MP1.PC.1

Reading Is Believing: Revisiting Language Bottleneck Models for Image Classification

Honori Udo*, Takafumi Koshinaka (Yokohama City University)

* Honori Udo is now with NTT Comware Corporation

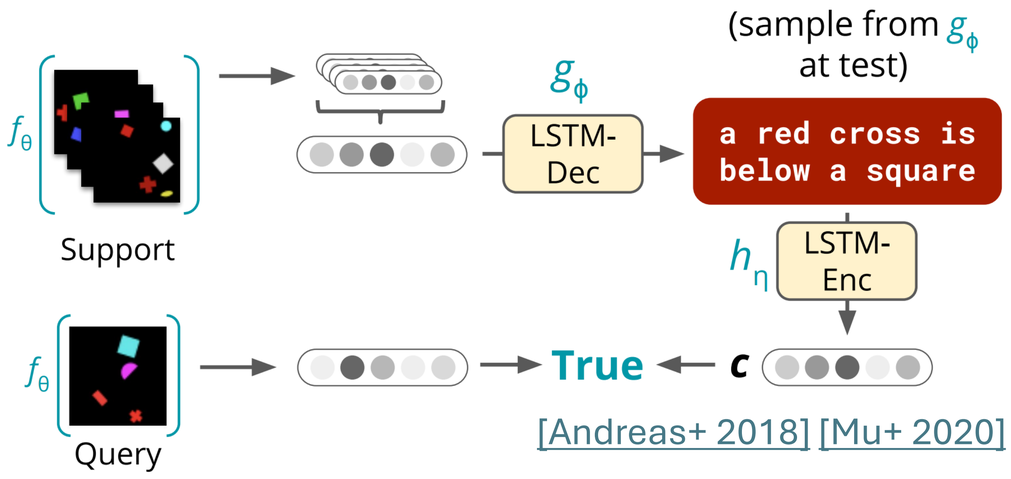
ICIP abu dhabi 2024
IEEE International Conference on Image Processing
27 - 30 October 2024 | Abu Dhabi, UAE

Background

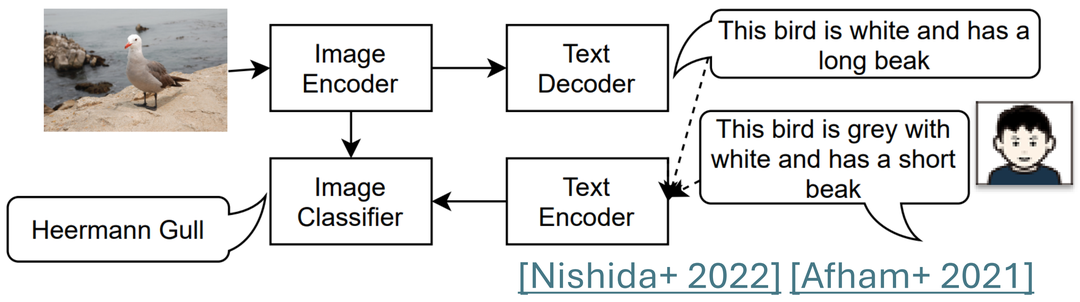
- The black-box nature of deep learning models often hinders the practical application of those models
- Using a set of human-readable features is a promising approach to eXplainable AI (XAI), e.g., concept bottleneck models [Koh+ 2020]

Language Bottleneck (LB) Models

- Have been studied in few-shot scenarios
- Linguistic knowledge within image captioning models helps understand abstract concepts as well as recognize unseen image classes



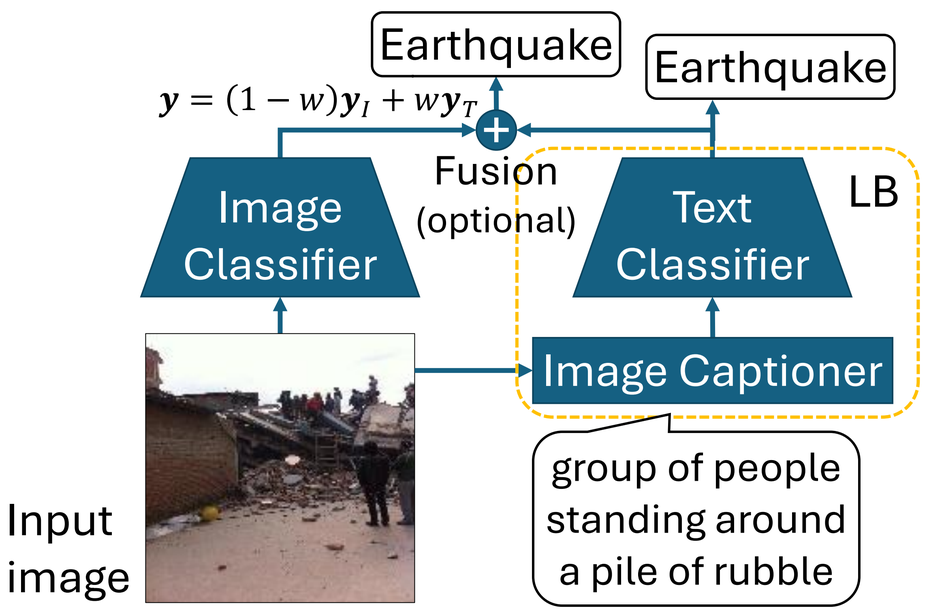
[Andreas+ 2018] [Mu+ 2020]



