A table is worth a thousand pictures: Multi-modal contrastive learning in image classification with tabular data

Iván Higuera-Mendieta*, Jeff Wen, Marshall Burke

Stanford | Doerr
School of Sustainability

Introduction

Multi-modal learning presents an opportunity to augment image classification tasks using new, efficient fine-tuning strategies while avoiding pre-training. We propose a CLIP-like architecture, where by using contrastive learning we put image and text embeddings into the same embedding space for classification. We apply this learner to a simple downstream binary task:

Is this house going to burn?

We use the modal approaches in the literature as baselines, and found that our proposed alternative behaves better across our experiments.

Data and baselines

NAIP Aerial Imagery

— 0.6 m/px —

Pre-fire imagery for houses in California
[+] ~5,000 [-] ~ 3,000

Change weights in CE loss

$$\left(w_c = \frac{N}{n_c * 2}\right)$$

Tabular data (GridMet + CalFire)

price	severity	avg_temp	wind_spd
860K	1.56	23	40
850K	0.8	26	9
1.6MM	0.4	29	18
750K	1.67	20	26

…transformed into text prompts

```
An {unburned} house value is  $1.6MM.
The temperature is 29C, and the wind
blows at 18 km/h.
```

Baselines [ResNet50 & XGBoost]

We use a RestNet50 and an XGBoost as baselines for image and text, respectively. These are the modal approaches to single-modal classification for similar tasks.

	Unbalanced	Acc	F1
ResNet50	Yes	0.61	0.56
ResNet50	No	0.72	0.67
XGBoost	Yes	0.76 ± 0.01	0.64 ± 0.03
XGBoost	No	0.61 ± 0.01	0.501 ± 0.02

All the models are trained with a 70/15/15 split for train/evaluation/validation with dropout and regularization. We completely fine-tune each model and in the case of the ViT, we picked the best set of parameters using a parameter sweep.

Methods and experiments

1. Fine-tuning multi-modal learner

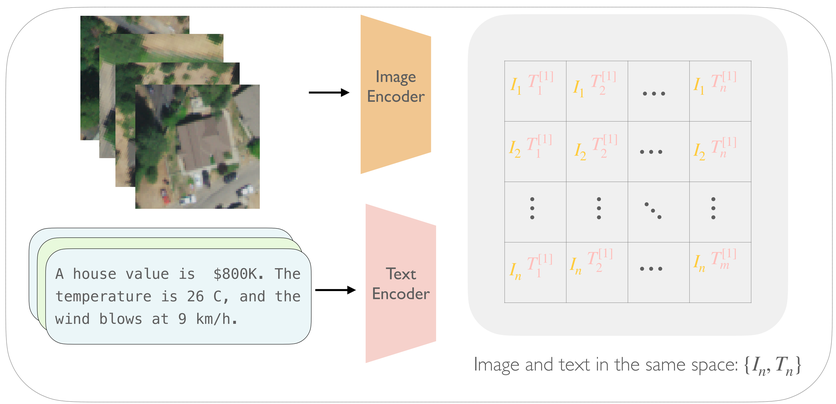


Image and text in the same space: $\{I_n, T_n\}$

2. Evaluate the model with prompts and images $\{I_n, p_n, \tilde{p}_n\}$

Test Image

True prompt p_n

```
This house value is $1.6MM. The
temperature is 29C, and the wind
blows at 18 km/h. This house burned
```

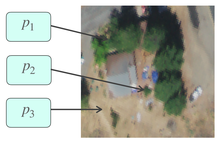
Fake prompt $\tilde{p}_n$

```
This house value is $1.6MM. The
temperature is 29C, and the wind
blows at 18 km/h. This house did not
burn
```

We classify the best image/prompt combination for each test set by calculating:

$$\mathbb{P}(y^{(i)} = 1) = \arg\max \sigma(z^{(i)})$$

Experiments



Many features, many prompts:
$\{x^{(i)}, t^{(i)} \in \mathcal{P}_n^{(i)}\}$

No prompt (ViT)	∅
All $X^{[N]}$	{column name 1}: {value}, and {column name 2}: {value}, …
House features	-Age -Location -Elev -Slope …
Fire weather features	-Temp -PVD -Humidity -Wind …

Results

Experiments F1-Scores

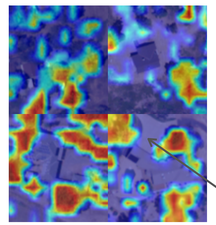
Text Encoder	House Features	Weather Features	All
BERT	0.839	0.609	0.734
RoBERTa	**0.867**	**0.638**	**0.859**
GTP-2 `medium`	0.810	0.589	0.738

All w/ vision encoder: ***`google/vit-base-patch16-224`***

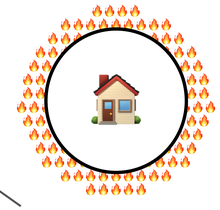
We use only 500 of our training observations to fine-tune all the models. We found that models plateau adding more shots. Under 10 shots gives low performances.

Weather features seem to be less explanatory than the house features! No prompting (only vision) F1 is: 0.689 ± 0.03.

Interpretation



GradCAM for our best vision/text model for images of households that burned



Defensible space [~30 ft.]

SHAP values for examples in the only-fire-weather features experiment.

As more pre-trained models using environment/climate datasets are available, we can improve the ability of foundation models to adapt to tasks and OOD data.