MAdVerse: A Hierarchical Dataset of Multi-Lingual Ads from Diverse Sources and Categories

Amruth Sagar
CVIT, IIIT Hyderabad, India

Rishabh Srivastava
CVIT, IIIT Hyderabad, India

Rakshitha R T
KLE Tech, India

Venkata Kesav Venna
CVIT, IIIT Hyderabad, India

Ravi Kiran Sarvadevabhatla
CVIT, IIIT Hyderabad, India

{amruth.sagar@research.iiit, rishabh.s@students.iiit ,
01fe20bcs107@kletech, venkata.kesav@students.iiit, ravi.kiran@iiit }.ac.in

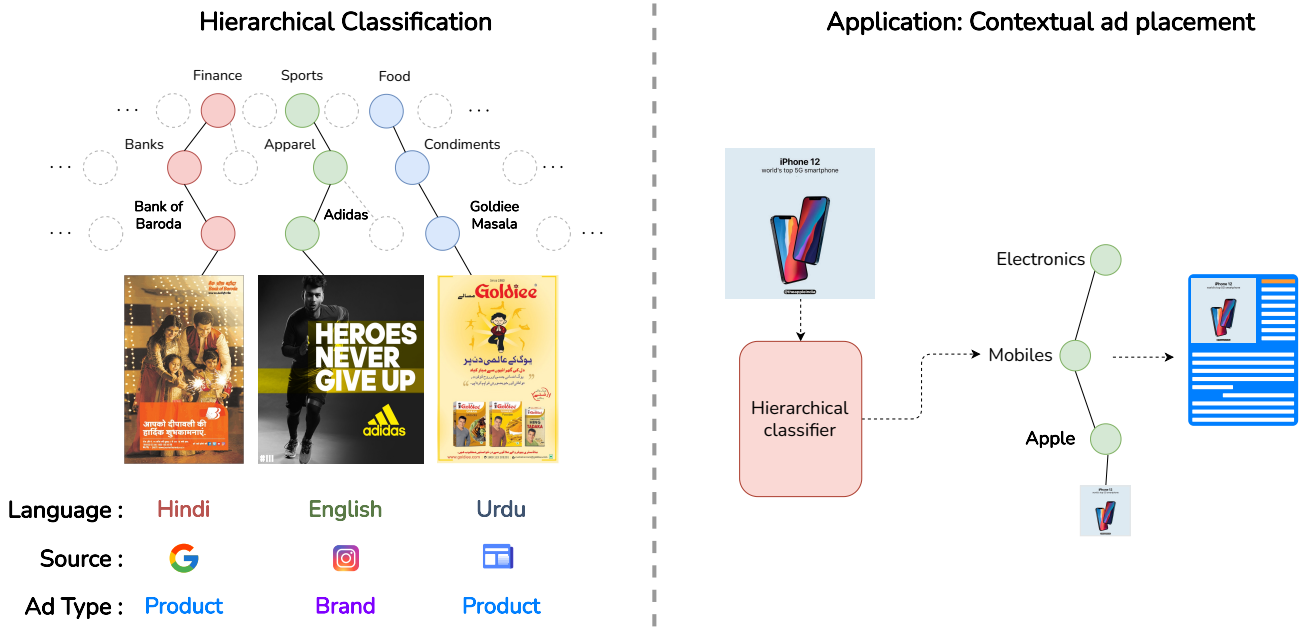

Figure 1. We introduce MAdVerse, an extensive multilingual dataset with over 50,000 ads. As shown above, our dataset can be used to design hierarchical classifiers which predict the ad category, language, ad source and type (left). As an application, a given ad could be placed alongside a media article (right) depending on the content of the article and *any/all* of the labels predicted by the hierarchical classifier.

Abstract

The convergence of computer vision and advertising has sparked substantial interest lately. Existing advertisement datasets are either subsets of existing datasets with specialized annotations or feature diverse annotations without a cohesive taxonomy among ad images. Notably, no datasets encompass diverse advertisement styles or semantic grouping at various levels of granularity. Our work addresses this gap by introducing MAdVerse, an extensive, multilingual compilation of more than 50,000 ads from the web, social media websites, and e-newspapers. Advertisements are hierarchically grouped with uniform granularity into 11 categories, divided into 51 sub-categories, and 524 fine-grained brands at leaf level, each featuring ads in various languages. We provide comprehensive baseline classifica-

tion results for prediction tasks within the realm of advertising analysis. These tasks include hierarchical ad classification, source classification, multilingual classification, and inducing hierarchy in existing ad datasets.

The dataset, code and models are available on the project page https://madverse24.github.io/

1. Introduction

In today's world, advertising and marketing spending have skyrocketed to record levels, reflecting their perceived effectiveness in boosting brand recognition and engaging consumers. On average, individuals are exposed to around 20,000 brands daily [13]. While traditional advertising channels such as television and print remain influential, the digital landscape has introduced innovative platforms such as social media, search engine marketing, and data-driven targeted advertising.

