MMVAE:

A Multi-modal Multi-task VAE on Misogynous Meme Detection

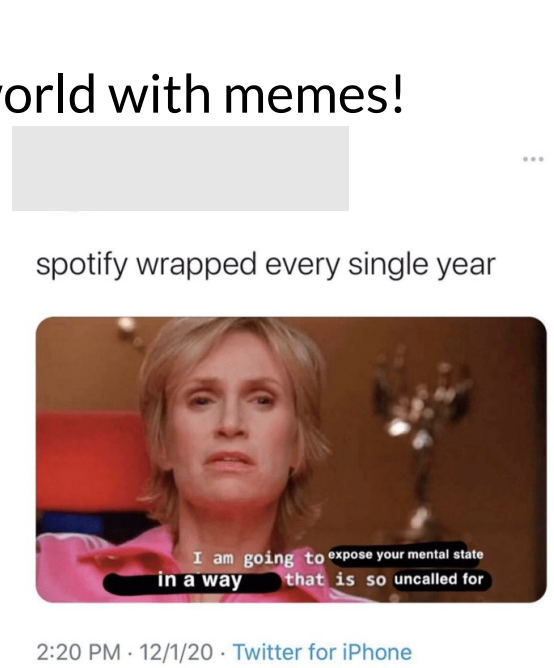
Yimeng Gu, Ignacio Castro, Gareth Tyson

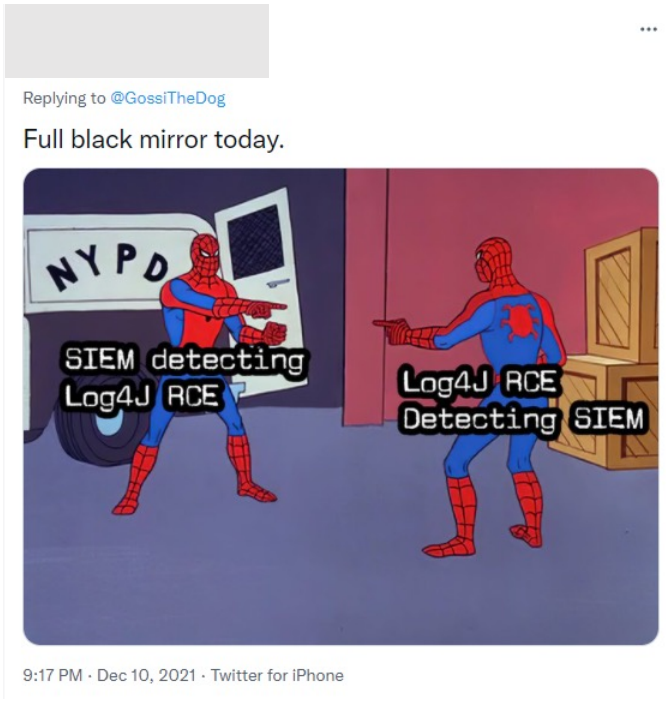
Outline of this talk

- Hateful meme detection and challenges
- Background
  - Pre-trained model
  - Multi-modal learning
  - Variational AutoEncoder (VAE)
  - Multi-task learning
- Our approach
- Evaluation

Hateful meme detection and challenges

A world with memes!





And hateful memes…

from Facebook Hateful Meme Challenge

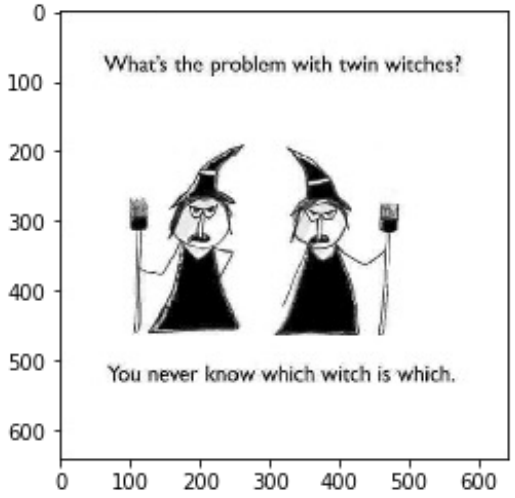
Hateful meme detection and challenges

Misogynous Meme Detection

- *Misogyny prediction*
- *Shaming*, *stereotype*, *objectification* and *violence prediction*

