INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

MULTI-CHANNEL CONVOLUTIONAL NEURAL NETWORK FOR PRECISE MEME CLASSIFICATION

[1]S.Prathap, [2]P.Yeshika, [3]K.Thapaswi, [4]S.Gnanaharshini, [5]E.varshitha

[1]Assistant Professor, Department of CSE(DS), Malla Reddy Engineering College for Women (Autonomous Institution – UGC, Govt. of India), Hyderabad, INDIA.

[2345]UG,Department of CSE(DS), Malla Reddy Engineering College for Women (Autonomous Institution – UGC, Govt. of India), Hyderabad , INDIA.

ABSTRACT

This paper proposes a multi-channel convolutional neural network (MC-CNN) for classifying memes and non-memes. Our architecture is trained and validated on a challenging dataset that includes non-meme formats with textual attributes, which are also circulated online but rarely accounted for in meme classification tasks. Alongside a transfer learning base, two additional channels capture low-level and fundamental features of memes that make them unique from other images with text. We contribute an approach which outperforms previous meme classifiers specifically in live data evaluation, and one that is better able to generalise 'in the wild'. Our research aims to improve accurate collation of meme content to support continued research in meme content analysis, and meme-related sub-tasks such as harmful content detection

1 INTRODUCTION

Internet memes are multi-modal content commonly shared online which reference cultural materials, catchphrases, jokes or images to communicate ideas. They exploit external knowledge in combination with text and image modalities to convey meaning; their unique properties make them easily editable or shareable, but difficult to collect or analyse using automated methods. Internet memes are the focus of ongoing research due to how quickly they circulate online and the complexity of detecting harmful, offensive, hateful or toxic messages in multi-modal content [5, 25, 32]. As meaning is generated through interactions in both modalities, humour and reference to external knowledge, meaning is no longer face value and difficult to decode without context. Varied and comprehensive meme datasets are therefore crucial for such

ISSN: 2057-5688

automated content detection tasks. Currently, available meme datasets are created through extensive manual annotation of collected content to determine whether an image is or is not a meme [29]. In some cases, artificially generated datasets are created for hate-speech detection using the popular superimposed text-over-image meme (image macro) format, which do not represent typical memes shared online that are more varied and contain noisy text [14, 15]. Alternative strategies include collating content from Twitter with the hashtag 'meme', though this approach assumes tagging and categorisation accurately represents that all content is a meme. Additionally, these datasets are 'static' and manual annotation requires updating – in the peculiar case of memes which rapidly evolve and develop new formats, static datasets do not capture emerging memes and may quickly become outdated. Datasets typically distinguish memes from images such as photographs and do not include other image-with-text (IWT) formats like advertisements, movie posters, online news articles or screenshots of posts which are also circulated online. Thus, models trained on such data perform poorly on live detection tasks, or the subsequent analysis of meme features are not accurate representations of only memes and real memes.

2 RELATED WORK

Meme definitions. We adopt the definition of memes from Shifman's Memes in the Digital World, which describes memes as, "a group of digital items sharing common characteristics of content, form, and stance; that [are] created in awareness of one another; that are circulated, imitated, and/or transformed via the internet by users." [34]. As user-generated content, memes are considered products of 'participation' by multiple users, which distinguishes them from other images and IWT combinations; as Shifman [33] further notes, they are "marked as textually incomplete or flawed, thus distinct from and perhaps defiant of glossy corporate content." In addition, Knobel and Lankshear [16] note that common features of memes are intentionally used to encourage participation via resharing and editing, thus meme formats evolve over time from continuous user participation. These attributes make memes distinct from other media content, such as viral videos – which are commonly reshared, but are not edited by participating users. Rogers and Giorgi [28] similarly argues that

ISSN: 2057-5688

memes are collections of technical content by analysing memes created using image generators.

