2023

Meme Sentiment Analysis Enhanced with Multimodal Spatial Encoding and Face Embedding

Muzhaffar Hazman
University of Galway, Galway, Ireland

Susan McKeever
Technological University Dublin, Ireland, susan.mckeever@tudublin.ie

Josephine Griffith
University of Galway, Galway, Ireland

Meme Sentiment Analysis Enhanced with Multimodal Spatial Encoding and Face Embedding

Muzhaffar Hazman[1(✉)] , Susan McKeever[2] , and Josephine Griffith[1]

[1] University of Galway, Galway, Ireland
{m.hazman1,josephine.griffith}@universityofgalway.ie
[2] Technological University Dublin, Dublin, Ireland
susan.mckeever@TUDublin.ie

Abstract. Internet memes are characterised by the interspersing of text amongst visual elements. State-of-the-art multimodal meme classifiers do not account for the relative positions of these elements across the two modalities, despite the latent meaning associated with where text and visual elements are placed. Against two meme sentiment classification datasets, we systematically show performance gains from incorporating the spatial position of visual objects, faces, and text clusters extracted from memes. In addition, we also present facial embedding as an impactful enhancement to image representation in a multimodal meme classifier. Finally, we show that incorporating this spatial information allows our fully automated approaches to outperform their corresponding baselines that rely on additional human validation of OCR-extracted text.

Keywords: Multimodal deep learning · Sentiment analysis · Internet memes

1 Introduction

The sentiment polarity classification task traditionally entailed analysing a piece of natural language text to classify its sentiment as negative, positive, or neutral. Sentiment analysis was initially performed on text. The growth of user-generated multimodal content (e.g., videos, image-caption pairs) has motivated the extension of affective computing techniques to input types beyond text [9]. Multimodal sentiment analysis poses the same questions as its text-only predecessor, but is extended to inputs comprising multiple modalities simultaneously. When faced with multimodal inputs, Poria et al. [9] describe unimodal encoders as crucial building blocks of multimodal systems, each encoder directly contributing to the resultant performance. Furthermore, the fusion of unimodal representations also plays a key role by providing "surplus information" to the classifier [9].
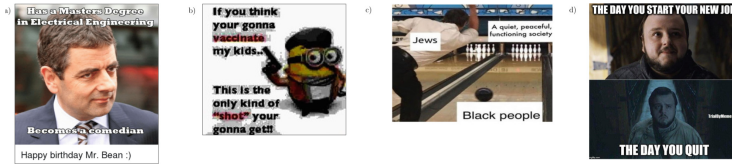
Fig. 1. Sample memes with a) *Positive* sentiment [8] and b) *Negative* sentiment [8], c) hateful spatial analogies [13], and d) spatial segments [14].

Along with the advent of other multimodal formats of user-generated content, Internet memes (or simply "memes") have proliferated. Memes are commonly found in various online communities to communicate ideas, incite humour, and express emotions. Automated analysis of memes allows for: including memes in automated opinion mining processes [9], taking action against meme-based hate speech [6,13], identifying disinformation campaigns [1], and investigating social and political cultures [5]. This work contributes to the underlying problem of **sentiment polarity classification of a meme**: "Given a meme in a visual format, comprising an image I with embedded text T, classify the meme as having the overall sentiment of either *Negative* (e.g., Fig. 1b), *Positive* (e.g., Fig. 1a), or *Neutral*".

Memes are challenging input in automated affective classification problems, as they typically exhibit very brief texts, references to popular culture, subtle intermodal semantic relations, and dependence on background context [11,13, 13,17]. Thus, solutions must consider the semantics of each, the textual and visual modalities, and their combinations [6]. The breadth of this challenge spans various affective goals, including sentiment polarity [8,14], offensiveness [6,8,14], sarcasm [8,14], and motivational intent [8,14].

