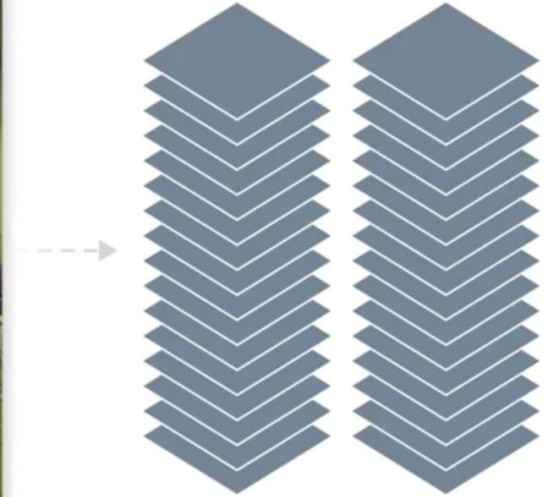# Hateful Meme Classification
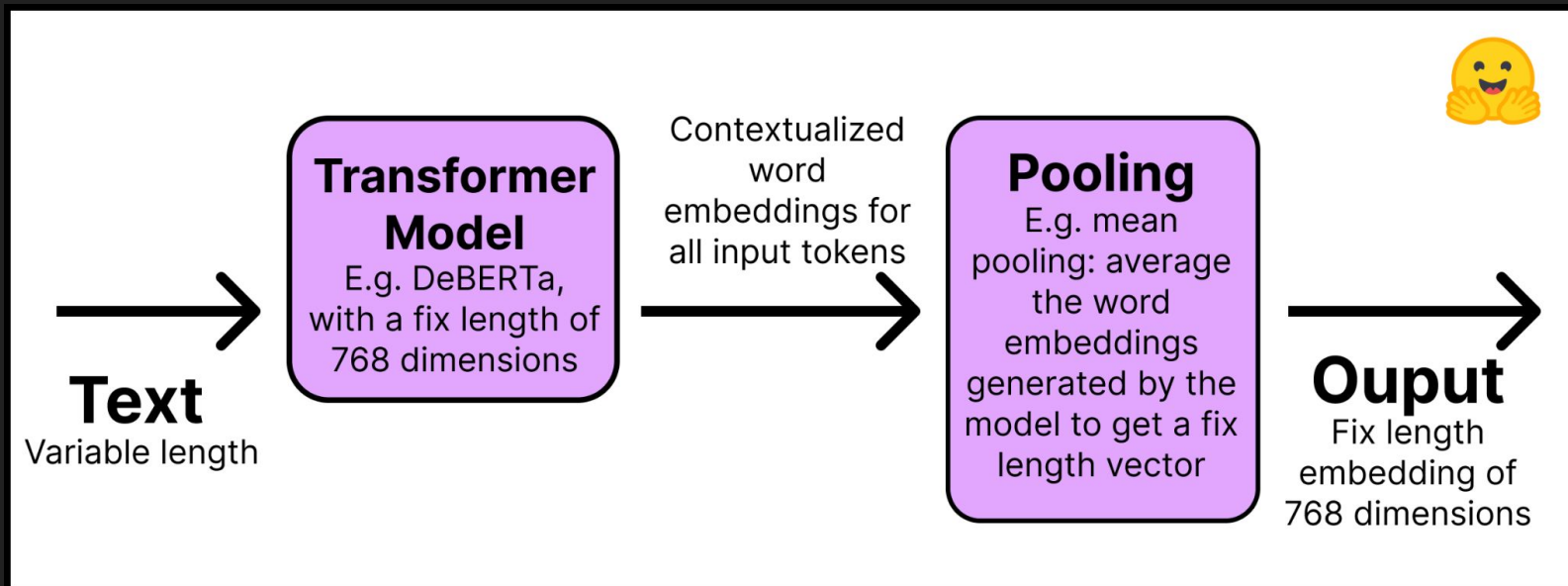
Akansha Lalwani
Data Science
A59019733