There is significant work identifying harmful or offensive memes as part of ongoing research in the detection and prevention of toxic/hateful online content. We therefore split meme classification into two categories: classification tasks concerned with identifying subsets of meme content (e.g., harmful vs non-harmful, propaganda vs truth) and classification concerned with distinguishing memes from non-meme content.

Hate speech detection. Afridi et al. [1] conducted a comprehensive study of multi-modal meme classification approaches, covering both meme vs non-meme classification and other classification tasks. Their survey noted that state-of-the-art multi-modal transformers perform poorly in meme related tasks; the authors suggest that, in standard vision and language tasks like image captioning, efforts are made to generate the best explanation for an image, but there is little semantic alignment between image and text in memes. Sharma et al. [2] conducted an extensive survey of harmful meme classification and available datasets, noting that the majority of state-of-the-art harmful content classification approaches use similarly large-scale pre-trained neural networks for visual and text content. However, the authors also outlined the complexity of the task and challenges including subjective label annotation, insufficient dataset size and rapid evolution of memes. Whilst our research does not address harmful content, it does aim to improve the availability of meme datasets, reduce annotation burden for meme detection and maintain better accuracy in live evaluation.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

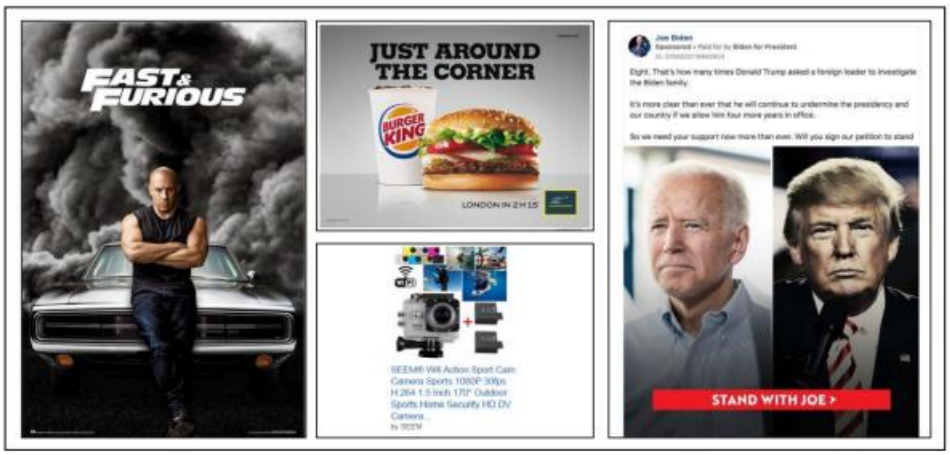

Figure 1: Example non-memes IWT formats and typical meme formats.

3 DEFINING MEME FEATURES

As indicated in Figure 1, there are numerous types of online content that share meme features but are not memes under Shifman's definition [34]. Non-meme types outlined here are not created, edited or transformed by internet users. They do not belong to a specific group of content (e.g., a subset of memes, such as image macros). Importantly, some of the content is designed to advertise or persuade; they are not opinions of users who created them, but rather the stance of brands intended to sell a product or idea. As noted by Kirk et al. [15] and in prior research, model performance is less accurate outside of competition or training scenarios due to the variety of IWT formats in online spaces compared to training data. The difficulty in collecting memes relates to their boundaries which are blurred, as memes take materials from existing artefacts and mimic them in form, structure, style, language and design - but repurposed to communicate a different message. The re-use of some images or catchphrases can be manipulated in ways that, in the context of a meme, carry an entirely different meaning to their origin. When performing deeper analysis of linguistic and visual meme features on datasets that incorrectly contain IWT formats, the subsequent analysis is likely to be a less accurate representation of memes circulated online. Whilst reusing cultural materials and effectively 'mimicking' other online content makes memes harder to detect, it is also this re-purposing and deliberate design to be re-shared and re-edited by other users that provide subtle visual and textual markers used in our architecture to detect memes.