Recent work has shown that incorporating additional relevant information improves the performance of meme affective classifiers [11], amongst which is positional information of words within text and visual objects within an image [13,17]. Unlike many other forms of multimodal content, the text within a meme is interspersed into its image, often either superimposed on the image or comprising a segment of the meme image, creating a shared visual medium. Meme authors intentionally position a grouping of words ("text clusters") to convey meaning, such as implying hateful analogies [13] (e.g., Fig. 1c); text clusters can be paired with image segments, with each pair signifying a different sentiment (e.g., Fig. 1d). Current approaches that use positional information in meme sentiment classification opt to omit intermodal positional relations, i.e. they consider the position of a word amongst text but not its position in relation to the meme image or vice versa.

This work proposes injecting the spatial information of features from both modalities of a meme into a deep learning multimodal classifier to improve sentiment classification performance. Crucially, we account for the interspersing of
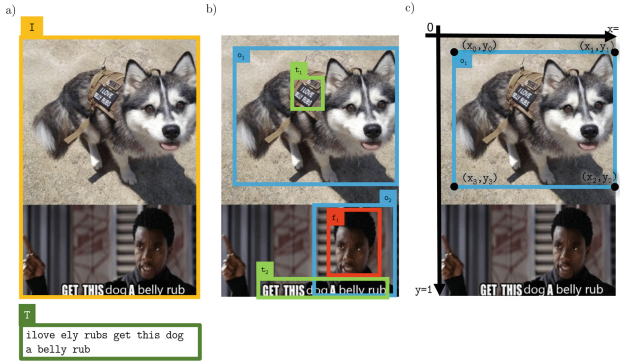
Fig. 2. Sample meme [8] a) showing the image and text modalities, I and T, as given in the dataset; b) bounding boxes generated for local features: text clusters (t_1 and t_2), objects (o_1 detected as "Dog" and o_2 as "Person"), and faces (f_1); and c) the coordinate system used to generate the spatial encoding for each bounding box (e.g. the vertices of o_1, p_{o_1}).

visual objects and text clusters by representing the spatial position of each on a shared coordinate system ("spatial encoding"). We append the spatial encoding of visual objects (e.g. o_1, o_2 in Fig. 2b), faces (e.g. f_1 in Fig. 2b), and text clusters (e.g. t_1, t_2 in Fig. 2b) to their local representations prior to multimodal fusion and classification. The performance implication of spatial encodings and local representations are systematically evaluated on two benchmark datasets using the seven models described in Sect. 3.2. To the best of our knowledge, this work is the first to use shared coordinate spatial encoding and deep representation of faces to tackle the sentiment classification of memes.

2   Related Works

2.1   Meme Affective Classifiers

Memes are distinct from other multimodal user-generated content types in several key ways. First, the text and image of a meme share a common visual medium, unlike the more common image-caption pairs. Text in memes is often intentionally located amongst other visual content to create meaning [13]. Second, memes use short text pieces and few foreground visual objects, relying on intermodal relations to convey meaning. Kiela et al. [6] show how harmless images and texts could be combined to create hateful memes. Furthermore, slight changes in either modality can change a hateful meme into a harmless one and vice versa. Therefore, meme classifiers must be able to learn subtle intermodal relationships with very limited input.

Architecturally, the current literature suggests that various affective classification tasks can be applied to memes without requiring entirely distinct

approaches. Most apparently, Bucur et al.'s [3] winning submission of the Memotion 2022 Challenge [8], was trained to simultaneously classify sentiment polarity, offensiveness, sarcasm, humour, and motivational intent. Their findings suggest that meme classification architectures exhibit adaptability across different affective computing tasks. Furthermore, Pramanick et al. [10], who reported the best-performing sentiment classification solution to the Memotion 1.0 dataset [14], showed that the same architecture outperforms all, or all but one, competing solution when individually trained on eight affect dimensions.

A typical approach to building a multimodal meme classifier is to generate unimodal representations of each modality before fusing these representations into a multimodal representation of the meme, such as in [3,10,11,13]. Furthermore, the literature presents a wide range of deep learning representations used for each visual and textual modality [6,8,14], with no clear evidence that any of the options would consistently outperform all others.

2.2   Positional Encoding

Positional encoding plays a central role in the Transformers architecture [15] and has seen wide adoption in tackling various natural language tasks. It describes the position of tokens, such as a word in a sentence or a region in an image, within the input. However, since most multimodal meme classifiers employ unimodal encoders, the positions of text and visual elements are encoded separately.