Ads are rich in visual and textual elements, including layout design, colors, logos, taglines, and product descriptions. This richness in information presents a challenge for understanding ads due to the intricate interplay between visual and textual components. Furthermore, advertisements for a single product or brand exhibit substantial variations influenced by factors such as advertising platform, design modifications, regional cultural nuances, and language.

To empower computer vision systems which excel in comprehending this diverse ad landscape, exposure to various sources and languages is essential. This exposure enables systems to recognize ad types across contexts, adapt to evolving advertising trends, and effectively process ads on diverse platforms. Surprisingly, existing works on advertising datasets [5, 11, 12, 16, 20] typically consist of singular source ads without diversity in sources, languages or label granularity.

In response to this gap, we introduce MAdVerse, a novel multi-source, multilingual and hierarchical dataset. MAdVerse captures the intricate relationships between products and brands, enriching our understanding of their attributes. The taxonomy includes hierarchical categories, ranging from broad product classifications to specific fine-grained brand names. MAdVerse consists of over 50,000 ads spanning diverse sources, including social media platforms, websites, and newspapers, providing insights into contemporary advertising trends. It also includes annotations related to both products and brands, facilitating a nuanced analysis of advertising strategies. MAdVerse spans 11 languages, offering insights into the linguistic and cultural aspects of advertising. In addition, we introduce and benchmark models for hierarchical ad classification, inducing hierarchy in other ad datasets and addressing auxiliary tasks such as ad source classification.

2. Related Work

2.1. Existing ad datasets

Pitt Ads [12] is a popular representative dataset and provides annotations related to ad topic, sentiment, ad strategy and symbols. However, the annotations do not include any notion of hierarchy. In general, most works typically operate within the confines of utilizing subsets of ad images sourced from Pitt Ads [12], which they then augment with task-specific annotations. Liang et al. [20] propose an ad dataset of 1000 images dubbed ADD1000 and explore saliency prediction on advertising images using eye movement data. More recently, Kumar et al. [16] use a subset of 3000 ad images from Pitt ads [12] and annotate them with persuasive strategies utilised in advertisements. A smaller subset of 250 images in their dataset also have segmentation masks for various ad regions.

Alternatively, some works employ ad images, not as their primary focus on advertisements but as supplementary material tools to address their specific research objectives. Fosco et al. [11] predict visual importance in graphic designs and include a subset of ad images from Pitt ads [12] since ads can be perceived as graphic designs. Cao et al. [5] generate image-conditioned advertisement layouts. This work also provides an ad poster layout dataset with saliency maps and layout annotations for all ad layouts.

A significant gap in this landscape is the absence of a comprehensive ad dataset that encompasses a wide array of advertisements from diverse sources, languages, and styles and also incorporates semantic grouping of brands or products.

2.2. Hierarchical image classification

Hierarchical image classification predicts classes for a given ad image at every level in the hierarchy. The initial step in hierarchical image classification is to establish the hierarchy structure. The nodes and the relationship between them can be represented in multiple ways (e.g. Directed Acyclic Graphs, trees where the presence of an edge between nodes indicates a type of relationship between them). Deng et al. [9] introduce the concept of Hierarchy and Exclusion (HEX) graphs, a graph formalism capturing different types of semantic relations between labels.

Works in hierarchical classification adopt different approaches based on the manner in which hierarchy is integrated. Some works employ a hierarchical architecture [7, 19, 24, 26, 27]. Some others, such as semantic embedding [3], HXE (Hierarchical cross entropy), and soft labels [4], incorporate the hierarchy into loss functions via tree-based metrics. Instead of using a single label structure as done in other works, MMF [19] integrates different label structures to obtain a diverse set of prior information about the categories and fuses them to achieve better hier-

Dataset	# images	Dataset source	Hierarchy present	Taxonomical levels	Languages
Pitt Ads [12]	64,832	Google images	✗	-	-
Persuasion strategies [16]	3000	Subset of Pitt Ads	✗	-	-
Fixation Prediction [20]	1000	Subset of Pitt Ads	✓	2	-
Visual Importance [11]	1000	Subset of Pitt Ads	✗	-	-
MAdVerse	**52443**	**Google, Instagram & Facebook, Newspapers**	✓	**5**	**11**

Table 1. Comparitive overview of related datasets

archical classification. Chen et al. [7] do hierarchical multi-granularity classification, in which images can be classified to a particular level of granularity in the hierarchy. Wang et al. [25] utilize cross-attention between hierarchical class word embeddings and image embedding, followed by pooling of output embeddings for score-based level and category prediction.

In the context of our ad dataset, we explore a combination of representative backbones, hierarchical/non-hierarchical architectures, and hierarchical /non-hierarchical losses. Incidentally, we are the first to benchmark hierarchical approaches on an ad dataset.

