A table is worth a thousand pictures: Multi-modal contrastive learning in house burning classification in wildfire events

Anonymous Author(s)
Affiliation
Address
email

Abstract

Wildfires have increased in frequency and duration over the last decade in the Western United States. This not only poses a risk to human life, but also results in billions of dollars in private and public infrastructure damages. As climate change potentially worsens the frequency and severity of wildfires, understanding their risk is critical for human adaptation and optimal fire prevention techniques. However, current fire spread models are often dependent on idealized fire and soil parameters, hard to compute, and not predictive of property damage. In this paper, we use a multimodal model with image and text embeddings that allows both image and text representations in the same latent space, to predict which houses will burn down in the event of wildfires. Our results indicate that the DE model achieves better performance than the unimodal baselines for image-only and text-only models (i.e. *ResNet50* and *XGBoost*), and text or vision only models. Moreover, following other models in the literature, it outperform these models also in low-data regimes.

1 Introduction

As the frequency and severity of wildfires surge around the world, so do its socio-economic consequences. According to the National Interagency Fire Center, wildfires generate more than 30 billion dollars in capital losses every year in the United States. In California alone, the 2022 fire season incurred 380 million dollars in losses from capital destruction and fire-fighting efforts. Property fuel management policies have been central to manage property burning risk. While changes in building codes have decreased the risks of property burning [2], these risks and its costs are projected to increase as the wildland-urban interface (WUI) footprint expands and climate change increases the frequency of wildfires around the globe [18, 8]. One of the most widely supported risk management strategies is to create a fuel-free *defensible space* surrounding houses and other structures [33], but often other property characteristics, such as the building materials, the spatial arrangement of the property footprint, or the fire weather can dramatically change the burning probabilities [20].

Literature exploring these property burning risk have rely on qualitative assessments [5] or regression analysis [33] combining remote-sensing outcomes and house features. Other literature focused on prediction tasks, has mainly pivoted around burned-area segmentation [4, 30], and fire spread modeling [10, 11], but not directly in property destruction as a prediction task. Fire spread and hazard models, while seemingly useful for this classification task, do not perform well when trying to predict property burning [34], and are not suitable for real-time fire estimation because of their computationally complexity. Thus, existing methods do not produce immediately actionable insights for land managers and emergency responders in wildfire-prone areas to minimize fire property damages.

Submitted to Computational Sustainability Workshop at NeurIPS 2023. Do not distribute.

Machine learning applications in sustainability have been predominantly dominated by vision tasks. These comes as satellite imagery has become abundant and readily accessible to researchers . Nonetheless, vision-only models forgo data available in more traditional formats for the social scientists and ecologists, such as tabular data. This presents trade-offs to researchers when training predictive models, where they would either featurize image data and combine it with tabular data in tree models (*i.e.* RandomForest or XGBoost) [15], or forgo tabular data and fine-tune deep learning vision models. The former strategy would miss possible spatial patterns that deep learning architectures can identify and generalize, while the latter will miss important non-visual data that can improve [36].

Multimodal models have been used for classification [12, 22, 21] and captioning [19, 26]. In these models, different data modes can be combined at different stages of the learning process [31]. In *early* fusion, the inputs (i.e. text and image) are combined and a common representation is learned, whereas in *late* fusion separate models learn each data mode before fusing the results into a single prediction. When data modes distributions lack a large common support, alternative fusion architectures can help to align data modes. CLIP [27], and other derivative models using contrastive approaches [22, 38, 37] have shown how we can use dual and multiple encoder models with a contrastive loss to cross-align different modes of data in the same latent space. Fine-tuning these models to new tasks, or adding projection heads after building embeddings [23] has shown performance gains [14, 24] while keeping its *few-shot* abilities.