To the best of our knowledge, a positional encoding that is shared between the text and image modalities on a common spatial coordinate system (a "spatial encoding") has not been applied to classifying meme sentiment. None of the architectures reportedly used to learn meme sentiment classification in [14] and [8] did so using a positional information from a coordinate system shared between modalities. Further, we were not able to find a pre-trained multimodal Transformer that readily supports such a shared encoding.

In this task, Pramanick et al. [10] showed performance gains by segmenting the text modality into text clusters but did not explicitly represent the spatial position of each cluster. To classify hateful memes, Zhu [17] employed a patch detector to divide each meme into "image regions". They then appended each text token with a representation of its surrounding image patch. However, they did not present the performance gains solely attributable to this approach. Further, we posit that such a patch-based definition of position would not be suitable where multiple text clusters are placed within the same image patch (e.g., Fig. 1c) or where a patch consists only of text (e.g. Fig. 1a).

Shang et al. [13] proposed a more general representation of spatial position by appending the spatial encoding of extracted visual objects and text clusters prior to input into an intermodal co-attentive pooling module based on a design from [7]. They attributed their model's outperforming of other leading hateful meme classifiers to its "awareness" of offensive intermodal analogies: the purposeful superimposing of a text cluster near to a visual object is used to represent an offensive conceptual comparison. While their approach is predicated solely on offensive spatial analogies, we posit that this approach could capture

a broader category of intermodal spatial relationships, including those captured by Pramanick et al.'s [10] and Zhu's [17] approaches.

2.3 Visual Feature Representations

While the image modality is commonly represented by passing the entire meme image through an image encoder [8], enhancing this representation with that of extracted visual objects has proven beneficial in classifying hateful memes [11,13,17]. One such approach is to input the meme image into Google Cloud Vision API's Web Entity Detection to create a corresponding description or set of attributes in text format [11,17]. Zhu [17] also demonstrated further performance improvement with the inclusion of Race and Gender tags for each face using a pre-trained FairFace classifier. Pramanick et al. [11] also showed improved performance by representing cropped images of visual objects and faces with VGG-19. Shang et al. [13] also found that their multimodal classifiers perform best when global and local visual feature representations are available.

The use of faces to convey sentiment is neither new nor unique to memes. Firstly, visual sentiment analysis [16] points to facial expressions as a valuable mid-level feature in classifying the sentiment conveyed by images from social networks. Second, facial expression emojis have been shown to be informative in supporting the sentiment classification of textual social media [2]. In memes, Zhu [17] argues that expecting a global image encoding to sufficiently recognise facial features that are predictive of hatefulness is unreasonable given the size of current meme datasets. Although we agree with Zhu's argument, we posit that their approach omits other information conveyed by faces that may indicate a meme's sentiment, such as emotion, expression, and identity.

3 Methodology

In this work, we evaluate the performance of seven novel multimodal classifier models. These models are separately trained on two competition datasets, Memotion 1.0 [14] and Memotion 2.0, [8], to classify the sentiment polarity of memes. We first designed and evaluated a multimodal deep learning model to establish baseline performance. This model is then repeatedly augmented to answer our research questions. Augmentations include incorporating spatial information of faces, visual objects, and text clusters and are described for each model in Table 3. Evaluation is conducted based on the differences in macro-averaged and weighted-averaged F1 scores – metrics prescribed by the authors of the datasets [8,14] – between pairs of models that respectively include and exclude each augmentation. This section presents details of the datasets and models used.

### 3.1 Dataset and Feature Extraction

This work utilises datasets presented in the SemEval 2019 Memotion 1.0 [14] ("**Memo1**") and AAAI 2022 Memotion 2.0 [8] challenges ("**Memo2**"). Both

Table 1. Samples per dataset.