3. Dataset

3.1. Collecting ad images

We created a diverse and high-quality dataset of image advertisements by scraping from multiple sources – Google images, social media platforms (Facebook and Instagram), dedicated websites (e.g. Advert Gallery [1]) and digital newspapers in various languages. Prior to scraping images from Google images and social media, we compiled a comprehensive list of popular brands to ensure representation across various advertisement topics such as food, clothing, sports, vehicles, etc. Our objective was to collect images for each brand from these sources.

For Google images, we used a fixed prompt, "Advertisement images of ", and prepended it to the brand name to query and retrieve the top 200 results associated with the given query. Images obtained through this process usually contain a variety of images that are not relevant to ads of the brand in question. To address this, we established specific criteria to distinguish genuine ads from non-advertisement content. A given image is an ad if it satisfies the following conditions:

- *Presence of Product and Graphics*: An ad should have a product and supplementary graphics that emphasize the product's features.
- *Brand Logo Inclusion*: The image must feature the brand's logo to be classified as an ad.
- *Tagline or Textual Explanation*: An ad should contain a tagline or some form of text that conveys information about the advertised product.

We manually annotated a subset of images according to the ruleset and used them to train an Ad and non-ad classifier incorporating a Vision Transformer (ViT) [10] and a classification head, which demonstrated accuracy of 93.77% and F1 score of 0.898 on the test set with train, test, val split distribution of 70%, 20%, 10%. The trained classifier was then applied to the unannotated set of images, extracting ad images from the dataset. Any misclassifications were manually corrected.

To ensure dataset quality and uniqueness, we aim to avoid duplicate images. For removing duplicates, we employed a CNN-based method from the imagededup library [14] to identify and remove exact duplicates and similar images above a similarity threshold of 0.95.

As part of our strategy to increase ad diversity, we scraped advertisements from popular social media platforms such as Instagram and Facebook using the Apify [2] scraping tool. The URLs obtained via the tool were used to scrape images, and we manually selected ads from these images. Additionally, we gathered ad images from Advertgallery [1], an image repository of ads from digital newspapers and websites of e-newspapers in various languages.

To further expand the dataset, we sourced advertisements from various digital newspapers, each originating from diverse languages and regions, obtained from careerswave [6] and dailyepaper [8]. Our data collection efforts encompassed the past six months of available content within each newspaper, and the collection of newspapers was conducted around July 2023. Consequently, the ads from newspapers included in our dataset span the time frame from January to July 2023. In total, we amassed a substantial dataset comprising approximately 12,000 newspapers, which collectively contain around 150,000 individual pages.

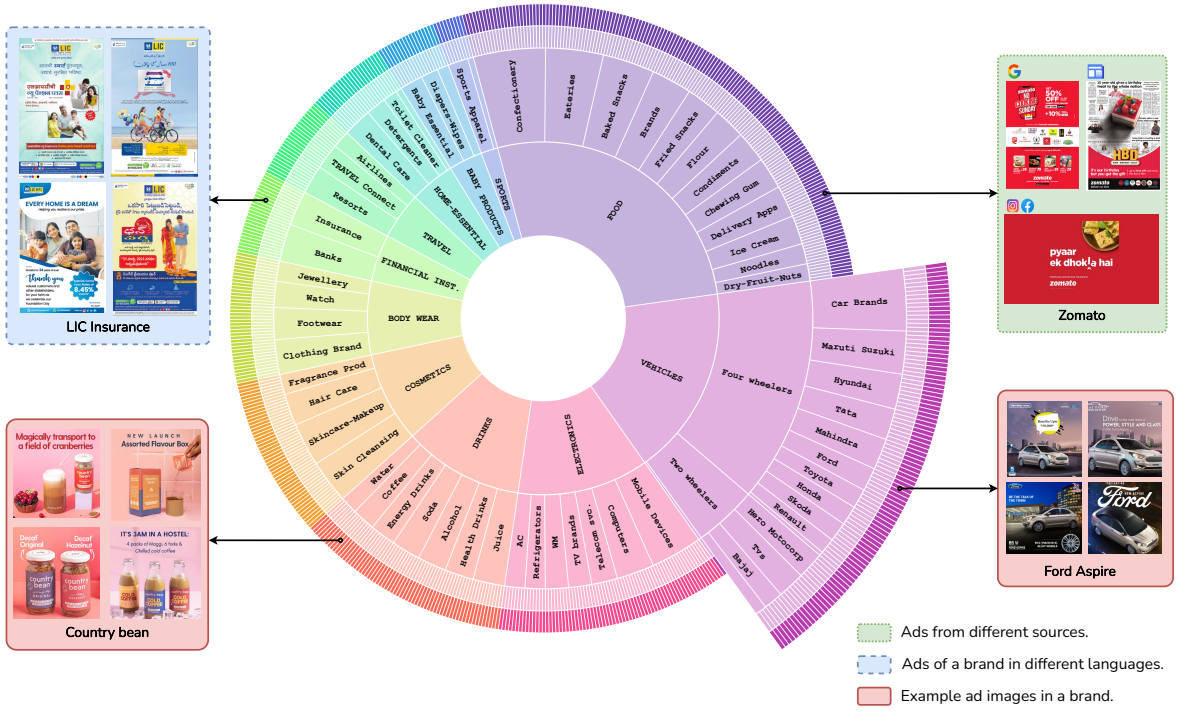

Figure 2. Hierarchical Taxonomy of the Dataset.

