Research Title: Multimodal Sarcasm Detection

Junzhe Huang, Wenyun Xiang

Introduction

Mainly focused on developing several methods to implement multimodal sarcasm detection from both images and text. Also, we quantify the performance of different models to give a comprehensive comparison.

Single-source methods:

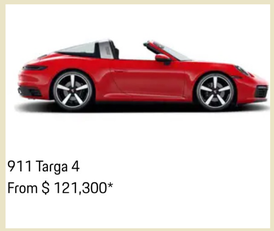
- Pretrained ResNet-50 and DenseNet-121 (Image only)
- Bi-LSTM and pretrained BERTweet(Text only)

Multimodal methods:

- Hierarchical Fusion Model from Cai et, al. as baseline
- Baseline improvement in both text modality and Image modality
- Multimodal learning with ResNet-50/DenseNet-121 and BERTweet
- Averaging ensemble learning

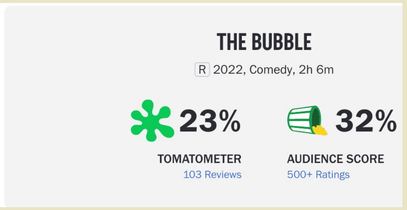
Sarcastic Examples

"what a lovely weather!"



"This car is so cheap! "

"I am satisfied 😊 "



"This movie is fantastic ! I love it!"

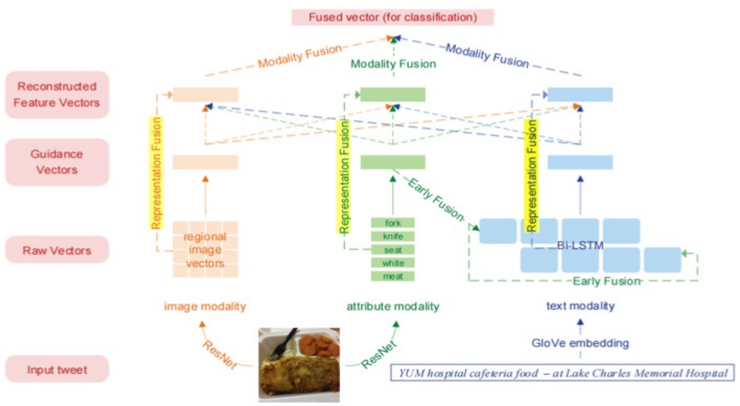
Dataset

Split	Text count	Image count
Train	29,040	19,816
Validation	2,410	2,410
Test	2,409	2,409