Dataset	Memo1			Memo2		
	Train	Val	Test	Train	Val	Test
`Original`						
Positive	4,156	–	1,099	1517	325	78
Neutral	2,204	–	584	4510	975	971
Negative	631	–	172	973	200	451
Total	**6,991**	**–**	**1,855**	**7,000**	**1,500**	**1,500**
`Filtered & Filtered-OCR`						
Positive	3,450	609	1,067	1,453	192	76
Neutral	1,837	324	572	4,363	951	939
Negative	518	92	169	941	317	442
Total	**5,805**	**1,025**	**1,808**	**6,757**	**1,460**	**1,457**

are collections of user-generated memes labelled with one of three exclusive sentiment classes. The authors of the datasets extracted text from each meme with an automated OCR tool and then manually corrected any erroneous text extraction. For our experiments, the samples from Memo1 and Memo2 are kept separate. Without filtering or pre-processing, these samples comprise our `Original` datasets that we use to compare our `Baseline` model to leading solutions.

For each meme in these datasets, we localised, extracted, and represented its text clusters, faces, and visual objects using the tools listed in Fig. 3. The maximum counts of text clusters, visual objects, and faces are set to 18, 10,
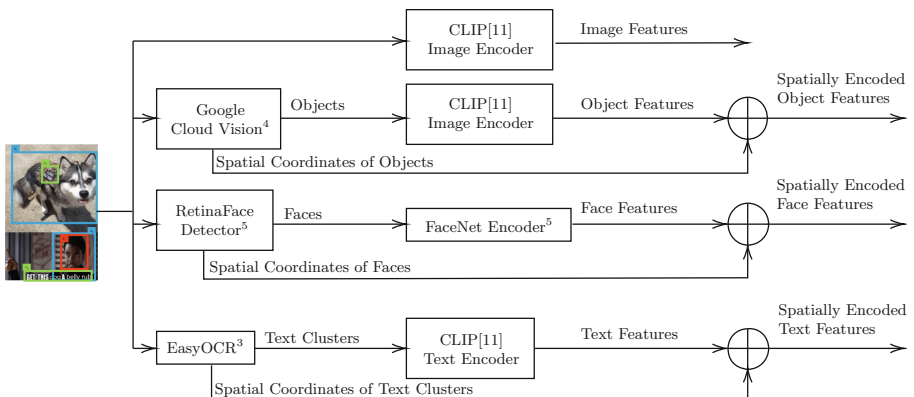
Fig. 3. Localisation and representation process applied to each meme to extract its Image, Object, Face and Text features. [3]https://github.com/JaidedAI/EasyOCR; `paragraph` option set to true. [4]https://cloud.google.com/vision/docs/object-localizer. [5]Using DeepFace wrapper from https://github.com/serengil/deepface.

and 5, respectively, with padding used for memes with fewer. Padding for text clusters is defined by passing an empty string into the CLIP text encoder, while that for visual objects is the CLIP encoding of a blank image, and zero–padding is used for faces.

Since this work applies to memes that contain identifiable visual objects and text clusters, we removed meme samples that do not meet these criteria to make up the `Filtered` datasets. This filtering is performed on all subsets of Memo1 and Memo2. As Memo1 did not contain a designated validation set, we defined one by splitting the training set – as reported by the authors of the Memo1 dataset and used in submissions to their competition [14] – with a random 85:15 sampling, weighted by the sentiment class, to maintain the target distribution. We maintained the train-validation-test splits defined for Memo2 [8]. Meme samples with identifiable visual objects but no detected faces are given face feature representation made up entirely of padding.