The process of isolating advertisements from the collected newspapers involved a two-stage approach. Specifically, 2,863 newspaper pages were manually annotated following the format defined by MS-COCO [21]. Subsequently, a Faster R-CNN [23] detector, which had been pretrained on NewsNavigator [17] data, was finetuned on the manually annotated subset. This fine-tuned model was used to extract candidate images for ads from each page of the newspapers. The candidate images were classified as ad or not ad using the classifier. To ensure the utmost accuracy of our dataset, a manual validation procedure was employed to eliminate any false-positive images. More details can be found in the supplementary material.

Finally, to address any remaining duplicates in the final dataset, we conducted an additional de-duplication step similar to the initial one. The final dataset is comprised of **52443** images. Of these, **21465** images came from WebAds, and **1,947** images were from Advert Gallery, and the remaining **29031** images were from the newspapers. The dataset images contain hierarchical annotation of categories and sub-categories, reflecting meaningful relationships between the images and their respective brands.

The hierarchy, characterized by varying heights of leaf nodes, aligns with real-world complexity and offers a comprehensive and well-organized dataset, reflecting the natural structure of the data. The statistics of the hierarchy are presented in Table 1. A pictorial illustration of the dataset can be viewed in Fig. 2.

3.2. Dataset annotations

Our dataset has a systematic organization of products and their associated brands into a structured taxonomy that captures their individual attributes and the relationships that exist between them. By structuring the data in this manner, we can provide a more comprehensive understanding of advertisements.

3.2.1 Hierarchical Annotations

The taxonomy comprises 5 levels that progressively detail the categorization of ad images. Each level in the hierarchy has the same level of abstraction. The structure of our hierarchy is shown in Fig. 2. The highest level has product categories such as “Food”, “Electronics”, and “Vehicles,” which get subdivided into fine-grained brand examples such as “Ice cream”, “Washing Machine”, and “Ford” as we go