New ways of representing tabular data as text using large language models (LLM) has opened new alternative for multimodal classification. TabLLM [17] have leveraged LLM for few-shot classification using tabular data by fine-tuning the `T0` model to different classification tasks. TabLLM serializes each row into a text prompt representation and a short description of the classification problem (*i.e.* `Is this house going to burn?`), and outperforms tree-based methods using fewer observations. LIFT [6] follows a similar approach by directly fine-tuning the LLM using the serialized tabular data to both classication and regression problems using a "no-code" interface where the prompt includes the prediction task (i.e `If x = 0.5 and y = 0.2, then z is`). As with TabLLM, LIFT has similar or better performances than tree-based models, although this performance decreases as the number of classes increases or if the features have large dimensions.

In this paper, we want to assess the prediction lift from adding tabular data as text prompts into a multi-modal classification task of house burning in California. To do this, we will combine pre-fire aerial imagery from houses, and tabular data including house characteristics, weather variables, and fire hazard scores. We will transform these data into different text prompts to be coupled with labeled images of houses [17]. We will run experiments combining different text-model encoders with a fixed vision encoder, and assess their performance against vision and text-only baselines.

2 Related literature

Prediction of property destruction in wildfire settings must account for different physical and property factors. Houses' fuel availability in their *defensible space* is not the only factor that affects their risk of fire, but also fire conditions and fire weather that can affect ember transport [5]. Fire modeling has been used by the United States Forest Service as the main tool to address property prediction damage and prioritize local fire suppression responses.

Numerical models that solve different fire spread and fuel-weather interaction equations to generate fire perimeters for a determined time frame are usually used on different time steps to predict the margins of a fire. Models like FARSITE [10] and FlamMap [9] are some of the production models used by the Forest Service for fire events in the USA. They use spatial information on weather, topography, fuels, and vegetation parameters. Although some of these information is near-real time available, some field critical information, as fuel consumption and fire spread rate, are often scant during fire events due to the risk to scientists on the field and measuring difficulty. To yield accurate results tuned to local conditions, numerical models' predictions need to be calibrated and these hard to collect critical fire features are the ones that the models are more sensible to [32].

More recently, MCTS-A3C [13] an agent-based model has been used to predict fire spread using a Markov Decision Process. Just like the numerical models, MTCS-A3C starts from an ignition point and generates a fire perimeter depending on weather and fire start characteristics. Other machine learning approaches include FireCast [29], a CNN-based approach using weather data to make fire predictions a day-ahead. Where as useful for fire boundary prediction, houses are often within fire boundaries and they do not necessarily burn, thus having a model that is able to predict burning within fire boundary is relevant for targeting fire responses and prevention.

3 Methodology

3.1 Data

For our binary classification task, we use a geo-referenced dataset of homes exposed to wildfire contacts in California between 2015 and 2020 ($n = 39,718$) collected by CALFIRE's the Damage Inspection program (DINS). The geo-referenced dataset contains an assessment of all burned and unburned properties within the boundaries of a wildfire with. We augment these data with high-resolution weather data ($\approx$ 4 km) from GridMet [1], to capture different weather variables during the wildfire event corresponding to each house in our sample. Since we want to predict fire destruction before the fire event, we use the average month weather variables before the fire event. We extract temperature, humidity, and wind-speed, although we are particularly interested in Vapor Pressure Deficit (VPD) since indicates the level of humidity saturation in the air and is predictive of fire spread [18].

Figure 1: **Sample of NAIP Labels**: These are four examples of our NAIP samples. The two houses on the first column were destroyed, whereas the ones in the second column survived the fire. Notice the image in the right-upper corner represents some of the labeling issues in out database, we remove all image labels where more than 95% of pixels are vegetation (using the NVDI).