Finally, the `Filtered-OCR` datasets replace the text of each meme in `Filtered` with that returned in our feature extraction OCR step. Unlike in [8,10,14], we excluded any additional human validation during the OCR extraction process. All models are trained, validated, and tested on the resultant `Filtered-OCR` datasets. The counts of memes in each dataset and sentiment labels are shown in Table 1.
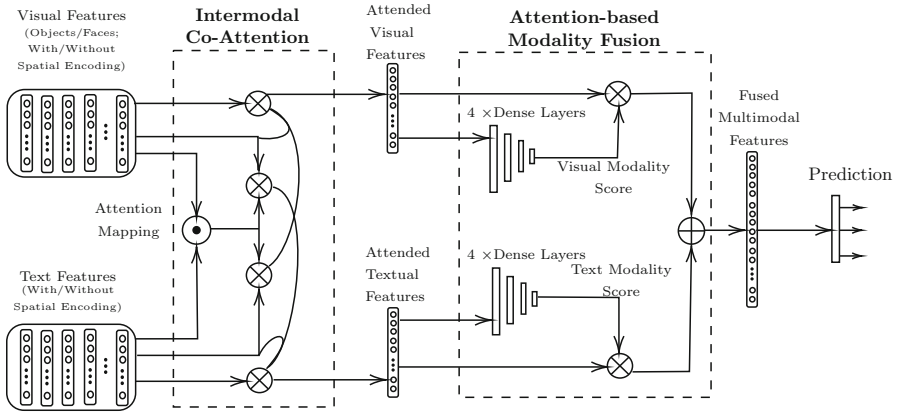
3.2   Models

This section describes the architectural characteristics of our models as listed in Table 2 and illustrated in Fig. 4. Each was built using PyTorch and trained with a triangular cyclical learning rate schedule ranging between $1e-4$ and $1e-3$ with a step size of 52 mini-batches of 512 samples. During training, validation performance was monitored for overfitting or until each model was trained for 100 epochs. Training is carried out using AdamW optimiser with weight decay of $5e-1$, betas of 0.1 and 0.25 to minimise negative log-likelihood loss with class weights inversely proportional to its sample count in the training dataset. All non-pretrained weights are initialised with a zero-mean Gaussian distribution with standard deviation 0.02, while pretrained weights are not fine-tuned. The same hyperparameter settings are maintained across all models as they are separately trained on the datasets.

Leading meme sentiment classifiers use a variety of architectures with little indication of which is most optimal. For our `Baseline` model, we drew inspiration from the typical overall approach used in leading solutions to the Memotion 2.0 Challenge [8]: each modality is represented using a pretrained encoder. Then, these representations are fused, often with a multimodal attention mechanism, and finally passed to a fully connected layer.

To encode the meme image and text (see I and T in Fig. 2a) in our `Baseline` model, we opted to use the pretrained image and text encoders of CLIP [12], respectively, which has shown comparable performance to other multimodal approaches [11]. In addition, CLIP image encodings have been shown to outperform various other image encoders in the zero-shot classification of hateful

Fig. 4. Architecture of our `Obj-NoSpatial`, `Obj-Spatial`, `Face-NoSpatial`, and `Face-Spatial` models. The Image features used in `Img-Obj-Spatial` and `Img-Face-Spatial` models bypasses the Intermodal Co-Attention module and requires the Attention-Based Modality Fusion module to be expanded with another set of dense layers. This work's `Baseline` model does not include the Intermodal Co-Attention module. Sources: Intermodal Co-Attention [7,13]; Attention-based Modality Fusion [4,10]

memes [12] and are used by the winning solution of the Memo2 challenge [3]. We chose the ViT–B/16 variant of CLIP while Pramanick et al. [11] and Bucur et al. [3] did not report their chosen variant.

Since attentive fusion has been shown to perform well on several meme problems [10], we included one in our models. Our `Baseline` model fuses the CLIP representations of the meme image and text using Gu's [4] attentive modality fusion mechanism, as used in [11]. We defined the sizes of the four dense layers as 256, 64, 8, and 1, which produces an attention score for each modality. The attention-weighted representation of each modality is concatenated and passed into a GeLU-activated dense layer followed by a log-softmax activation to output predicted logits of each sentiment class.

This model is trained on the `Original` dataset to allow performance comparisons with previously published works. We then evaluated this model on the `Filtered` and `Filtered-OCR` datasets. In the latter, the content of all text clusters t_n is concatenated and entered into the text encoder. The difference in the performance of this model on these two datasets allows us to measure the performance impact resulting from our OCR-based text extraction output relative to the human-curated approach used by the authors of the datasets [8,14].

