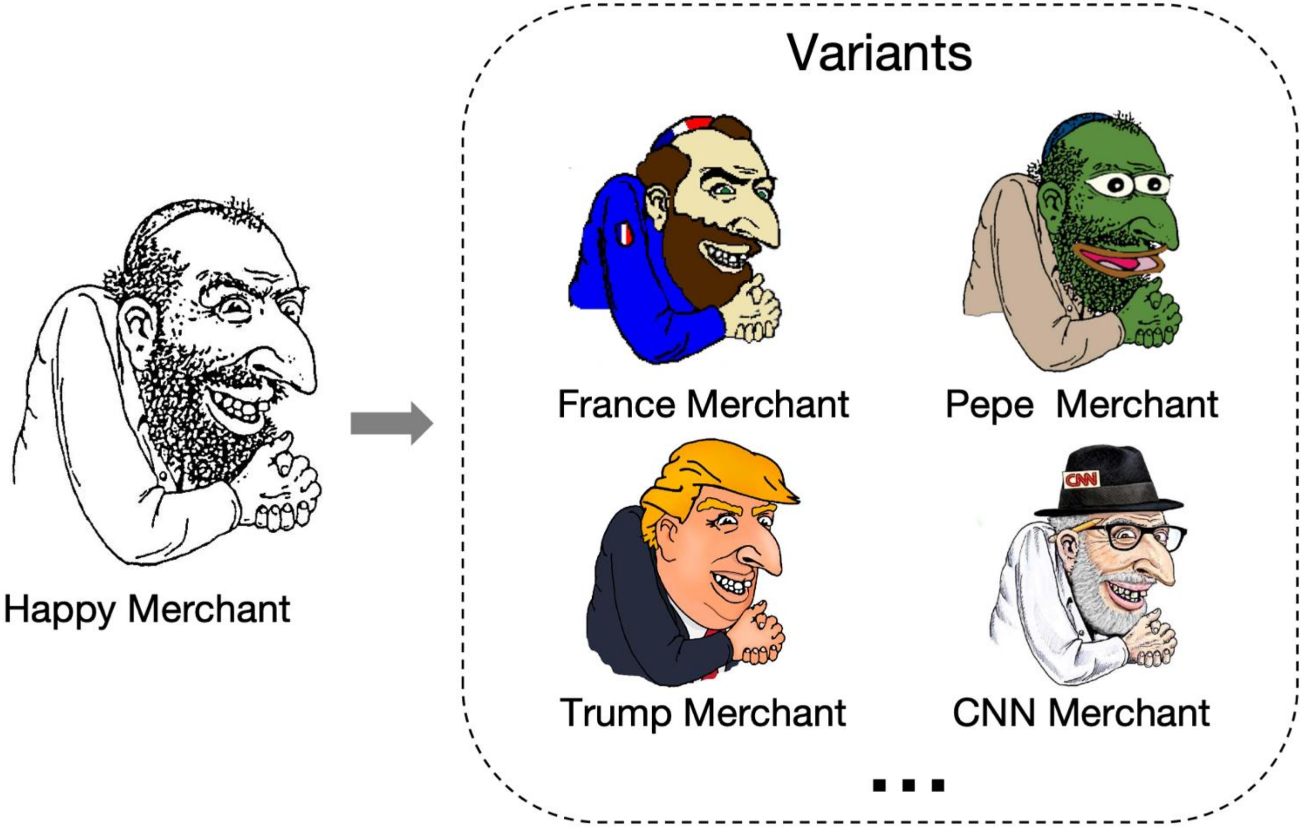
On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning

IEEE S&P 2023.
Joint work with Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, and Yang Zhang (CISPA)

TU Delft

Motivation

- (Hateful) memes have an evolutionary nature
  - New hateful meme variants can emerge by combining other cultural ideas/symbols



How can we systematically identify these variants and study meme evolution?

Identifying variants/influencers using visual semantic regularities

+

≈

Known hateful meme **Influencer** **Hateful meme variant**

We aim to identify influencers and variants given a known hateful meme

Identifying variants using visual semantic regularities

For each image *i*
in our 4chan dataset

Calculate $sim_i = cos(e_0, e_i)$

Select image *i* if $t_{lower} \leq sim_i \leq t_{upper}$

Known hateful meme

e_o

Variants

France Merchant

Pepe Merchant

Trump Merchant

CNN Merchant

...