Mayank Sharma
ECE
A59019439

Problem Definition and Motivation
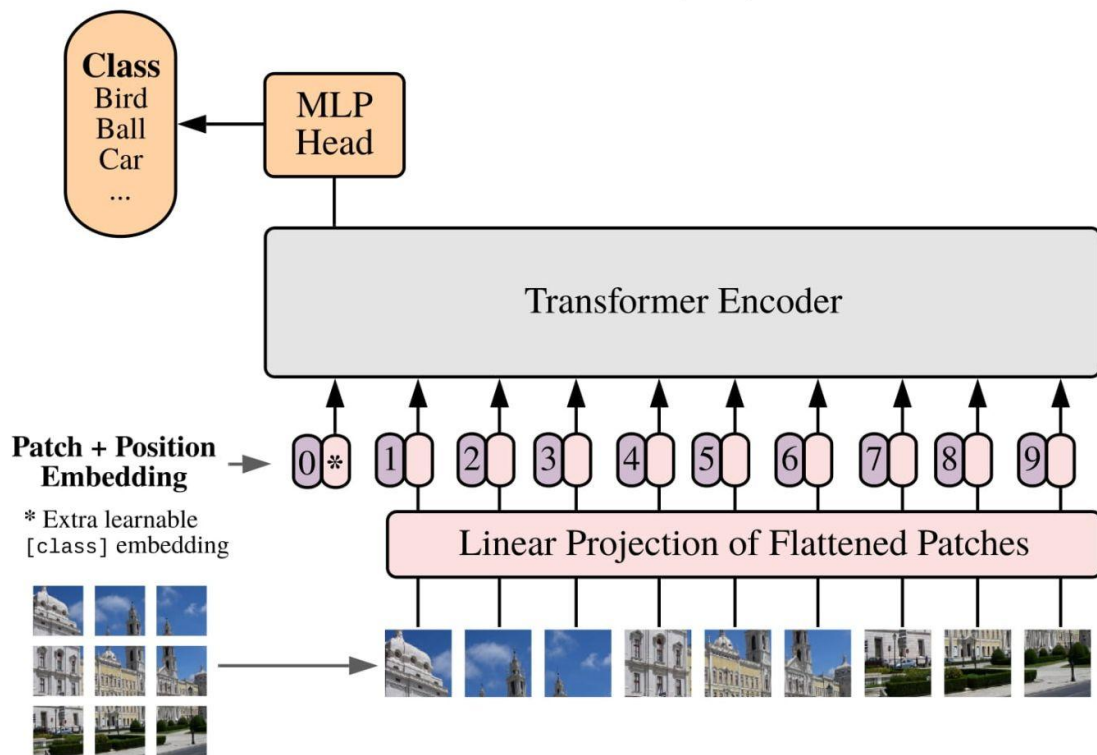
Sentence Transformer