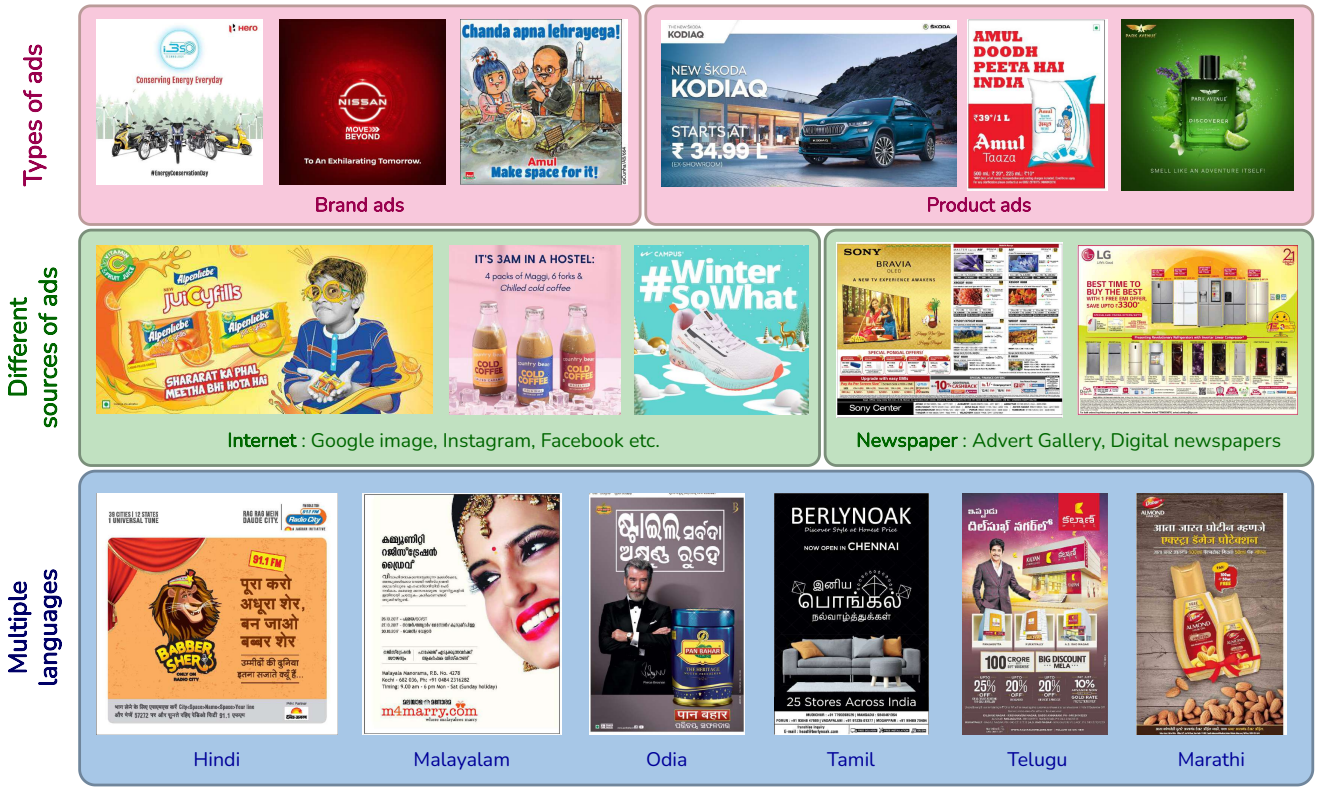

Figure 3. An illustration of the diverse annotations in the MAdVerse dataset.

down the hierarchy. The taxonomy is designed such that the leaf nodes are ad images, while the non-leaf nodes are clustered into a hierarchy on the basis of their product features.

For images from the internet, the images are mainly retrieved through keyword query-based techniques. The keywords used for extraction were ordered in a hierarchical structure, so the data downloaded is ordered in the provided hierarchy. More details on the keywords used are present in the supplementary material.

3.2.2 Ad Source Annotation

Our dataset has a wide array of ad images collected from various online sources, each contributing to the diversity and comprehensiveness of our dataset. These sources offer a range of ad formats, including banners, carousel ads, and sponsored posts (see Fig. 3). In particular, the sources consist of popular social media platforms such as Facebook, Instagram, and Google Images, as well as other websites.

Additionally, ad images sourced from e-newspapers further expand the dataset's scope by incorporating images from a publically available website, the advert gallery, and those appearing in publically available newspapers. Ads from e-newspapers stand out due to the presence of more text compared to digital ads featuring more extensive written content. Having multiple sources for ads can help investigate differences in various advertising formats and analyze cross-platform strategies.

In order to categorize the images in our dataset by their source, we provide source annotations. Images obtained from the internet or social media sources are labelled as 'internet,' while those originating from newspapers are labelled with the source 'newspaper.' This categorization helps distinguish between images collected from diverse sources, including popular social media platforms such as Facebook, Instagram, and Google Images, as well as e-newspapers, each contributing to the comprehensive nature of our dataset.

3.2.3 Product vs Brand Annotations

Our dataset contains two types of images: one for products and another for brands. Each type has its own role in showing the different aspects of advertising content. Product images constitute a category that revolves around highlighting specific offerings or services. These images centre on showcasing the distinct features, benefits, and selling points of individual products. In contrast, brand images occupy a different sphere within the dataset. These images are crafted to enhance a company's overall brand image and identity,

transcending the promotion of individual products.

By distinguishing between product and brand annotations, our dataset provides a representation of advertising content strategies (see Fig. 3). This separation allows us to analyze how brands convey their messages differently when focusing on specific products versus building a cohesive brand identity. This distinction allows us to explore the varying impacts these approaches may have on consumer perception and engagement.

3.2.4 Multilingual Annotations

Including ads in different languages is crucial because it helps us understand how advertising messages are adjusted to suit various languages and cultural settings. Multilingual annotations enable researchers to study how advertisers use different languages to connect with people. They can look at how words and meanings change and how ads consider different cultures to reach people who speak different languages. This multilingual perspective is crucial for uncovering region-specific dynamics that underlie effective advertising campaigns and consumer behaviour.

Our dataset is a testament to the linguistic diversity inherent in modern advertising campaigns, featuring ads in 11 distinct languages: Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Tamil, Telugu, and Urdu. The fact that our dataset includes multiple languages adds a special aspect that matches the multilingual nature of advertising and accommodates a wider range of research questions.

We provide multilingual annotations for all images, with an ad being one of the 11 languages. The majority of ads are in English, Hindi, and Marathi. The distribution of the dataset can be seen in Fig. 4.

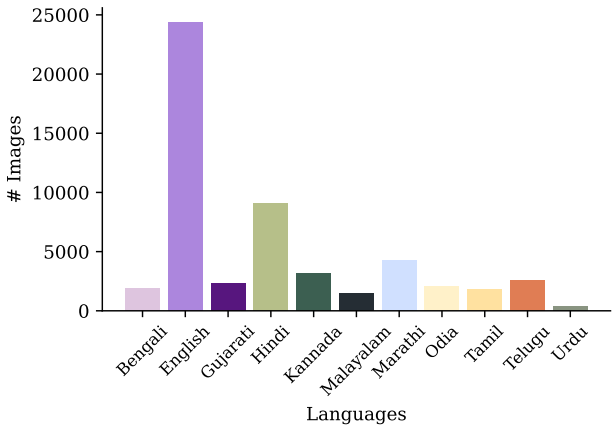

Figure 4. Distribution across multiple languages in the dataset.