The `Obj-NoSpatial` and `Face-NoSpatial` models remove the meme image and text, I and T per `Baseline`. As inputs, the former takes CLIP-encoded visual objects, $o_1, o_2, ..., o_j$, and text clusters extracted from a meme, $t_1, t_2, ..., t_i$. Instead of objects, the `Face-NoSpatial` model takes the FaceNet representation of faces, $f_1, f_2, ..., f_k$. Then, the j visual objects or k face representations are

passed through co-attentive weighted pooling against i text clusters as used in [13] but without spatial encodings. This step allows the models to learn attention maps between each object/face and each text cluster; producing a one-dimensional vector representing each modality. This representation replaces that of the image modality as input into the attentive fusion mechanism described for the `Baseline` model.

The `Obj-Spatial` and `Face-Spatial` models introduce the spatial encodings of each text cluster, p_{t_i}, as well as for visual objects, p_{o_j}, and faces, p_{f_k}, respectively. We augment the co-attentive pooling module in `Obj-NoSpatial` and `Face-NoSpatial` into the co-attentive analogy alignment module proposed in [13]. This is performed by appending each object's and cluster/face's representation vector with its spatial encoding. The padding for spatial encodings is defined as zeros for all coordinates.

The `Img-Obj-Spatial` and `Img-Face-Spatial` models each combine the CLIP representation of the meme image, I, into `Obj-Spatial` and `Face-Spatial`, respectively. Since these models make use of three representations per meme – image, text clusters and objects/faces – we extend Gu's [4] fusion mechanism to accommodate three inputs by introducing a third set of dense layers.

Table 2. Goals of each experimental model.

Model	Dataset	Goal
Baseline	Original	Benchmarks our chosen modality encodings and fusion mechanism against leading solutions
	Filtered	Establishes baseline performance on samples with detectable text clusters and visual objects
	Filtered-OCR	Measures the impact of replacing human-curated text replaced with text clusters returned by automated OCR. Also, establishes a baseline for our fully automated approaches
Obj-NoSpatial	Filtered-OCR	Measures the performance of representing the image modality using only CLIP-encoded localised visual objects without spatial encodings
Obj-Spatial	Filtered-OCR	Measure the performance impact of including spatial encodings of objects and text clusters
Img-Obj-Spatial	Filtered-OCR	Maximises available visual information by augmenting image input with objects and text clusters and respective spatial encodings
Face-NoSpatial	Filtered-OCR	Measures the performance of representing the image modality using only embeddings of localised faces without spatial encodings
Face-Spatial	Filtered-OCR	Measure the performance impact of including spatial encodings of faces and text clusters
Img-Face-Spatial	Filtered-OCR	Augments image input with faces and text clusters and respective spatial encodings

Table 3. Performance of our `Baseline` model against leading solutions on the Memo1 dataset. Sources: [10,14].

Solution	Macro-F1
MHA-Meme[a] [10]	**0.3762**
Vkeswani IITK	0.3547
Our `Baseline`	0.3546
Guoym	0.3520
Aihaihara	0.3502
Sourya Diptadas	0.3476
Irina Bejan	0.3469

[a] Not a competition submission; results based on subset of the original dataset

Table 4. Performance of our `Baseline` model against leading solutions on the Memo2 dataset. Source: [8].

Solution	Weighted-F1
BLUE	**0.5318**
BROWALLIA	0.5255
Yeti	0.5088
Little Flower	0.5081
Greeny	0.5037
Our `Baseline`	0.5035
Amazon PARS	0.5025

4 Results

Evaluating the `Baseline` model on the `Original` datasets places it within the top six highest performing solutions on each respective dataset; see Tables 3 and 4.

The performance of the `Baseline` model on the `Original, Filtered` and `Filtered-OCR` datasets are shown in Table 5. The lower performance of the model on the `Filtered` dataset than on the `Original` dataset likely stems from the removal of samples that contain only text on an object-less background. Classifying such samples is similar to discerning the sentiment of unimodal text inputs and is beyond the scope of this work. We attribute the performance decrease of the `Baseline` model on the `Filtered-OCR` vs. `Filtered` datasets to the lower quality of the text extracted with our automated OCR process relative to human-curated text. Despite this, our spatially aware models are able to overcome this performance penalty. The model that performs best on each dataset – as seen in Table 6 – constitutes **fully automated approaches** that outperform their respective `Baseline` models trained on the human-curated text

Table 5. Weighted F1 (F1-W) and Macro F1 (F1-M) for the `Baseline` model on all datasets.