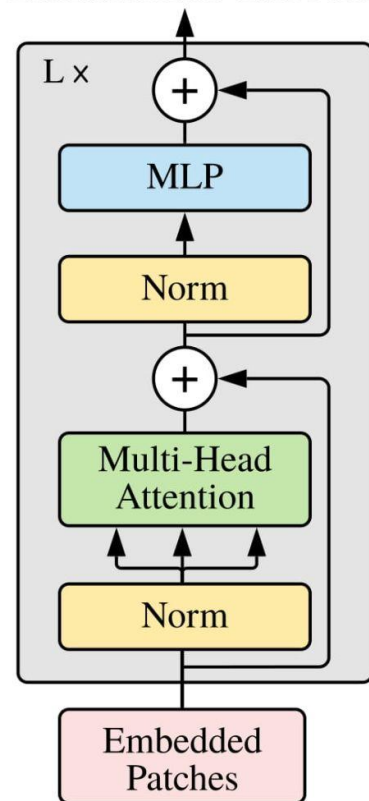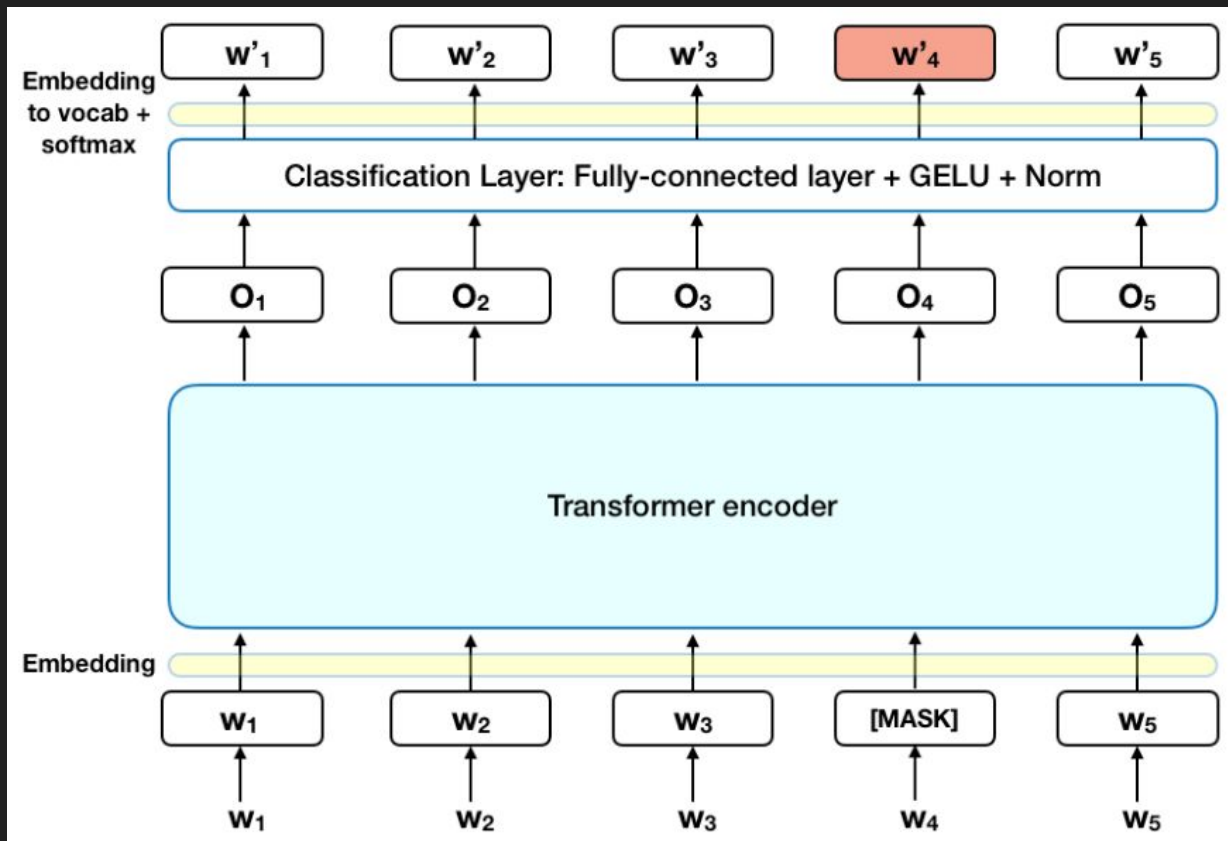
Vision Transformer (ViT)

