FUSING VISUAL AND TEXTUAL INFORMATION TO DETERMINE CONTENT SAFETY

Rodrigo Leonardo[1] Amber Hu[2] Mohammad Uzair[3] Qiujing Lu[4]
Iris Fu[5] Keishin Nishiyama[5] Sooraj Subrahmannian[5] Divyaa Ravichandran[5]

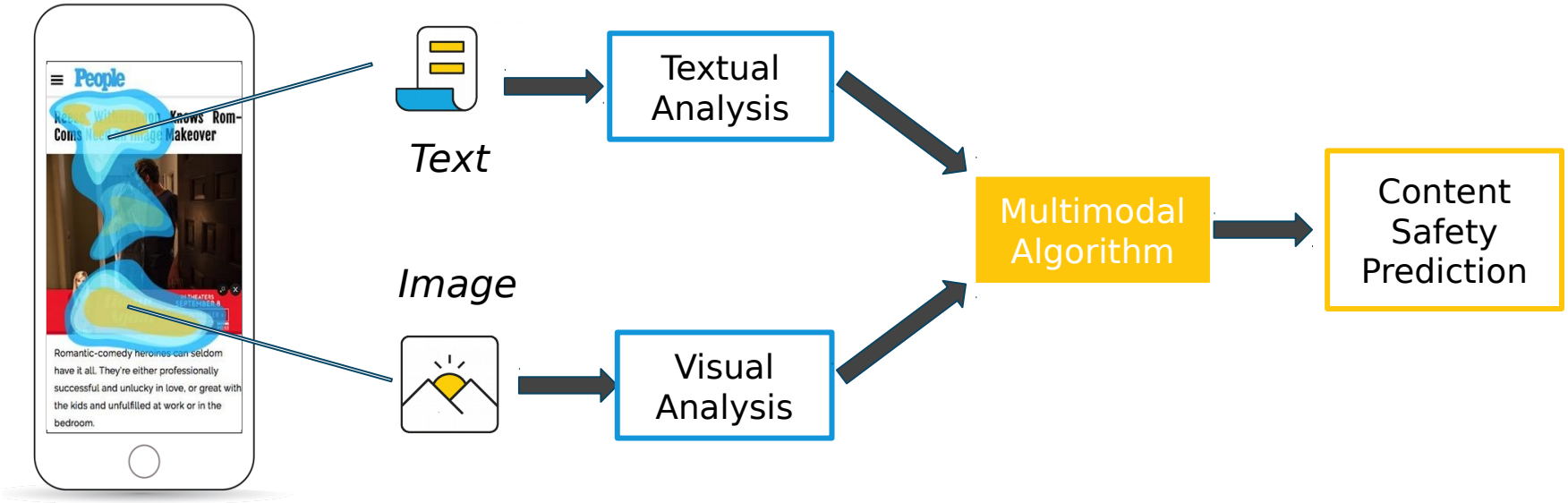
[1]Universidad del Valle de Guatemala
[2]Yale University
[3]National University of Science and Technology
[4]University of California Los Angeles
[5]GumGum Inc.

18th IEEE International Conference on Machine Learning and Applications
December 16-19, Boca Raton, Florida, USA

Project jointly sponsored by the Institute for Pure and Applied Mathematics at UCLA, GumGum Inc. and NSF Grant (DMS) 1440415

MOTIVATION & PROBLEM STATEMENT

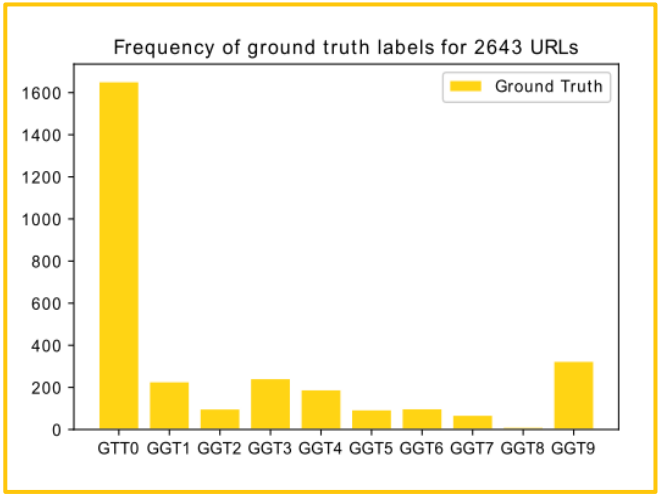
- Our aim is to classify the content safety of web pages containing images and text using multimodal machine learning algorithms
  - Fusion of information from computer vision and natural language processing models