Misogynous

Stereotype



Non-misogynous

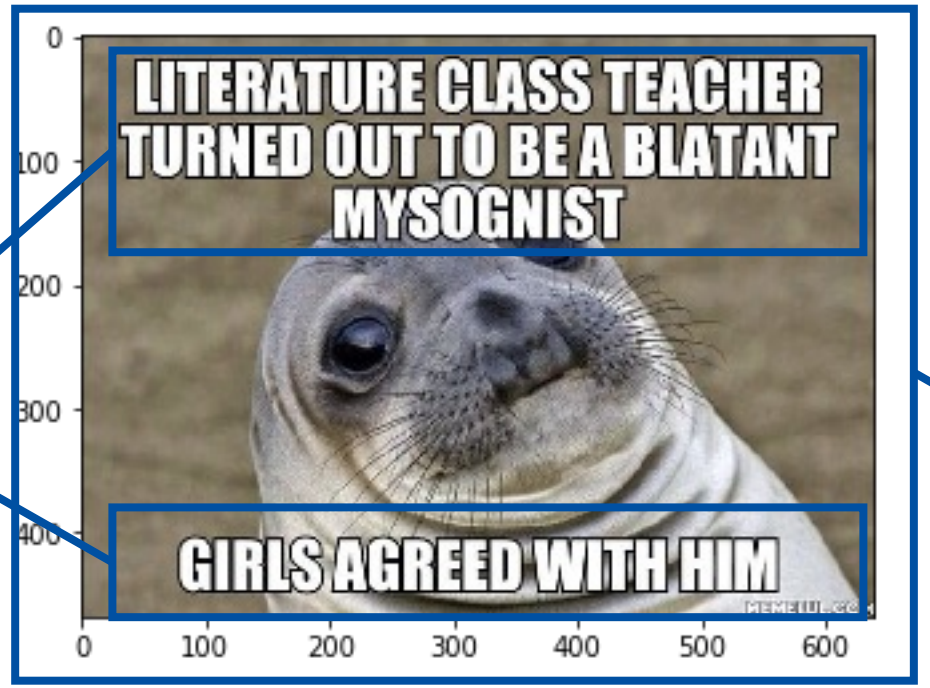
Misogynous

Shaming,
Stereotype,
Objectification

Hateful meme detection and challenges

Challenge #1

- With image and text



Text

Image

\+ = Multimodal!