Dataset	Memo1		Memo2	
	F1-W	F1-M	F1-W	F1-M
`Original`	**0.481**	**0.355**	**0.504**	**0.325**
`Filtered`	0.475	0.327	0.503	0.314
`Filtered-OCR`	0.462	0.326	0.439	0.283

Table 6. Weighted F1 (F1-W) and Macro F1 (F1-M) for all models on the Memo1 and Memo2 `Filtered-OCR` datasets. **Rel.** indicates relative performance to model stated in the **Comparison** column on each given dataset.

Model	Comparison	Memo1			Memo2		
		F1-W	F1-M	Rel.	F1-W	F1-M	Rel.
Baseline	-	0.462	0.326	-	0.439	0.283	-
Obj-NoSpatial	vs. Baseline	0.452	0.307	↓	0.427	0.271	↓
Obj-Spatial	vs. Obj-NoSpatial	0.481	0.317	↑	0.482	0.305	↑
Img-Obj-Spatial	vs. Obj-Spatial	**0.489**	0.336	↑	0.499	0.300	↑↓
Face-NoSpatial	vs. Baseline	0.476	0.340	↑	0.471	0.298	↑
	vs. Obj-NoSpatial			↑			↑
Face-Spatial	vs. Face-NoSpatial	0.485	**0.341**	↑	0.496	0.310	↑
	vs. Obj-Spatial			↑			↑
Img-Face-Spatial	vs. Face-Spatial	0.473	0.332	↓	**0.509**	**0.314**	↑
	vs. Img-Obj-Spatial			↓			↑

from the `Filtered` datasets. By removing the need for manual intervention, fully automated models improve the feasibility of conducting sentiment classification of memes at scale, and reduce the effort necessary for creating future meme datasets.

The results show that spatial encoding improves performance. `Obj-Spatial` and `Face-Spatial` each outperforms `Obj-NoSpatial` and `Face-NoSpatial` respectively. These results point to intermodal spatial information being informative for the problem task and not sufficiently represented by the CLIP encodings of the whole meme image. This finding holds significance to applying deep learning solutions on memes in particular, as the text modality is incorporated and interspersed within the image. Although the importance of token positions in leading solution architectures has been well established, the lack of a shared visual medium for image and text modalities in many other vision-language tasks has resulted in leading multimodal architectures with separate positional representations for each modality. Based on our results, we argue that spatial encodings should also be considered for other vision-language tasks where visual objects and text share a common visual medium.

The performance benefit of representing the image modality with localised visual feature representations depends on whether the features are defined as objects or faces. CLIP-encoded object representation performs worse than `Baseline`. This results from a reduction in the visual information available to the image encoder. However, `Face-NoSpatial`, which uses FaceNet embeddings to represent faces, outperforms both `Obj-NoSpatial` and `Baseline` while also suffering from the same, if not greater, reduction in available visual information. Furthermore, `Obj-Spatial` showed mixed results against `Baseline`, while `Face-Spatial` outperforms `Baseline` in both datasets. Notably, faces are not entirely excluded from models based on visual objects, as many meme sam-

ples had "Person" as a detected object. Thus, we believe that the performance difference between the two approaches arises from the more fine-grained facial embedding provided by FaceNet and the inherent exclusion of non-face visual objects that emphasises the contribution of faces to the sentiment of a meme.

We found that augmenting the meme image with local representations of either objects or faces and their spatial encodings consistently outperforms models that rely on the image alone. However, choosing between CLIP-encoded objects versus FaceNet-encoded faces as augmentations to the meme image proved inconsistent and dependent on the dataset. Although `Img-Obj-Spatial` and `Img-Face-Spatial` perform the best in the Memo1 and Memo2 datasets, respectively, their performance relative to `Obj-Spatial` and `Face-Spatial` appears to depend on the dataset. Drops in performance here may stem from redundant intermodal information (e.g. between global image and objects-based representations). Unlike in [10], we did not employ any form of learned cross-modal filtering.