Identifying influencers using visual semantic regularities

Hateful Meme variant

e_v

Known hateful meme

e_o

For each image *i*
in our 4chan dataset

Calculate $sim_i = cos(e_0 + e_i, e_v)$

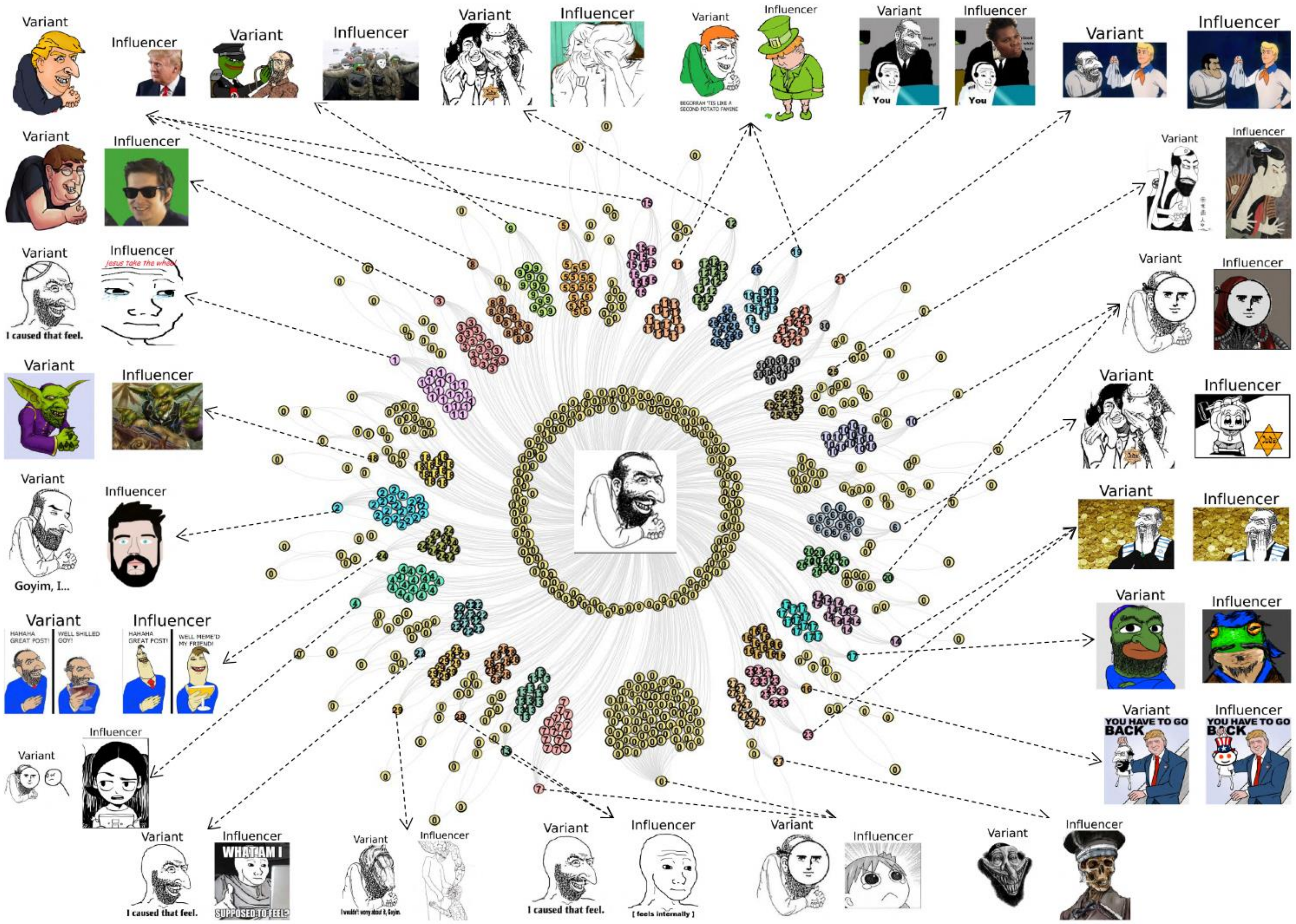
Select the image *with the highest* sim_i
if $sim_i \geq t_{lower}$

Influencer

Variants and Influencers of Happy Merchant Meme

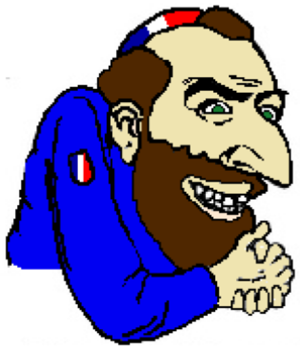
- We run our methodology starting from the Happy Merchant Meme
- Identified 3.3K variants along with their influencers
- Three authors performed annotations to assess the performance
  - A random sample of 100 variants/influencer pairs
  - 78% of the identified variants are correct
  - 53% of the influencers are accurate

Visualization of Variants/Influencers of Happy Merchants



Identifying variants using visual-linguistic semantic regularities

+ France ≈



Known hateful meme

Textual Influencer

Hateful meme variant

We aim to identify variants given a known hateful meme and a set of textual influencers