Hateful meme detection and challenges

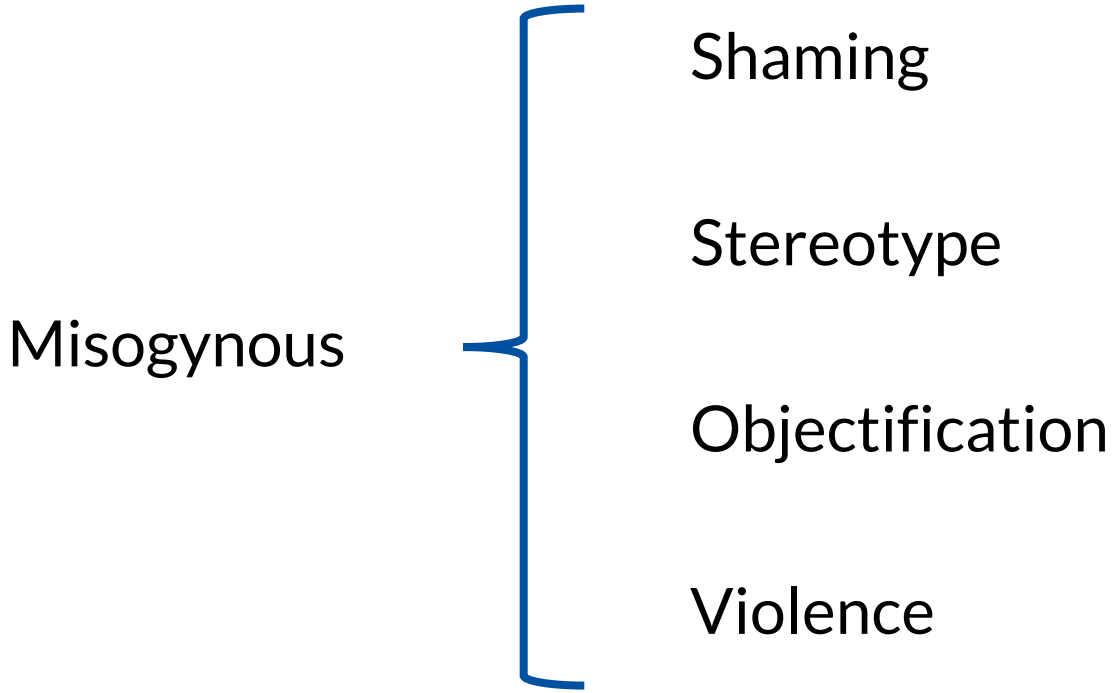
Challenge #2

- Same image, different texts

Hateful meme detection and challenges

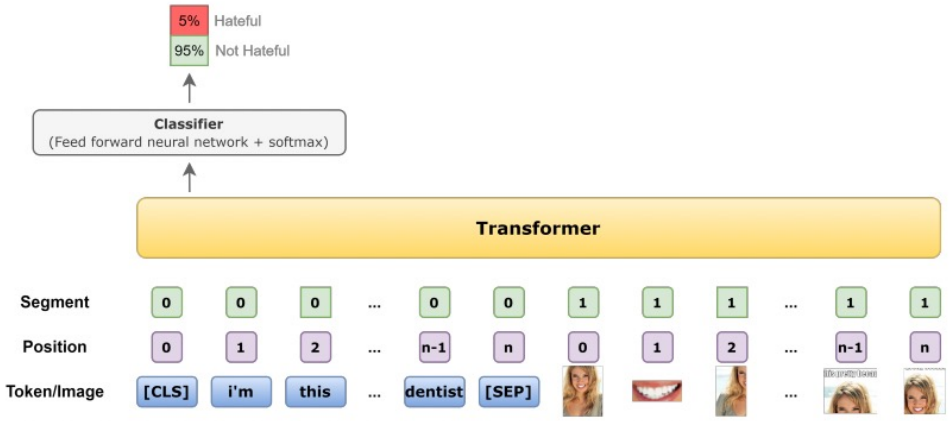
Challenge #3

- Granular labels on hateful message

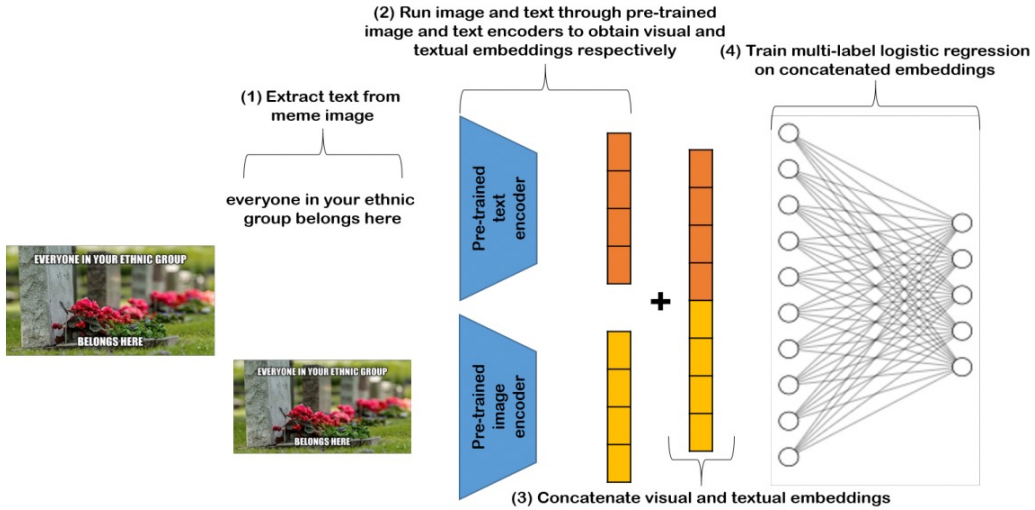


Hateful meme detection and challenges

Previous solutions



Velioglu et al., 2020



Zia et al., 2021

How to encode text and image?