- Below 300 characters length
- Below 50 words count
- Data Augmentation

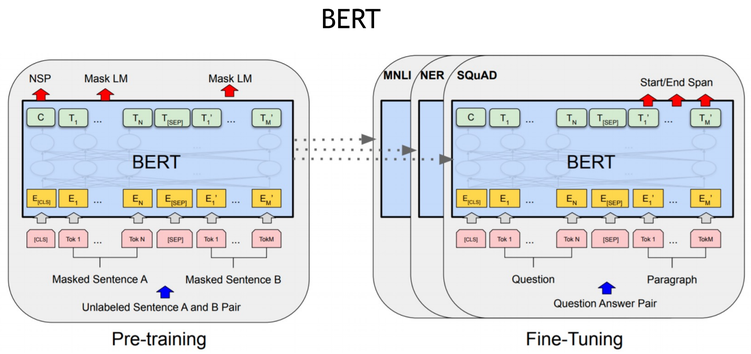
Methodology

Baseline: hierarchical fusion model

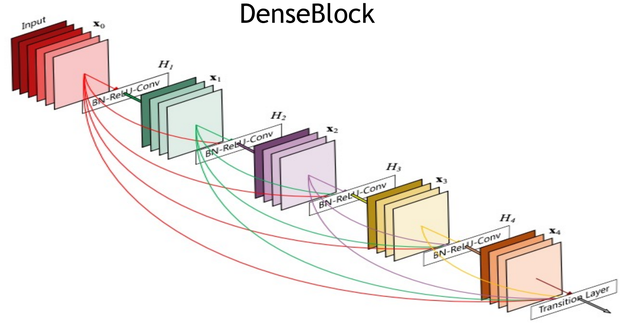


Methodology

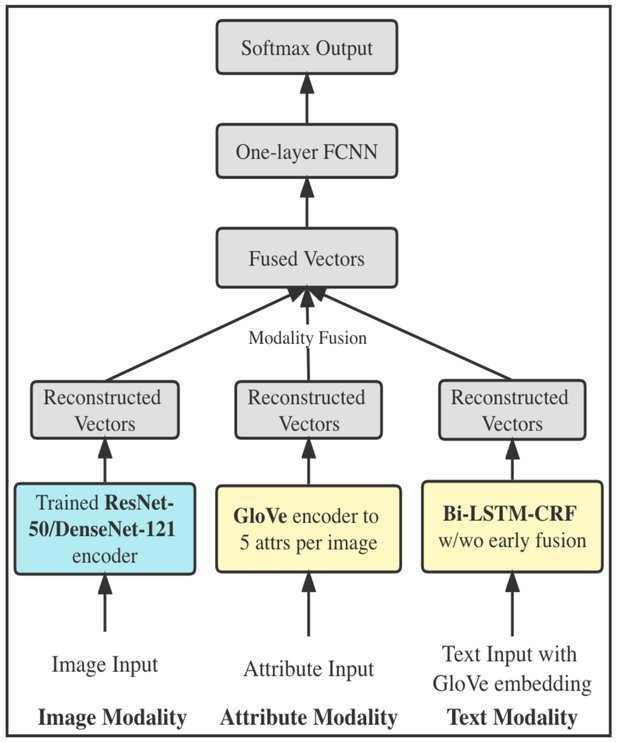
BERT



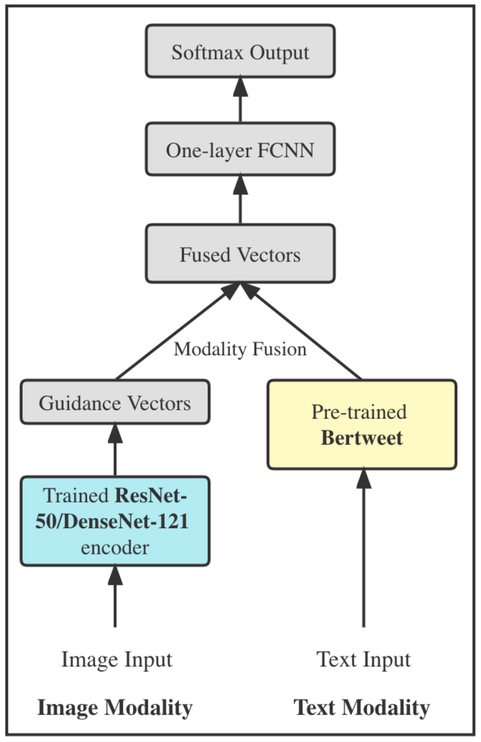
DenseBlock



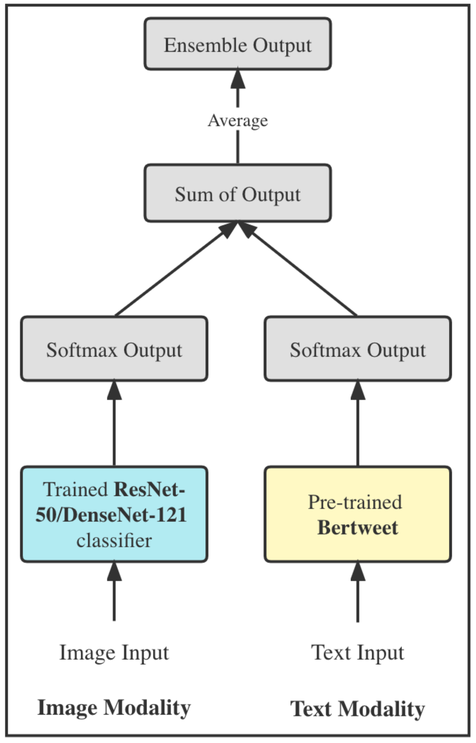
Improved Hierarchical Fusion Model



Multimodal learning



Ensemble learning



Experimental Results

Model	Accuracy	F1	Model	Accuracy	F1	Model	Accuracy	F1
ResNet - Image only	0.6650	0.6412	Baseline with CRF (without early fusion)	0.8172	0.7670	Ensemble learning with ResNet	0.9110	0.8940
Bi-LSTM + Attribute - Text only	0.8190	0.7753	Baseline with CRF (with early fusion)	0.8382	0.8154	Ensemble learning with DenseNet	**0.9130**	**0.9080**
BERTweet - Text only	0.9020	0.8820	Baseline with DenseNet + CRF (without early fusion)	0.8321	0.8140	Multimodal learning with ResNet	0.8732	0.8330
Baseline	0.8313	0.8119	Baseline with DenseNet + CRF (with early fusion)	**0.8582**	**0.8315**	Multimodal learning with DenseNet	0.8770	0.8344