Identifying variants using visual-linguistic semantic regularities

1. Geo-Political Entities (GPE)
2. People
3. Organizations (ORG)
4. Nationalities, Religious, or Political Entities (NORP)

Textual Influencers

t_k

Known hateful meme

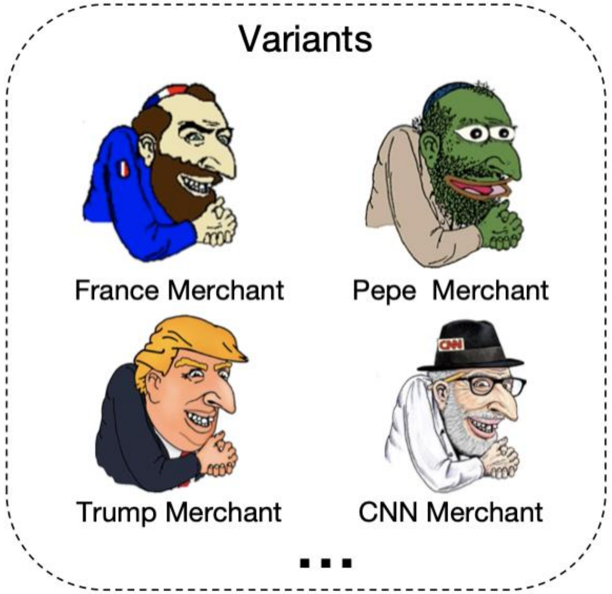
e_o

For each image *i* in our 4chan dataset

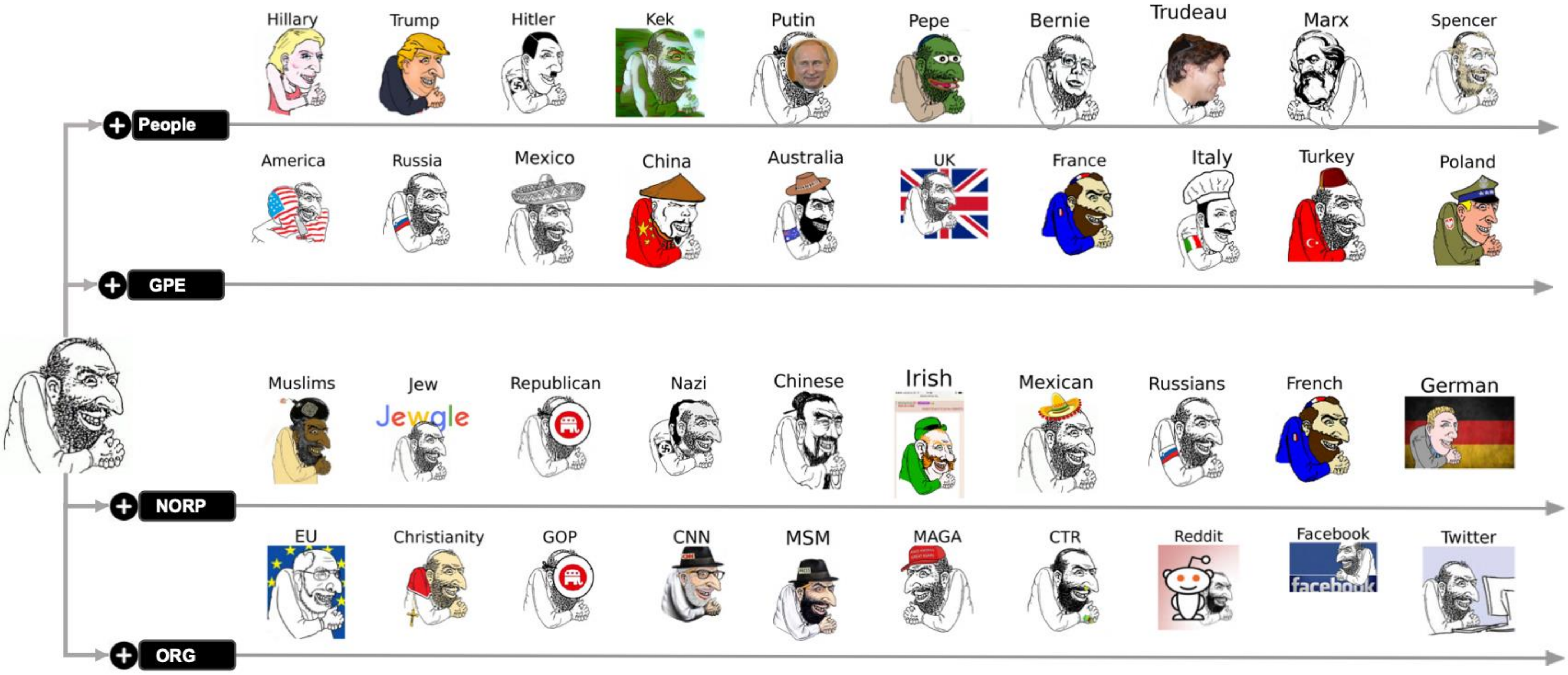
$$e_f = 0.2 * e_o + 0.8 * t_k$$

Calculate $sim_i = cos(e_f, e_i)$

Select the image *with the highest* sim_i if $sim_i \geq t_{lower}$



Variants identified using visual-linguistic semantic regularities



Annotation of variants using visual-linguistic semantic regularities

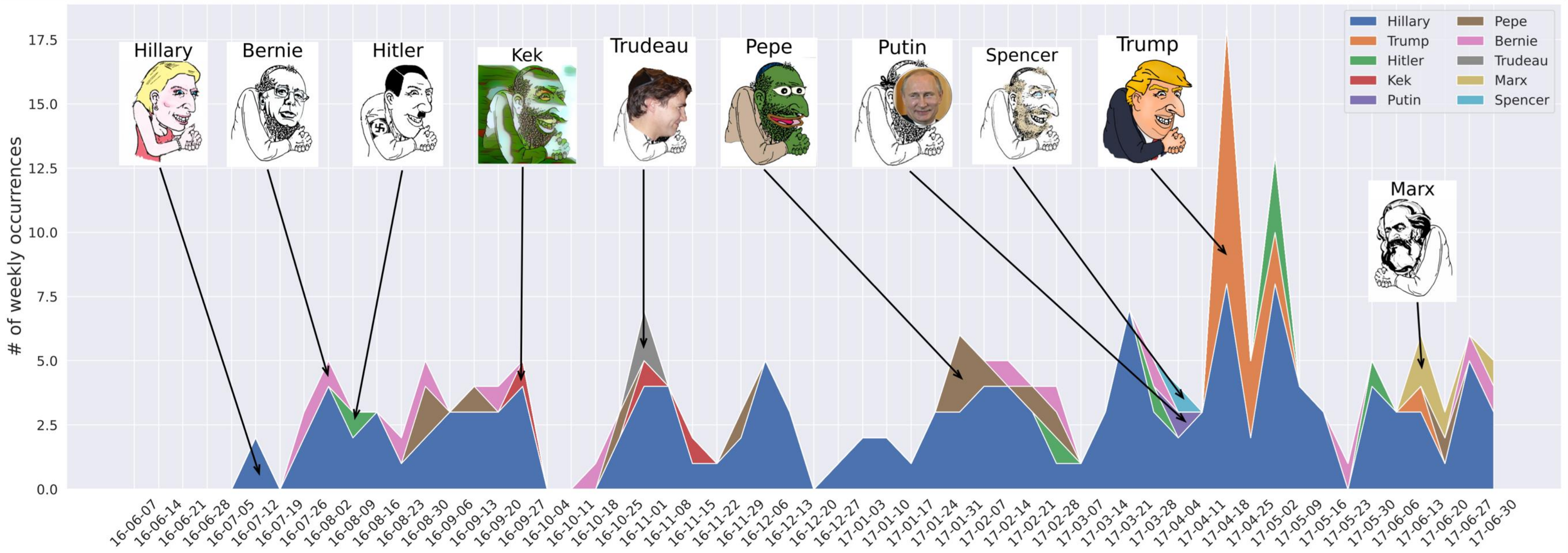
- Extracted the variants obtained using 120 textual influencers
  - Top 30 in GPE, People, NORP, and ORG based on their mentions in the text

- Three authors annotated the images to find out for how many entities we have variants
  - 48% for People
  - 76.7% for GPE
  - 80% for NORP
  - 63.5% for ORG

- **These percentages depend on whether 4chan users actually shared an image that is based on the textual influencer**
  - There is no generative part in our method

TU Delft

Temporal Analysis



TU Delft

Summary

- CLIP model can play a role in identifying hateful content
  - Our simple classifier outperforms previous classifiers focusing on identifying hateful images

- Systematic analysis of the evolution of memes using CLIP's semantic regularities
  - Visual semantic regularities
  - Visual-linguistic semantic regularities

- We envision our work to be used for aiding human moderators to identify hateful content

- Can be used to identify orchestrated hate campaigns

TU Delft

Conclusion

- CLIP is a *general-purpose multimodal model and* its ability to align text and images makes it useful across several tasks

- **Considerations:**
  - Training data bias
  - Context and prompt sensitivity
  - Compute needs

- CLIP is powerful, but its biases and prompt sensitivity mean it must be used with **critical awareness** and often in combination with other methods
  - Always validate the performance of CLIP under your dataset/task conditions

Thank you for your attention

Savvas Zannettou