Pre-trained model

→ Image pre-trained model → Image embedding

Pre-trained model

→

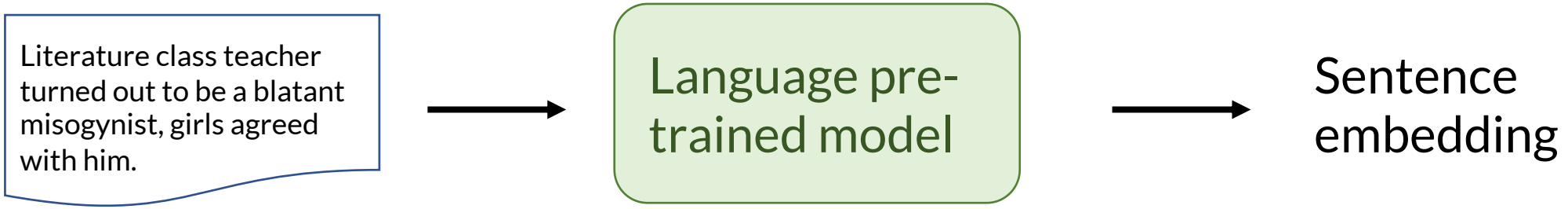
ResNet He et al., 2016

CLIP Radford et al., 2021

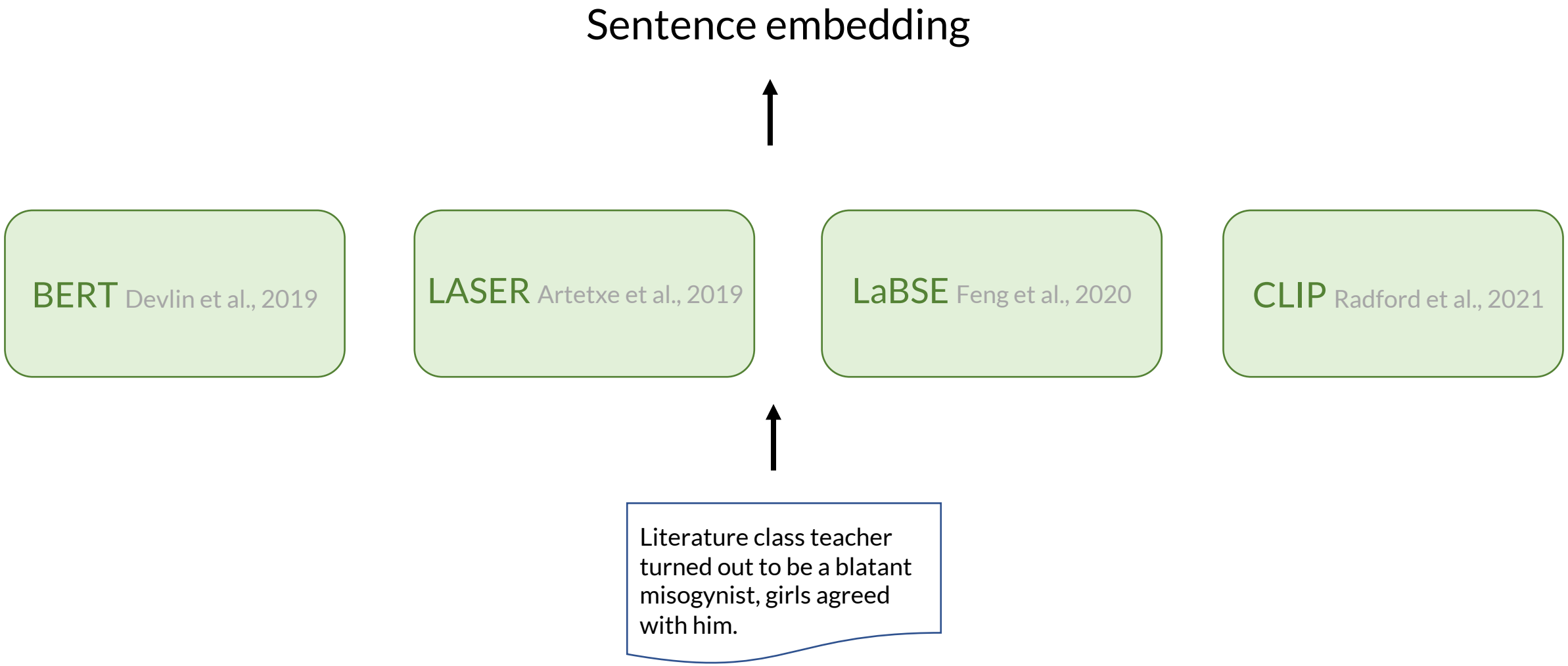
→

Image embedding

Pre-trained model



Pre-trained model