Vision Transformer (ViT)

Class
Bird
Ball
Car
...

MLP Head

Transformer Encoder

Patch + Position Embedding

* Extra learnable [class] embedding

0 * | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

Linear Projection of Flattened Patches

Transformer Encoder

L ×

MLP

Norm

Multi-Head Attention

Norm

Embedded Patches

BERT

ResNet101
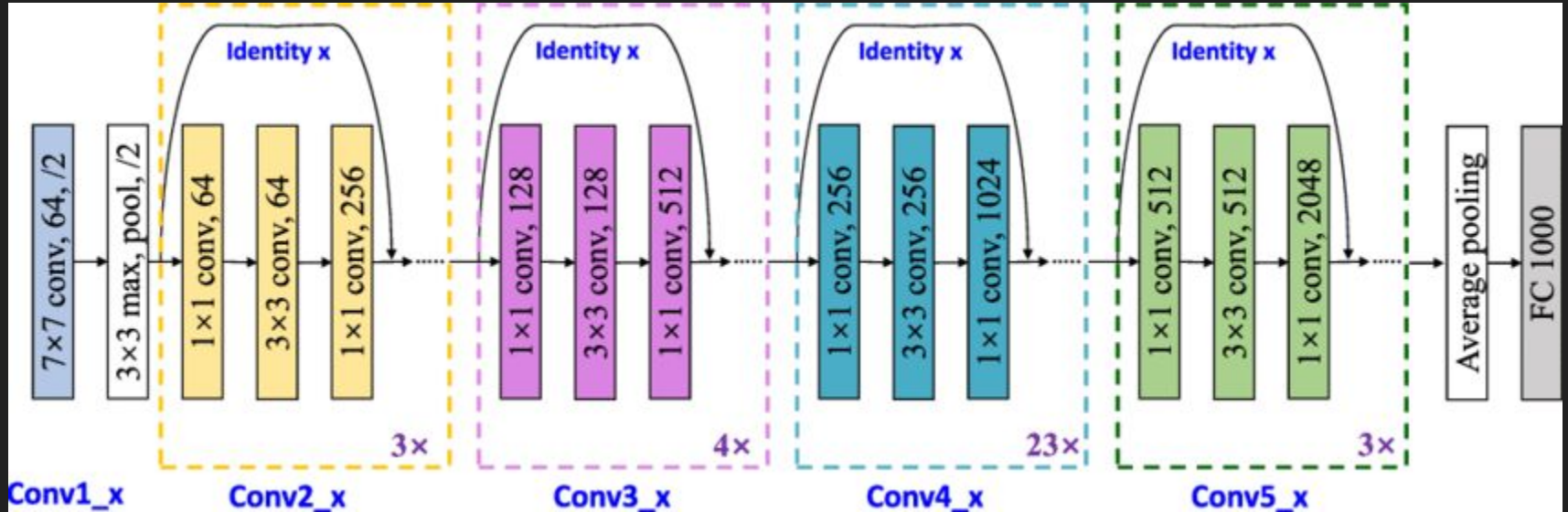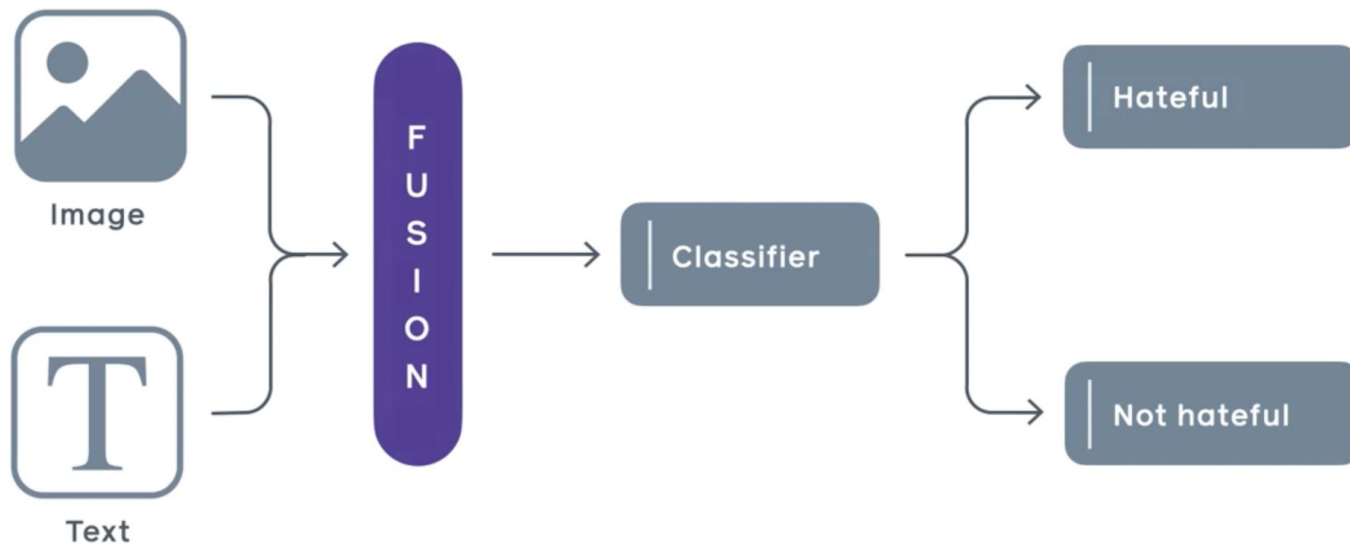
Early Fusion