UNDERSTANDING THE DATA

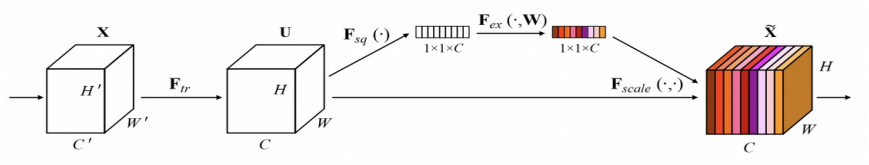
Threat Category	Description
GGT0	Safe
GGT1	Violence/Gore
GGT2	Criminal
GGT3	Drugs and Alcohol
GGT4	Sexually Charged
GGT5	Obscene/Disgust
GGT6	Hate
GGT7	Disasters
GGT8	Malware
GGT9	Illness

Web pages



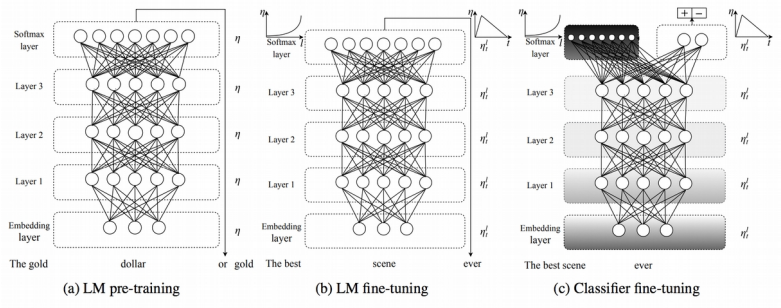
COMPUTER VISION & NLP MODELS

Computer Vision model: Squeeze-and-excitation network[1]



$$[0.9, 0.0, 0.0, \ldots, 0.2]$$

Natural Language Processing model: Universal Language Model Fine-tuning[2] (ULMFiT)



$$[0.7, 0.3, 0.0, \ldots, 0.0]$$

Each model produces a 10-dimensional vector, where each entry represents the probability that the sample is classified into a threat category (e.g. 0th entry is probability of GGT0)

[1]J. Hu, L. Shen, and G. Sun, ``Squeeze-and-excitation networks,'' in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-–7141.
[2]J. Howard and S. Ruder, ``Fine-tuned language models for text classification,'' Jan. 2018.