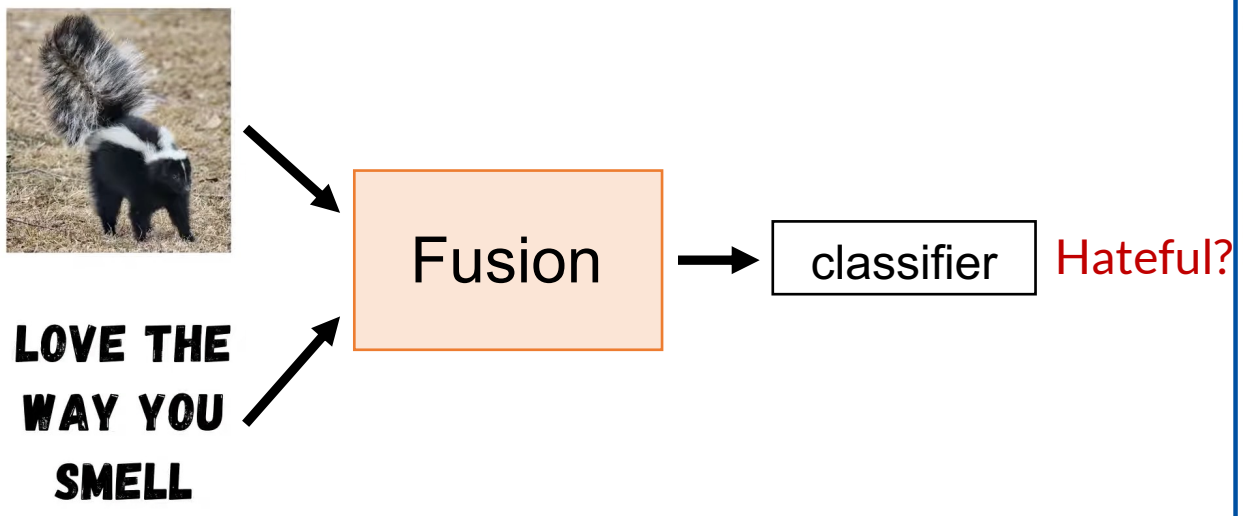


How to represent multimodal data?

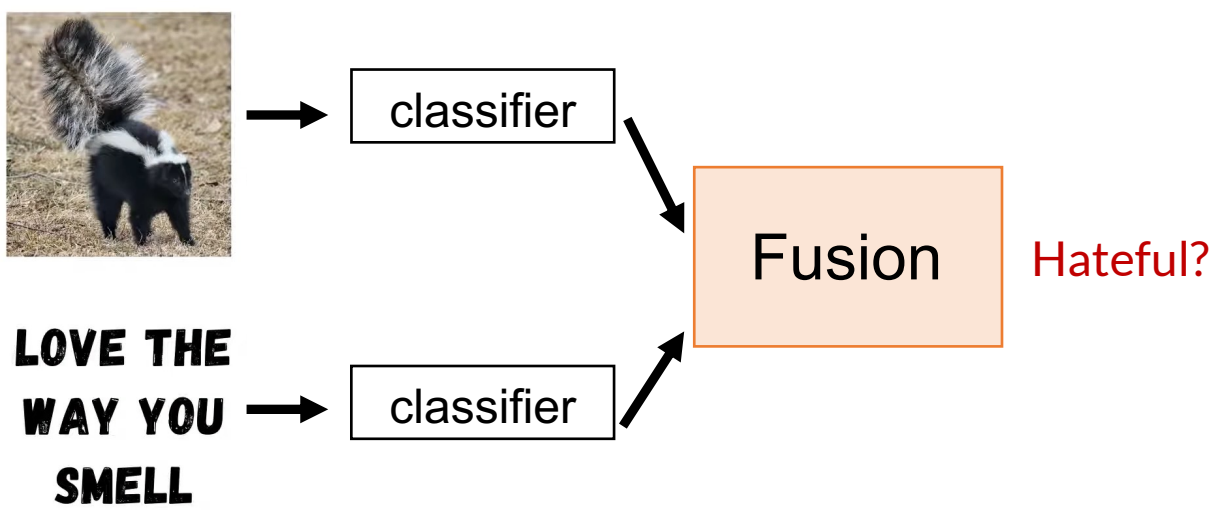
Multimodal learning

Fusion mechanism

- Early at feature level: concatenation, cross-modal attention, outer product…
- Late at decision level: voting