4. Experiments

We have discussed the dataset, the data collection process, its multiple sources and annotations and the different types of ad images. We present baselines for hierarchical classification, inducing hierarchy in other ad datasets and other tasks.

4.1. Hierarchical classification

Hierarchical classification predicts classes for images at various levels of granularity in an underlying structured hierarchy. The structure of a hierarchy depends on the relationship between the classes in the dataset. In our case, nodes present in each level are mutually exclusive classes, and our hierarchy is a tree, and every node has only one parent. To conduct a comprehensive evaluation, we consider various combinations of backbones, classifier configurations, loss functions and metrics - see Fig. 5.

Backbones: We consider ViT-L [10], ConvNeXT large [22] and BLIP-2 [18] as representative backbones. While the ViT model has shown great results in tasks such as object detection and classification tasks, a CNN-based backbone such as ConvNeXT [22] is robust to noise in input and is deemed efficient compared to transformer-based vision backbones. BLIP-2 is a backbone which is aware of both the visual and the textual modality. The latter is crucial since ads consist of pictures and text.

Classification Architectures: We consider two classifier architectures to perform hierarchical classification. One approach is to directly predict the fine-grained brands (i.e. leaves in the hierarchy) and then obtain parent predictions by traversing from the predicted leaf to the topmost level. Another approach is to have a classification branch for each level of the hierarchy to obtain class predictions, as shown in Fig 5.

Loss functions: The set of losses used for each approach is given in Fig 5. Irrespective of the approach, hierarchy-agnostic losses do not incorporate the knowledge of the hierarchy. They are simple cross-entropy-based losses, whereas hierarchy-aware losses use tree-based metrics as a tool to include hierarchy in loss computation. We also consider the label embedding methods such as Soft labels [4] and Semantic embedding [3], since they add hierarchical information to the target labels, indirectly inducing the hierarchy in loss calculation.

Metrics: We use two types of metrics. Hierarchy agnostic metrics are regular metrics such as Accuracy, Precision, Recall and F1 score. They do not use any information about the underlying hierarchy. The other type of metrics are hierarchy aware and typically utilize the tree structure of the dataset to compute associated performance measures.

Refer to supplementary material for a detailed description of the losses and the metrics.

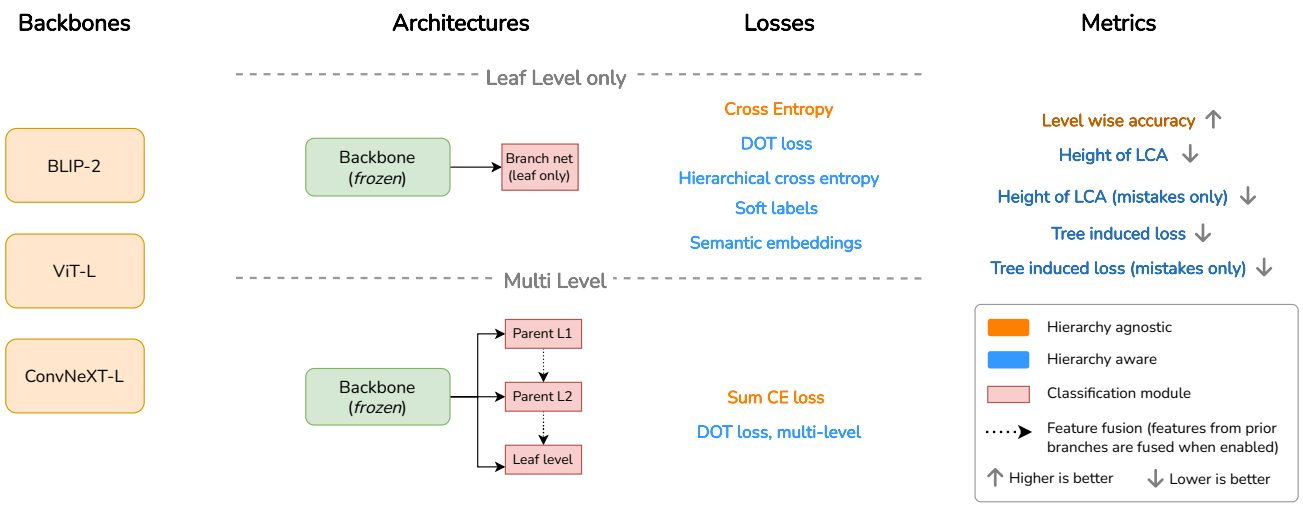

Figure 5. All combinations used in the experiment.

4.2. Evaluation setup

We create 70, 10, and 20 percentage splits for train, validation and test sets using 21k ads taken from the internet, which predominantly contain ads in English. From this point forward, we will refer to this subset of ad images as 'English ads'. All the configurations of backbones, architectures and losses are trained for 40 epochs, with a batch size of 32, 1e-3 learning rate with Adam [15] optimizer.