BASELINE & LATE FUSION

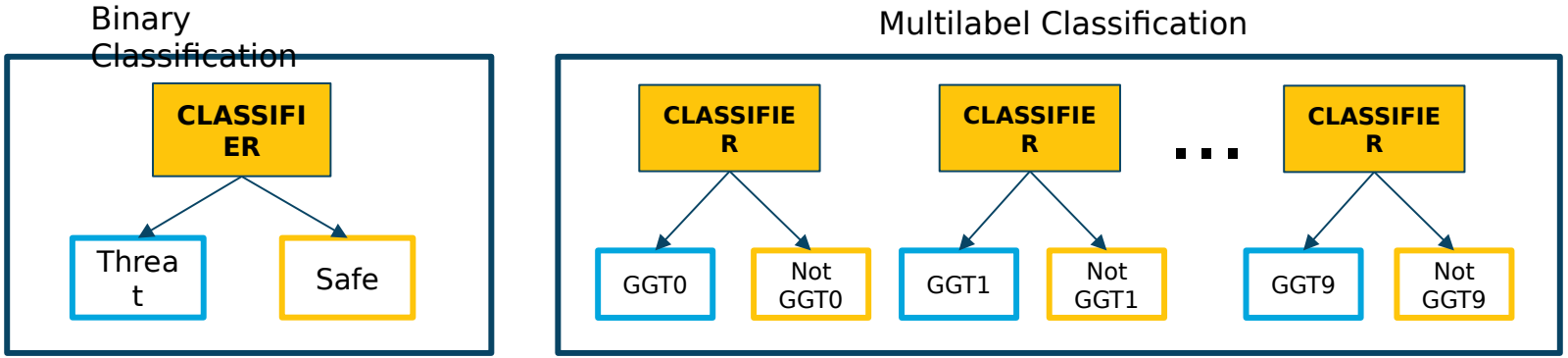
Results from NLP model:
Text: SAFE

OR Logic

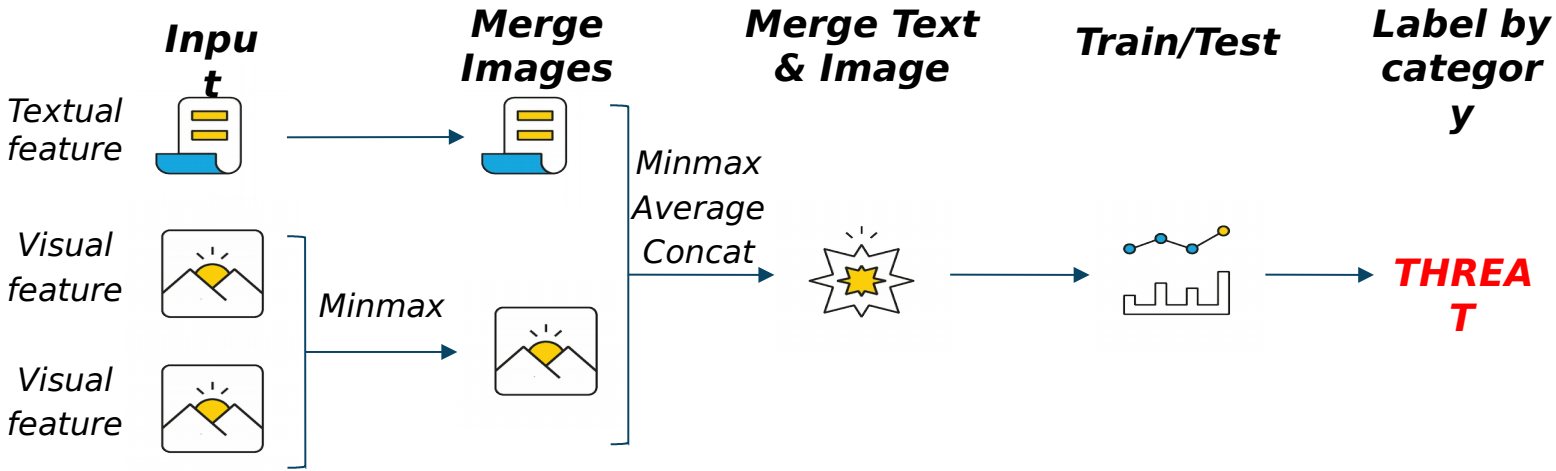
Web page classified as ***THREAT***

Results from CV model:
Image 1: SAFE
Image 2: SAFE
Image 3: ***THREAT***
Image 4: SAFE

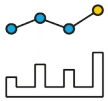
Late fusion: extract **output** features from the CV and NLP models, then train a classifier on these features for two classification tasks:



OUR LATE FUSION APPROACH

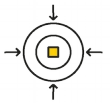


Using these classifiers:



SUPPORT VECTOR CLASSIFIER

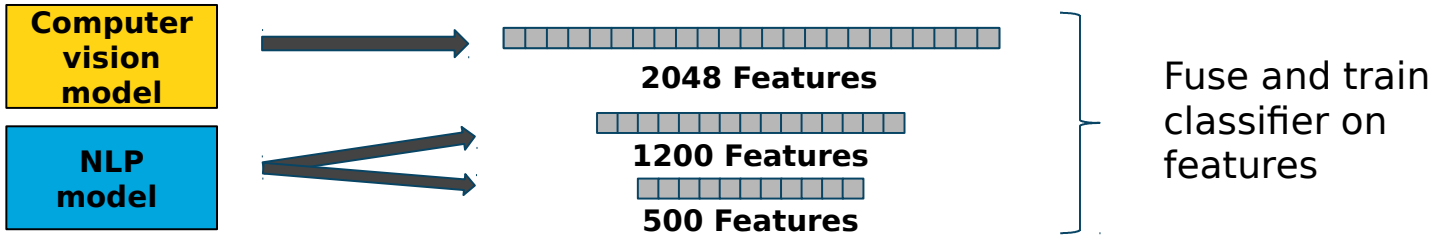
RANDOM FOREST CLASSIFIER



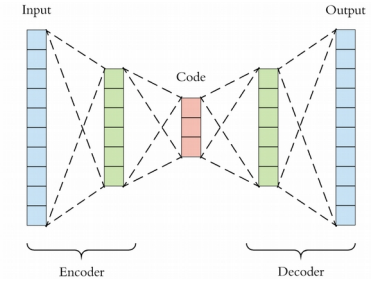
CONVEX OPTIMIZATION

EARLY FUSION BACKGROUND

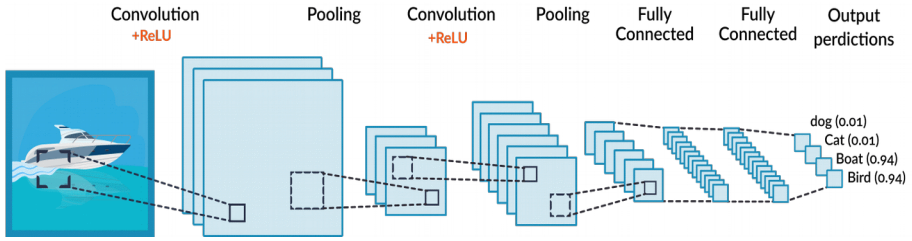
Early fusion: extract **intermediate** features from the CV and NLP models, then train a classifier on these features