[Nishida+ 2022] [Afham+ 2021]

Objectives

- However, in more fundamental many-shot settings, LB models generally perform worse than standard (black-box) CV models
  - Because of information loss incurred in the step of converting images into language
- Recent foundation models for image captioning, on the other hand, are capable of describing images with great accuracy and detail



We (1) evaluate LB with modern image captioning models in a many-shot setting and (2) try fusing them with standard CV models

Experimental Setup

- Dataset: CrisisNLP [Alam+ 2021]
  - Natural disaster images shared on social media
  - "Disaster Types" is a classification task with 7 classes: earthquake, fire, flood, hurricane, etc.

# images	
Train	12,724
Dev	1,574
Test	3,213

- Models

Classifiers		Image captioners	Specs
Image-based (black-box)	ResNet-50/101	InceptionV3+RNN	TF tutorial
	ViT-Base/Large	BLIP	ViT-L
Text-based	$BERT_{BASE}$	BLIP-2	ViT-g + OPT-2.7B
		CLIP Interrogator	BLIP + CLIP

Experimental Result

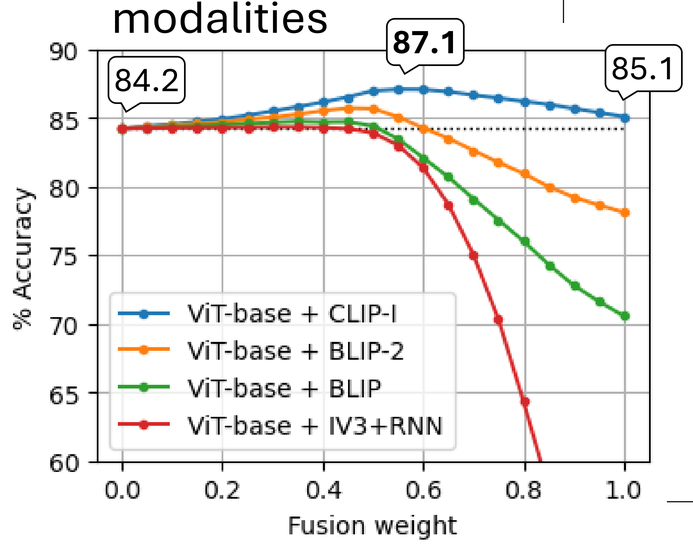
- Performance of image/text single-modal models

Model	%acc
ResNet-50	78.38
ResNet-101	79.71
ViT-Base	84.22
ViT-Large	82.01
IV3+RNN	42.38
BLIP	70.55
BLIP-2	78.11
CLIP-I	**85.09**

Example input Image (hurricane)

ViT-Base	→ **not disaster**
IV3+RNN	a snowboard near another wave in the water → **flood**
BLIP	an oil rig in the middle of the ocean on a foggy day → **not disaster**
BLIP-2	a large wave is crashing over the ocean → **hurricane**
CLIP-I	a large body of water with a boat in the distance, stormy seas, stormy sea, rough seas, tumultuous sea, rough sea, violent stormy waters, storm at sea, rough water, apocalyptic tumultuous sea, a violent storm at sea, towering waves, sea storm, in rough seas with large waves, rough seas in background, stormy wheater → **hurricane**

- Fusion of image/text modalities



Summary

- Modern captioning models can be powerful and explainable feature extractors for image classification
- A captioning model and a standard CV model see images differently so that fusing the two achieves even better performance
- We plan to verify our finding with more diverse datasets from different domains

Takafumi Koshinaka Laboratory
School of Data Science, Yokohama City University
22-2 Seto, Kanazawa, Yokohama 236-0027, Japan

横浜市立大学 YOKOHAMA CITY UNIVERSITY