Using the coordinates from each house plot, we extract images for each house before a fire event from the National Agricultural Imagery Project (NAIP), a yearly aerial imagery survey with very-high resolution ($0.6m/px$) for all the continental US run by the US Department of Agriculture. NAIP covers California during the growing season, April to August, which overlaps with the state's fire season. The NAIP labels might contain more than one house in the case of plots overlap (i.e. houses in a *cul-de-sac*) introducing the possibility for false-positive or false-negative events. We try to alleviate this problem by excluding houses that overlap with other houses within a 40 meter radius, this reduces the sample of total houses to $9,256$. Figure 1 shows some of the sample labels in our dataset.

3.2 Baselines

To build a vision-only baseline we full fine-tune a ResNet50[1] using our dataset. During learning, We use the *Adam* optimizer with an decaying schedule learning rate, and a weight decay of 10^{-3} for L2 regularization in our loss. Given the nature of our dataset, and the local randomness of fire exposure, we have an unbalanced data set. To correct for this we changed the batch sampling to always have a balanced sample or change the weights on the cross-entropy loss to give more weight to the minority class (in our case the `destroyed` class). For the tabular data baseline, and following similar approaches in Ecology, we include the featurized pixel data for each house (*i.e.* mean, standard deviation, and variance for each of the bands) and used an XGBoost model to classify each of our labels using 10-fold cross-validation for each of the years in our sample. Our best vision baseline achieved a 0.61 F1-score, while our best tabular baseline had an F1-score of 0.66.

[1]We use `V2` weights from PyTorch vision, which enhance the original paper weights using new optimizations during train and test time.

This house is {} years old. It is located {} meters above sea level with a slope of {}. Temperature is {} degrees. Relative humidity is {}. Wind speed is {}. The vapor pressure deficit is {} and the fuel moisture was {}. The risk to structure is {}. The fire name is {}

Figure 2: Template to transform tabular data to text prompt

Vision Encoder	Text Encoder	F-1 (All sample)	F-1 (1% sample)
ViT	-	0.71	0.67
-	GPT-2	0.65	0.62
-	RoBERTa	0.73	0.67
	Multimodal Models		
ViT	GPT-2	0.64	0.61
ViT	RoBERTa	0.77	0.75

Table 1: F1 Scores for all the unimodal and multimodal models. The last column captures the few-show abilities of each model using the 1% of our sample (n = 92).

3.3 Experiments

1. **Vision**: To test the leverage from the DE model, we first fine-tune a vision transformer *ViT* [7]:`vit-base-patch16-224-in21k` to our house dataset for a binary classification task. We do a grid search to pick the best learning rate, batch size and dropout combination during fine-tuning. Our best model had a a LR of 5×10^{-5}, and a dropout probability of 1×10^{-3} with a batch size of 64. As with the baselines, we test both upsampling the batches to have balanced sets and changing the weights of the CE loss function. We follow a '80-20-20' split policy for training, validation and testing sets.
2. **Text**: Following [17] best performing prompting strategies, we picked a template prompting, as seen in Figure 2. For our binary classification task we fine-tune two LLMs: GPT-2 [28] and RoBERTa [25]. Both models have a similar number of parameters (`gpt2-medium` and `roberta-large` have around 340M parameters), but RoBERTa is trained using significantly more data than GPT-2. For both models, we pass the suggested prompt and follow a similar grid search with a LR of 5×10^{-3}, with a batch size of 64.
3. **Multimodal**: Pre-trained versions of CLIP do not have not expressive text encoders [26]. To augment CLIP's text encoding-decoding abilities, we fine-tune the `VisionTextDualEncoder` class from HuggingFace [35] and change the text encoders to the same ones used in our text experiments. We always use the same ViT encoder (`vit-base-patch16-224-in21k`) and use the same prompt we described in Figure 2 and similar training parameters.

3.4 Multimodal Model Evaluation

To evaluate the classification abilities from our DE model after the DE fine-tuning we pass during test time a tuple: $\{(\mathcal{I}^{(i)}, p_t^{(i)}, p_f^{(i)})\}^{(n)}$ with an image: $\mathcal{I}^{(i)}$ and two text prompts with the same information, but with different label, a true label: $p_t^{(i)}$, and a false label: $p_f^{(i)}$ using the same template we used during training. Now, we will calculate the probabilities of matching image to text using a the softmax function and pick the image-prompt label with the highest probability: $\mathbb{P}(y^{(i)} = 1) = \arg\max \sigma(z^{(i)})$ where σ is the softmax function.