ISSN: 2057-5688

A baseline model was trained on memes from the Memotion competition [29], memes collected from Reddit.com [3] and non-memes from the Flickr8k data-set [11], with 8,000 images each for memes and non-memes. We trained a convolutional neural network (CNN) on this dataset as a baseline model to compare the impact of excluding IWT images on classifier accuracy. An additional data-set was created to include IWT non-meme formats commonly circulated online; as expected, the baseline classifier achieved significantly poorer performance on this dataset (see Section 5 - Results). The analyses presented in this paper are the results of architectures trained this extended data of of memes and non-memes with the inclusion of IWT formats.

4 MULTI-CHANNEL APPROACH

Visual salience analysis of the baseline classifier indicated this model focused on the presence of text in images to classify memes, which would be unlikely to distinguish memes from IWT nonmemes. We therefore explored individual text and image channels in more detail to understand which features were important in each modality and could be used alongside the visual features extractor from an image-only CNN. Given their usual format as one incorporating text, a model based on text-only features examines what textual attributes were unique to memes in comparison to non-meme IWT formats. We also tested variations of histogram channels, including local binary patterns (LBP), histogram of orientated gradients (HOG) and Haar wavelet transformations, which have previously seen promising results [24, 30]. However, these were not used in the final architecture proposed as our histogram variation outperformed these approaches. Other potential channels were explored, including template matching, however this was deemed less useful as meme formats change over time. Object detection and facial recognition/detection were also ruled out, as in the case of movie posters and adverts individuals who appear in memes may also appear in non-meme IWT formats as memes tend to re-appropriate available icons and material.

ISSN: 2057-5688

5 RESULTS

Combinations of individual and dual channels were tested as well as the proposed MC-CNN. The baseline model is also provided for comparison, which is a single-channel CNN without VGG16 as a backbone and trained on a dataset of memes and non-memes without IWT samples Although Sharma and Pulabaigari [30] present a three-channel classifier, the authors superimpose the captions of non-text images to artificially convert all images to IWT formats, whereas captions are not available as a feature in the datasets used in our study. Semantic similarity is calculated from either superimposed caption text or extracted meme text, against a generated scene descriptor of the image; in the case of Flickr8k images, the captions of Flickr8K used for superimposition are descriptions of the image and thus will have a high semantic similarity to a scene descriptor. The authors also train their classifier on a much smaller dataset and is unavailable to re-train with our dataset for a fair comparison. We suggest the MC-CNN is better able to generalise than comparison models, and that multimodal approaches based only off input text and image is not sufficient to accurately predicted challenging dataset like images circulated on Twitter. Du et al. [8] focused primarily on detecting memes with text from non-memes IWT with image and text input only; in the case of live detection, memes and non-memes can contain both modalities or only one, and in live evaluation text data is likely to be noisier. The additional histogram channel of our classifier performs the same function regardless of whether both modalities exist, and is better able to identify instances of poor image alteration innate to many memes. For the MC-CNN, non-memes incorrectly classified as memes tended to be examples of user-generated content (e.g., a digital drawing, screenshot of other viral content, a user generated advert or design) but not necessarily considered a meme. Given the original training data did not contain user-generated images and only corporate content, lower performance on this type of data is expected. There is some difficulty defining memes themselves. For example, the practice of screenshotting and re-sharing humourous content is popular on Twitter, though not necessarily following the principles of altering or editing to make content a meme; however, such content shares features of memes in their format and

linguistic attributes. In the Twitter evaluation dataset, these images were not considered a meme