Late Fusion
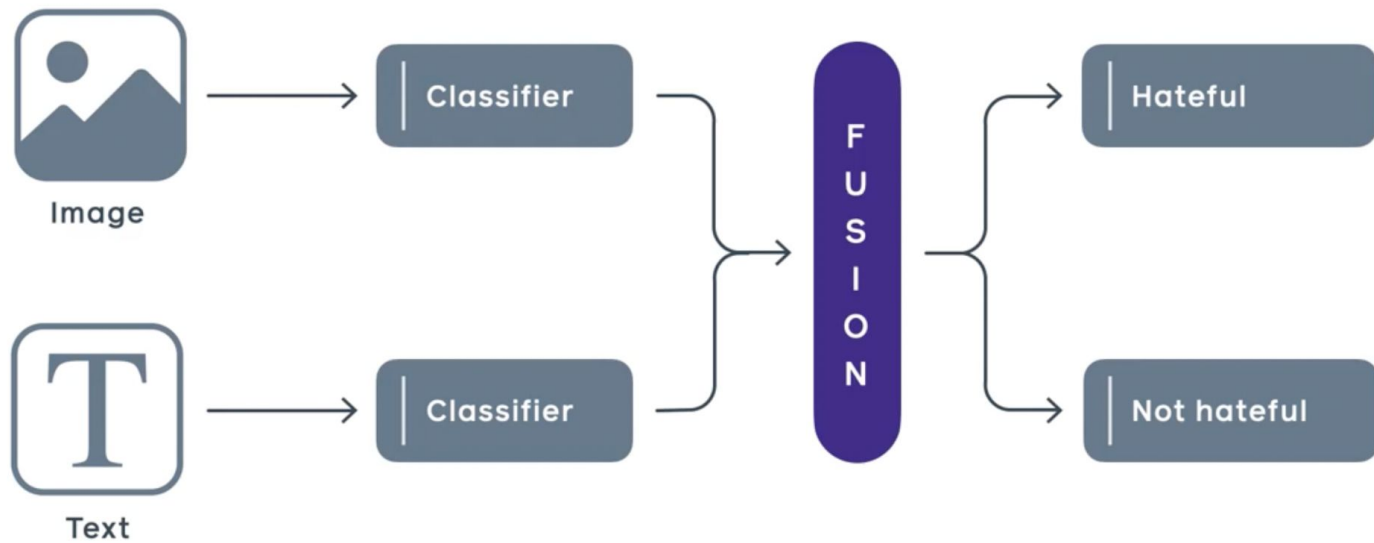
Experiment Results

Table 1: Final model results

Type	Model	Validation		Test	
		Acc.	AUROC	Acc.	AUROC
	Human	-	-	84.70	82.65
Unimodal	Sentence Transformer (ST)	54.80	54.80	55.00	54.60
	Vision Transformer (ViT)	61.60	51.28	52.00	51.10
Multimodal (Unimodal Pretraining)	Early Fusion: BERT + ViT	52.00	52.6	52.6	51.70
	Early Fusion: BERT + ResNet	53.40	53.39	53.20	52.80
	Late Fusion: ST + ViT	52.00	52.00	52.60	52.07

Thank you