4.3. Results

The top two performers from each approach are given in the table 2. As expected, fine-grained classification is generally harder, given the number of classes and per-class image distribution. Since BLIP-2 is trained on various vision-language tasks, it exhibits the best performance across the board, demonstrating the advantage of a language-and-text-based backbone choice. Compared to hierarchy-agnostic cross-entropy losses, hierarchy-aware label embedding methods such as Semantic embedding [3] and Soft Labels [4] enable superior performance in predicting the classes accurately across the tree levels.

Performance scores for the remaining configurations are available in the supplementary material.

4.4. Inducing hierarchy in other ad datasets

To assess the generalization of our hierarchical classifiers and to demonstrate the efficacy of providing hierarchical annotations to ads beyond our dataset, we conducted experiments by inducing hierarchy onto a sampled set of images from Pitt ads [12], as it does not provide a test set explicitly, we selected a sample of more than 200 images from their dataset. We then removed all such images from the test set already in our dataset. In the end, we had 234 images.

The images were then manually annotated till the subcategories level, aligning them to the same hierarchy structure of our dataset. We chose the BLIP-2 as the backbone with multilevel architecture with feature fusion, trained with the level-wise sum of cross-entropy loss on the English ads.

Table 3 shows the level-wise accuracy of the model used for hierarchical placement. The results show that the classifier trained on our dataset shows reasonable cross-dataset generalization, which is particularly relevant for researchers seeking to leverage hierarchical annotations in diverse domains.

4.5. Inducing hierarchy on multilingual ad images

Similar to the previous task, we wanted to assess the ability of our hierarchical classifiers trained on English ads to classify multilingual ads. We selected a sample of 300 images from the multilingual ads in our dataset. We repeated the same step of removing images from the sampled set that might be present in the English ads. Finally, we had 291 images. Similar to the previous task, we chose BLIP-2 as the backbone with multilevel architecture with feature fusion, trained with Sum_CE loss, since it has the best level-wise accuracy out of all the configurations tried. Table 3 shows the level-wise accuracy of the model on multilingual images. The classifier performs very well in predicting the first level of the hierarchy, but the prediction accuracy drops slightly. The main reason behind this is that the images used for inference contain text in multiple languages, and most of these images are from newspapers. Newspapers tend to have more text content in various languages, which makes it a bit harder for the classifier to classify these images accurately.

Architecture		Backbone	Loss	Accuracy (%)			Height of LCA			Height of LCA (mistakes only)			TIE			TIE (mistakes only)		
				L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Leaf level		BLIP-2	Semantic embedding	98.17	94.56	52.54	0.02	0.07	0.55	1.00	1.34	1.15	0.04	0.15	1.09	2.00	2.67	2.31
		BLIP-2	Soft-labels	97.46	93.74	85.25	0.03	0.09	0.24	1.00	1.41	1.60	0.05	0.18	0.47	2.00	2.81	3.19
Multi level	with FF	BLIP-2	Sum CE	95.16	89.25	81.73	0.05	0.15	0.29	1.00	1.41	1.60	0.10	0.30	0.58	2.00	2.82	3.20
	without FF	BLIP-2	Sum CE	95.51	89.34	80.87	0.04	0.15	0.31	1.00	1.41	1.61	0.09	0.30	0.62	2.00	2.83	3.23

Table 2. A comparison between leaf-only and multilevel approaches, with a focus on the top two performers

Metrics	Hierarchy Agnostic								Hierarchy Aware							
	Accuracy (%)		Precision		Recall		F1 Score		Height of LCA		Height of LCA (mistakes only)		TIE		TIE (mistakes only)	
	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2
Pitt's Ad Dataset	86.67	72.15	0.74	0.49	0.73	0.47	0.73	0.45	0.12	0.41	1	1.46	0.24	0.81	2.93	5.63
Multilingual Ads	96.91	87.29	0.71	0.61	0.70	0.62	0.71	0.60	0.03	1.62	1	1.27	0.06	0.32	2.00	2.54

Table 3. Results of inducing hierarchy on a subset of Pitt's Ad Dataset and our multilingual dataset.

4.6. Ad source classification

This task deals with the classification of ad sources such as social media, newspapers, and Google databases. Leveraging our dataset's comprehensive ad image annotations from online platforms such as Facebook, Instagram, Google images, and newspapers, we chose two separate types of grouping of data for our classification tasks. For the first task, we grouped the images of the ads into newspaper ads and web ads, giving a total of 30978 newspaper ads and 21465 web ads. For the second task, we grouped the images into three classes: Google ads, social media ads (Instagram/Facebook), and newspaper ads. For this task, we have 12260 images from Google Images, 9205 images from social media websites, mainly Facebook and Instagram, and 30,978 newspaper images combining both the advert gallery [1] and the e-papers. For both classification tasks, we split the data into training, validation, and testing sets, allocating 70%, 20%, and 10%, respectively.