4 Results

As seen in Table 1, our results suggest that the DE model performs better than our two vision and tabular baselines (F1: 0.61 and 0.66, respectively). Following [3], we run our experiments using only 1% of our sample obtaining a comparable performance than with the full sample. This results align with experiments with TabLLM [17], LIFT [37], and CLIP [24] that have shown good few-shot performance in reduced data regimes. Compared to our baselines, all our models, including the vision and text only models, do perform better, with the exception of GPT-2.

5 Discussion

We have explored the use of multi-modal classification to solve a practical problem in fire management in fire-prone areas in the United States. We found that DE models are able to perform better than single-mode models (only vision or tabular data) and our baselines, giving a promising result to apply contrastive learning and CLIP-like models to environmental problems that involve multiple data modes and rely on small label samples. RoBERTa showed better performance overall compared to GPT-2, we still need to test larger or science-domain LLMs. Despite our results, is still needed to experiment the optimal fine-tuning strategies in DE models, not only to explore more computationally efficient strategies, such as LoRa [16], but also to exploit the adaptability of LLMs embeddings to the house burning task classification.

We have not explore the ability of these models to adapt to lower resolution imagery or its performance to do inferece with out-of distribution samples. These are still widely present problems in the remote sensing classification literature. Is importance to notice that each of our experiments were less computational demanding than the fire-spread model FARSITE [10], and it can serve as a test bed for land management interventions during wildfires in furture.

References

[1] J. T. Abatzoglou. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/joc.3413.

[2] P. W. Baylis and J. Boomhower. Mandated vs. Voluntary Adaptation to Natural Disasters: The Case of U.S. Wildfires, Dec. 2021.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020. arXiv:2002.05709 [cs, stat].

[4] E. Chuvieco, F. Mouillot, G. R. van der Werf, J. San Miguel, M. Tanase, N. Koutsias, M. García, M. Yebra, M. Padilla, I. Gitas, A. Heil, T. J. Hawbaker, and L. Giglio. Historical background and current developments for mapping burned area from satellite Earth observation. *Remote Sensing of Environment*, 225:45–64, May 2019.

[5] J. D. Cohen. Preventing disaster: Home ignitability in the wildland-urban interface. *Journal of Forestry 98(3): 15-21.*, 2000.

[6] T. Dinh, Y. Zeng, R. Zhang, Z. Lin, M. Gira, S. Rajput, J.-y. Sohn, D. Papailiopoulos, and K. Lee. LIFT: Language-Interfaced Fine-Tuning for Non-Language Machine Learning Tasks, Oct. 2022. arXiv:2206.06565 [cs].

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[8] A. Duane, M. Castellnou, and L. Brotons. Towards a comprehensive look at global drivers of novel extreme wildfire events. *Climatic Change*, 165(3):43, Apr. 2021.

[9] M. Finney. An Overview of FlamMap Fire Modeling Capabilities. 2006.

[10] M. A. Finney. FARSITE: Fire Area Simulator-model development and evaluation. *Res. Pap. RMRS-RP-4, Revised 2004. Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 47 p.*, 4, 1998.

[11] A. Forghani, B. Cechet, J. Radke, M. Finney, and B. Butler. Applying fire spread simulation over two study sites in California lessons learned and future plans. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, pages 3008–3013, July 2007. ISSN: 2153-7003.

[12] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[13] S. Ganapathi Subramanian and M. Crowley. Combining MCTS and A3C for Prediction of Spatially Spreading Processes in Forest Wildfire Settings. In E. Bagheri and J. C. Cheung, editors, *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 285–291, Cham, 2018. Springer International Publishing.