6 CONCLUSION

In this paper we proposed a multi-channel convolutional neural network for meme and non-meme classification, which outperforms models trained on a similar dataset of IWT non-memes in live data evaluation. The individual channels that comprise the MC-CNN were developed from analysis of meme text and colour features in relation to IWT non-memes. Whilst we propose an image channel with transfer learning as other models have, the two additional channels exploit different features of memes than previous studies, focusing on the visual and textual markers that make such content 'textually incomplete and flawed' compared to other IWT content [33]. Our architecture retains better performance in live evaluation tests, a crucial step for classifier in meme-related tasks which often perform poorly outside of training [15]. The boundaries between memes and other content is not always clear; memes mimic and reuse cultural materials from other images, and their formats continually evolve through participation. The architecture presented is better able to generalise those varied formats by focusing on the markers of user-edited content rather than image or object detection. A classifier that can accurately collate more memes would improve tasks relating hate speech detection, harmful content or propaganda detection by increasing the availability of data representative of real memes and facilitating accurate analysis of features that make multi-modal content like memes offensive. Currently, this is less possible when datasets include incorrect IWT formats, as the strategies used by memes to generate meaning are unique to user-generated meme content.

REFERENCES

[1] Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2020. A Multimodal Memes Classification: A Survey and Open Research Issues. (Sept. 2020). https://doi.org/10.48550/arXiv.2009.08395 arXiv:https://arxiv.org/abs/2009.08395

ISSN: 2057-5688

[2] Library of Congress American Folklore Centre. [n. d.]. Meme Generator: collected datasets. Available at: https://www.loc.gov/item/2018655320/ (2022-05-10).

[3] Kate Barnes, Tiernon Riesenmy, Minh Duc Trinh, Eli Lleshi, Nóra Balogh, and Roland Molontay. 2021. Dank or not? Analyzing and predicting the popularity of memes on Reddit. Applied Network Science 6, 1 (March 2021). https://doi.org/ 10.1007/s41109-021-00358-7

[4] David M. Beskow, Sumeet Kumar, and Kathleen M. Carley. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. Information Processing & Management 57, 2 (March 2020), 102170.https://doi.org/10.1016/j.ipm.2019.102170 Number: 2.

[5] Tanmoy Chakraborty and Sarah Masud. 2022. Nipping in the bud: detection, diffusion and mitigation of hate speech on social media. ACM SIGWEB Newsletter Winter (2022), 1–9.

[6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20, 1 (1960), 37–46.

[7] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021.

[8] Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. Understanding Visual Memes: An Empirical Analysis of Text Superimposed on Memes Shared on Twitter. Proceedings of the International AAAI Conference on Web and Social Media 14, 1 (May 2020), 153–164. https://doi.org/10.1609/icwsm.v14i1.7287

[9] Abhimanyu Dubey, Esteban Moro, Manuel Cebrian, and Iyad Rahwan. 2018.MemeSequencer: Sparse Matching for Embedding Image Macros. In Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18. ACM Press, Lyon, France, 1225–1235. https://doi.org/10.1145/3178876.3186021

[10] Marta Dynel. 2016. "I has seen image macros!" Advice animals memes as visualverbal jokes. International Journal of Communication 10 (2016), 29.

[11] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Flickr8k Dataset. Journal of Artificial Intelligence Research 47 (2013), 853–899.

ISSN: 2057-5688

[12] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In Proceedings of the IEEE conference on computer vision and pattern recognition (Honolulu, HI, USA). IEEE, 1705–1715. https://doi.org/10.1109/CVPR.2017.123

[13] JaidedOCR. 2022. EasyOCR. Available at: https://www.jaided.ai/easyocr/ (2022-05-10).

[14] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In Advances in Neural Information Processing Systems NIPS'20 (Vancouver, BC, Canada), Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, Article 220, 2611–2624 pages.

[15] Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the Wild: Assessing the Generalizability of the Hateful Memes
Challenge Dataset. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). Association for Computational Linguistics, Online, 26–35. https://doi.org/10.18653/v1/2021.woah-1.4

[16] Michele Knobel and Colin Lankshear. 2007. Online memes, affinities, and cultural production. A new literacies sampler 29 (2007), 199–227. Publisher: New York.