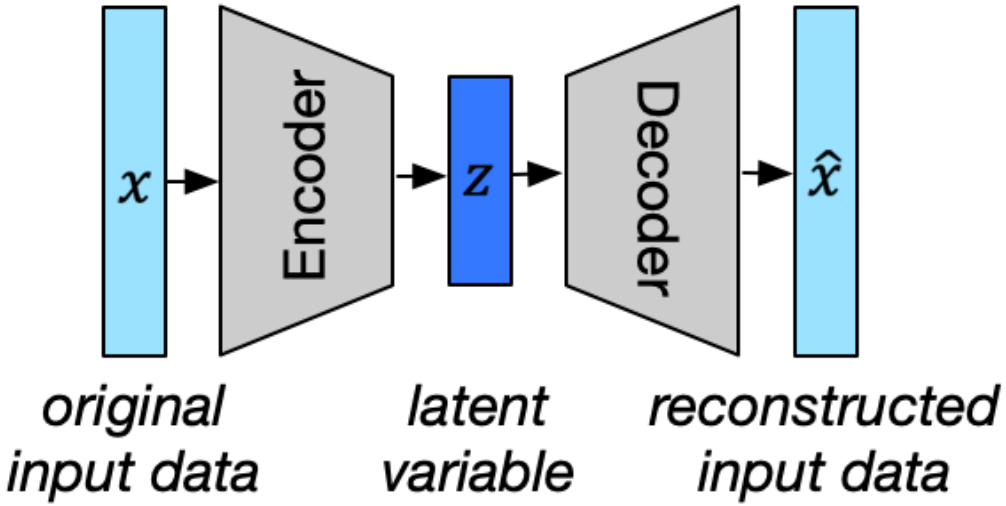
Early at feature level



Late at decision level

The approach we use to represent multimodal data

Variational AutoEncoder (VAE)



Aim:
Finding latent variable z that captures meaningful ***factors*** of variation in the data

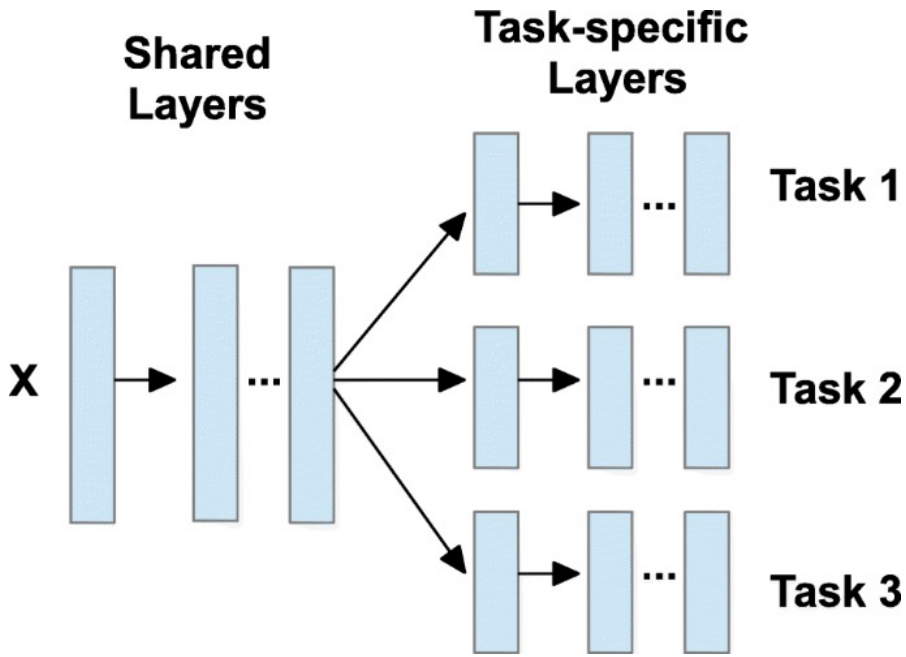
Kingma and Welling, 2014

How to benefit from the learning of the other task?