Our chosen architecture consisted of a straightforward, non-frozen backbone with a classification head. We had a separate learning rate for each backbone and classification head in order to benefit the training process. For the backbones, ViT-large [10] and ConvNeXT [22] were chosen. Table 4 shows the results of models on ad source classification.

The results of the experiment clearly show that our classifier works well in both situations. We noticed that telling apart ads from Instagram and Google Images is harder than distinguishing between web and newspaper ads.

Classification	Metrics	Architecture	
		VIT	Convnext
Two-class classification (Web Ads vs Newspaper Ads)	Accuracy (%)	97.90	96.96
	Precision	0.98	0.96
	Recall	0.97	0.94
	F1 score	0.97	0.95
Three-class classification (Google Ads vs Social Media Ads vs Newspaper Ads)	Accuracy (%)	90.07	90.70
	Precision	0.85	0.87
	Recall	0.84	0.87
	F1 score	0.85	0.87

Table 4. Results of ad source classification

5. Conclusion

There has been a lack of datasets encompassing diverse advertisement styles and semantic grouping at various levels of granularity. To bridge this gap, we have introduced MAdVerse, a vast and multilingual collection of over 50,000 ads sourced from the web, social media platforms, and e-newspapers. These advertisements are systematically organized as a coarse-to-fine hierarchy. Importantly, this dataset spans multiple languages, providing a comprehensive view of advertising across linguistic and cultural boundaries. Additionally, we have presented baseline classification results for various crucial prediction tasks in the field of advertising analysis, including hierarchical ad classification, source classification, multilingual classification, and inducing hierarchy in existing ad datasets. With MAdVerse and our classification results, we aim to facilitate further research and innovation in the realm of advertising analysis, offering a valuable resource for researchers and practitioners to advance our understanding of advertising and its impact on consumer behaviour.

References

[1] Advert gallery. https://www.advertgallery.com/. 3, 8

[2] Apify. https://apify.com/. 3

[3] B. Barz and J. Denzler. Hierarchy-based image embeddings for semantic image retrieval. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 638–647, Los Alamitos, CA, USA, jan 2019. IEEE Computer Society. 2, 6, 7

[4] L. Bertinetto, R. Mueller, K. Tertikas, S. Samangooei, and N. A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12503–12512, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. 2, 6, 7

[5] Yunning Cao, Ye Ma, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. Geometry aligned variational transformer for image-conditioned layout generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1561–1571, New York, NY, USA, 2022. Association for Computing Machinery. 2

[6] careerswave. https://www.careerswave.in/newspaper-download-pdf-free/. 3

[7] Jingzhou Chen, Peng Wang, Jian Liu, and Yuntao Qian. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4848–4857, 2022. 2, 3

[8] dailyepaper. https://www.dailyepaper.in/news-home/. 3

[9] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 48–64, Cham, 2014. Springer International Publishing. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3, 6, 8

[11] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. Predicting visual importance across graphic design types. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 249–260, New York, NY, USA, 2020. Association for Computing Machinery. 2, 3

[12] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110, 2017. 2, 3, 7

[13] Investment in advertising. https://www.theworldcounts.com/economies/global/spending-on-advertising/. 2

[14] Tanuj Jain, Christopher Lennan, Zubin John, and Dat Tran. Imagededup. https://github.com/idealo/imagededup, 2019. 3

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 7

[16] Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. Persuasion strategies in advertisements. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):57–66, Jun. 2023. 2, 3

[17] Benjamin Charles Germain Lee, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld. The newspaper navigator dataset: Extracting headlines and visual content from 16 million historic newspaper pages in chronicling america. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3055–3062, New York, NY, USA, 2020. Association for Computing Machinery. 4

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 6

[19] Xiaoni Li, Yucan Zhou, Yu Zhou, and Weiping Wang. Mmf: Multi-task multi-structure fusion for hierarchical image classification. In Igor Farkaš, Paolo Masulli, Sebastian Otte, and Stefan Wermter, editors, *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 61–73, Cham, 2021. Springer International Publishing. 2

[20] Song Liang, Ruihang Liu, and Jiansheng Qian. Fixation prediction for advertising images: Dataset and benchmark. *Journal of Visual Communication and Image Representation*, 81:103356, 2021. 2, 3

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 4

[22] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society. 6, 8

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 4

[24] Yian Seo and Kyung shik Shin. Hierarchical convolutional neural networks for fashion image classification. *Expert Systems with Applications*, 116:328–339, 2019. 2

[25] Peng Wang, Jingzhou Chen, and Yuntao Qian. Semantic guided level-category hybrid prediction network for hierarchical image classification. *International Journal of Wavelets, Multiresolution and Information Processing*, 0(0):2350023, 0. 3

[26] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: Hierarchical deep convolutional neural networks for large scale visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2740–2748, 2015. 2

[27] Xinqi Zhu and Michael Bain. B-CNN: branch convolutional neural network for hierarchical classification. *CoRR*, abs/1709.09890, 2017. 2