Two methods to deal with these higher-dimensional features:



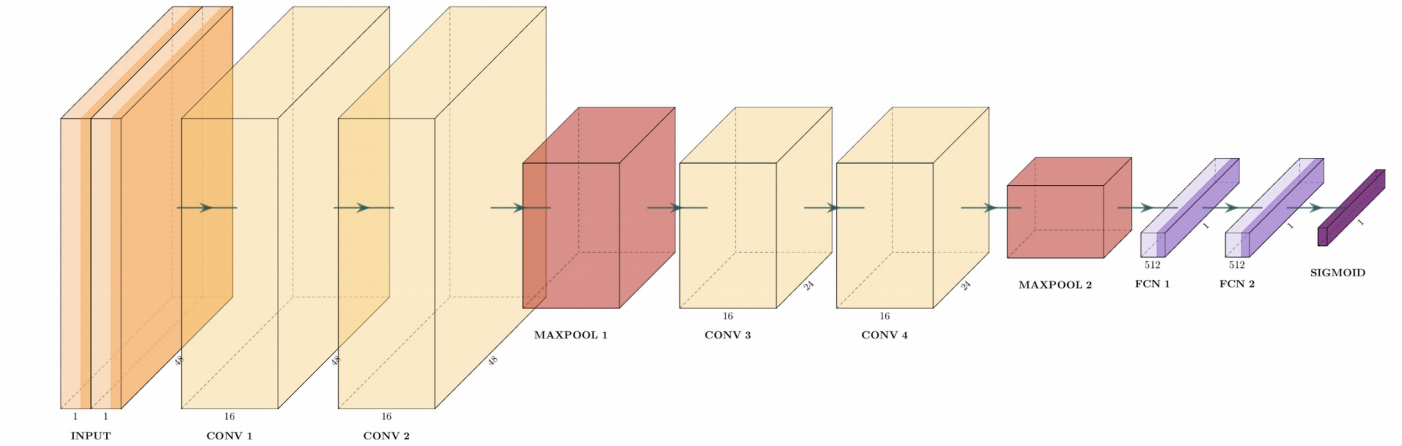
Autoencoder



Convolutional Neural Network

OUR EARLY FUSION APPROACH

- **Method 1:** Autoencoder for dimension reduction, followed by classification
  - Separate autoencoders for visual and textual features
  - Multi-layer autoencoder architecture to reduce visual and textual features each to length 128
  - Concatenation to fuse features
  - Classification using random forest
- **Method 2:** CNN for binary classification

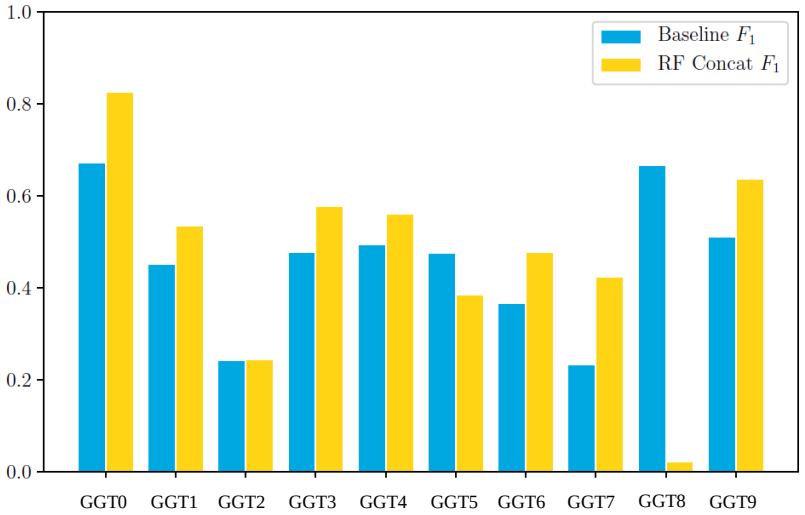


RESULTS

Binary Classification

	Model	F_1	F_2	Precision	Recall
LF	CO Minmax	0.615	0.605	0.633	0.598
	CO Average	0.641	0.672	0.596	0.694
	SVC Minmax	0.656	0.711	0.581	0.753
	SVC Average	0.673	0.727	0.600	0.767
	SVC Concat	0.697	**0.753**	0.621	**0.795**
	RF Minmax	0.691	0.742	0.620	0.781
	RF Average	0.689	0.744	0.614	**0.785**
	RF Concat	**0.716**	0.748	0.668	0.772
EF	Dim. Red.	0.702	0.700	**0.705**	0.699
	CNN	**0.753**	**0.795**	0.691	0.826
	Baseline	0.622	0.739	0.492	0.845

Multilabel Classification



Thank you!
Questions?