Multitask learning

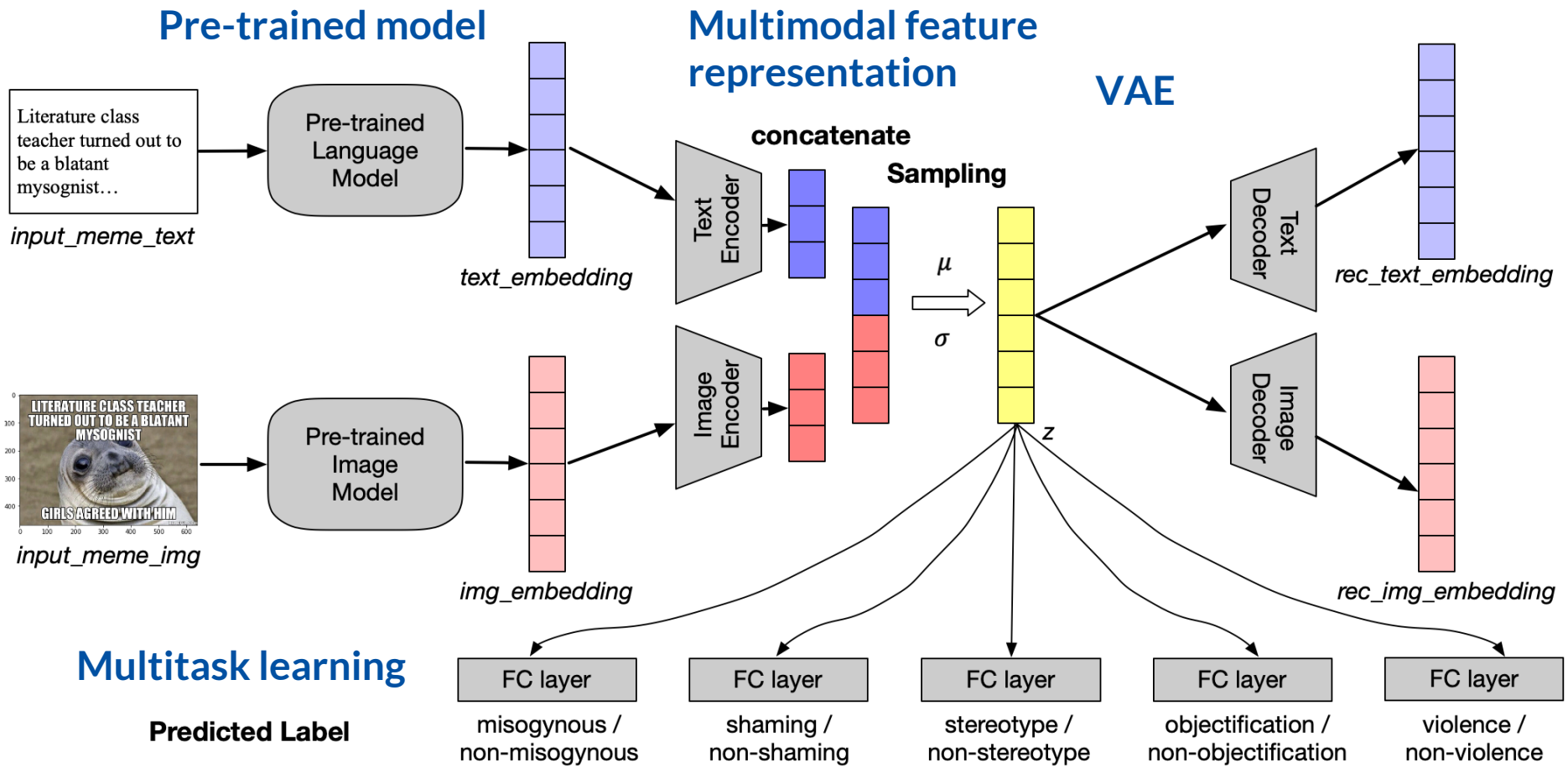
Benefit:

- Knowledge transfer



Multitask learning

Our approach

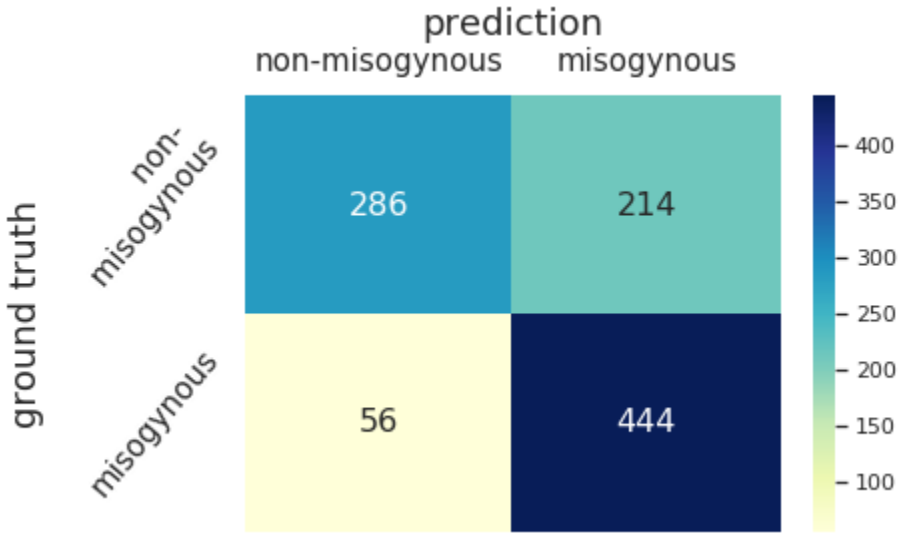


$$\mathcal{L}_{total} = \lambda_{img}\mathcal{L}_{img} + \lambda_{txt}\mathcal{L}_{txt} + \lambda_{kl}KLD + \sum_t \lambda_t \mathcal{L}_t$$