5   Conclusions

In this work, we addressed spatial encoding and facial embedding in classifying sentiment polarity of internet memes. We developed seven novel architectures, and evaluated each on two challenge datasets. For both datasets, our proposed baseline multimodal classifier ranked within the top six of leading state-of-the-art solutions on both datasets. While we found that representing the image modality with visual objects alone does not consistently offer performance benefits, a face-based representation does. Furthermore, the incorporation of spatial information of these visual features grants performance improvements over both image-only and faces-/objects-only approaches. For each of the Memotion datasets, our top performing solution comprises augmenting the image modality with spatially encoded visual features and text clusters. We propose these solutions as fully automated competitive alternatives to current state-of-the-art solutions that rely on manual validation of OCR-based text extraction.

References

1. Al-Rawi, A.: Political memes and fake news discourses on instagram. Media Commun. **9**(1), 276–290 (2021). https://doi.org/10.17645/mac.v9i1.3533
2. Ayvaz, S., Shiha, M.: The effects of emoji in sentiment analysis. Int. J. Comput. Electr. Eng. **9**, 360–369 (2017). https://doi.org/10.17706/ijcee.2017.9.1.360-369
3. Bucur, A.M., Cosma, A., Iordache, I.: BLUE at memotion 2.0 2022: you have my image, my text and my transformer. In: De-Factify @ AAAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection. CEUR Workshop Proceedings, AAAI (2022)
4. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I.: Hybrid attention based multimodal network for spoken language classification. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, pp. 2379–2390. Association for Computational Linguistics (2018)

5. Joshi, A., Buntain, C.: Exploiting the right: inferring ideological alignment in online influence campaigns using shared images. In: Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media (2022). https://doi.org/10.36190/2022.45

6. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C.A., et al.: The hateful memes challenge: competition report. In: Proceedings of the NeurIPS 2020 Competition and Demonstration Track. Proceedings of Machine Learning Research, vol. 133, pp. 344–360. PMLR (2021)

7. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances in Neural Information Processing Systems, vol. 29. Curran Associates, Inc. (2016)

8. Patwa, P., Ramamoorthy, S., Gunti, N., Mishra, S., Suryavardan, S., Reganti, A., et al.: Findings of memotion 2: sentiment and emotion analysis of memes. In: De-Factify @ AAAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection. CEUR Workshop Proceedings, AAAI (2022)

9. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017). https://doi.org/10.1016/j.inffus.2017.02.003

10. Pramanick, S., Akhtar, M.S., Chakraborty, T.: Exercise? I thought you said 'extra fries': leveraging sentence demarcations and multi-hop attention for meme affect analysis. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, no. 1, pp. 513–524 (2021)

11. Pramanick, S., Sharma, S., Dimitrov, D., Akhtar, M.S., Nakov, P., Chakraborty, T.: MOMENTA: a multimodal framework for detecting harmful memes and their targets. In: Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, pp. 4439–4455. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.findings-emnlp.379

12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)

13. Shang, L., Zhang, Y., Zha, Y., Chen, Y., Youn, C., Wang, D.: AOMD: an analogy-aware approach to offensive meme detection on social media. Inf. Process. Manag. **58**(5), 102664 (2021). https://doi.org/10.1016/j.ipm.2021.102664

14. Sharma, C., Bhageria, D., Scott, W., PYKL, S., Das, A., Chakraborty, T., et al.: SemEval-2020 task 8: memotion analysis- the visuo-lingual metaphor! In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, pp. 759–773. International Committee for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.semeval-1.99

15. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Nips 2017, pp. 6000–6010. Curran Associates Inc., Red Hook (2017)

16. Yuan, J., Mcdonough, S., You, Q., Luo, J.: Sentribute: image sentiment analysis from a mid-level perspective. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. Wisdom 2013. Association for Computing Machinery, New York (2013). https://doi.org/10.1145/2502069.2502079

17. Zhu, R.: Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution (2020). https://doi.org/10.48550/arxiv.2012.08290