[14] S. Goyal, A. Kumar, S. Garg, Z. Kolter, and A. Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models, Dec. 2022.

[15] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? June 2022.

[16] X. He, C. Li, P. Zhang, J. Yang, and X. E. Wang. Parameter-efficient Model Adaptation for Vision Transformers, Dec. 2022. arXiv:2203.16329 [cs].

[17] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag. TabLLM: Few-shot Classification of Tabular Data with Large Language Models, Oct. 2022.

[18] W. M. Jolly, M. A. Cochrane, P. H. Freeborn, Z. A. Holden, T. J. Brown, G. J. Williamson, and D. M. J. S. Bowman. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6(1):7537, July 2015. Number: 1 Publisher: Nature Publishing Group.

[19] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, Apr. 2017.

[20] J. E. Keeley, H. Safford, C. Fotheringham, J. Franklin, and M. Moritz. The 2007 Southern California Wildfires: Lessons in Complexity. *Journal of Forestry*, 107(6):287–296, Sept. 2009.

[21] D. Khattar, J. S. Goud, M. Gupta, and V. Varma. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference*, WWW '19, pages 2915–2921, New York, NY, USA, May 2019. Association for Computing Machinery.

[22] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine. Supervised Multimodal Bitransformers for Classifying Images and Text, Nov. 2020. arXiv:1909.02950 [cs, stat].

[23] G. K. Kumar and K. Nandakumar. Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features, Oct. 2022. arXiv:2210.05916 [cs].

[24] H. Liu, S. Xu, J. Fu, Y. Liu, N. Xie, C.-C. Wang, B. Wang, and Y. Sun. CMA-CLIP: Cross-Modality Attention CLIP for Image-Text Classification, Dec. 2021.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. arXiv:1907.11692 [cs].

[26] R. Mokady, A. Hertz, and A. H. Bermano. ClipCap: CLIP Prefix for Image Captioning, Nov. 2021.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision, Feb. 2021.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[29] D. Radke, A. Hessler, and D. Ellsworth. FireCast: Leveraging Deep Learning to Predict Wildfire Spread. pages 4575–4581, 2019.

[30] S. T. Seydi and M. Sadegh. Improved burned area mapping using monotemporal Landsat-9 imagery and convolutional shift-transformer. *Measurement*, 216:112961, July 2023.

[31] W. C. Sleeman, R. Kapoor, and P. Ghosh. Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *ACM Computing Surveys*, 55(7):150:1–150:31, Dec. 2022.

[32] R. D. Stratton. Guidance on spatial wildland fire analysis: models, tools, and techniques. *Gen. Tech. Rep. RMRS-GTR-183. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 15 p.*, 183, 2006.

[33] A. D. Syphard, T. J. Brennan, J. E. Keeley, A. D. Syphard, T. J. Brennan, and J. E. Keeley. The role of defensible space for residential structure protection during wildfires. *International Journal of Wildland Fire*, 23(8):1165–1175, Oct. 2014. Publisher: CSIRO PUBLISHING.

[34] A. D. Syphard, J. E. Keeley, A. B. Massada, T. J. Brennan, and V. C. Radeloff. Housing Arrangement and Location Determine the Likelihood of Housing Loss Due to Wildfire. *PLOS ONE*, 7(3):e33954, Mar. 2012. Publisher: Public Library of Science.

[35] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.

[36] P. Xu, X. Zhu, and D. A. Clifton. Multimodal Learning with Transformers: A Survey, May 2023. arXiv:2206.06488 [cs].

[37] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. LiT: Zero-Shot Transfer with Locked-image text Tuning, June 2022. arXiv:2111.07991 [cs].

[38] H. B. Zia, I. Castro, and G. Tyson. Racist or Sexist Meme? Classifying Memes beyond Hateful. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219, Online, Aug. 2021. Association for Computational Linguistics.