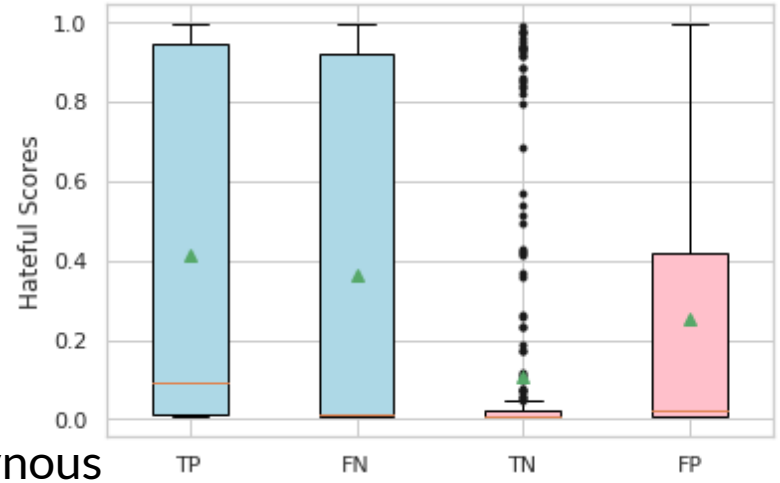
Evaluation

Model	Misogyny prediction			Subcategories prediction		
	Precision	Recall	F1	Precision	Recall	F1
BERT	0.608	0.632	0.589	-	-	-
ResNet-50	0.635	0.656	0.622	-	-	-
CNN-VAE	0.526	0.550	0.462	0.514	0.545	0.469
$MMVAE_{BERT+ResNet}$	0.640	0.653	0.632	0.543	0.590	0.532
$MMVAE_{BERT+CLIP}$	0.707	0.752	0.693	0.586	0.633	0.589
$MMVAE_{LASER+CLIP}$ ★	0.721	0.756	**0.711**	0.594	0.648	**0.600**
$MMVAE_{LaBSE+CLIP}$	0.707	0.751	0.694	0.578	0.658	0.592
$MMVAE_{CLIP+CLIP}$	0.712	0.760	0.698	0.587	0.658	0.592
$MMVAE_{+dropout=0.5}$	0.724	0.759	0.714	0.606	0.656	0.616
$MMVAE_{+dropout=0.2}$ ★★	0.730	0.756	**0.723**	0.613	0.647	0.622
$MMVAE_{+concat}$	0.721	0.751	0.712	0.602	0.657	0.609
$MMVAE_{+more\ layers}$ ★★	0.710	0.750	0.698	0.631	0.649	**0.634**
$MMVAE_{+img\ transform}$	0.710	0.756	0.696	0.605	0.651	0.615

Performance analysis



Correctly classifies 88.8% of misogynous memes yet only 57.2% of non-misogynous memes



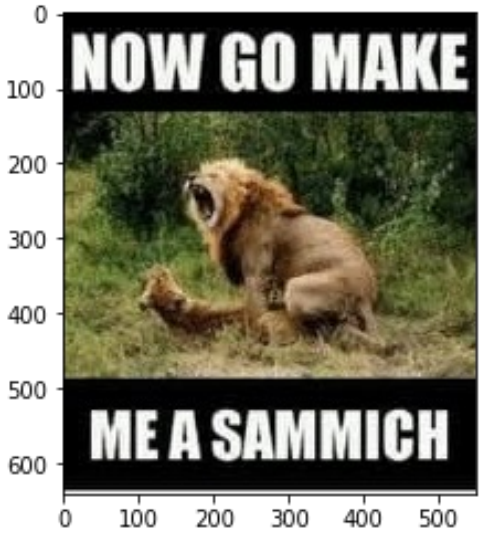
Blue: misogynous
Pink: non-misogynous

Hate score (Pérez et al., 2021)

- Non-misogynous -> lower hate score
- Hate score has some correlation with prediction but not the only factor

Misclassification examples

Model prediction: Misogynous
Ground truth: Non-misogynous



Model prediction: Non-misogynous
Ground truth: Misogynous

Summary

- Goal
  - Build a multimodal hateful meme detection model that gives accurate predictions on granular hateful labels

- Our approach
  - Propose a novel model leveraging multimodal and multitask learning
  - Learn an effective multimodal representation using Variational AutoEncoder

- MMVAE at SemEval-2022 Task 5: A Multi-modal Multi-task VAE on Misogynous Meme Detection
  - Ranked 16/67 in SemEval 2022 task 5

https://github.com/MMVAE-project/MMVAE

Yimeng Gu yimeng.gu@qmul.ac.uk