

Unimodal Intermediate Training for Multimodal Meme Sentiment Classification

Muzhaffar Hazman¹, Susan McKeever², and Josephine Griffith¹

¹University of Galway, Ireland

²Technological University Dublin, Ireland

{m.hazman1, josephine.griffith}@universityofgalway.ie
susan.mckeever@TUDublin.ie

Abstract

Internet Memes remain a challenging form of user-generated content for automated sentiment classification. The availability of labelled memes is a barrier to developing sentiment classifiers of multimodal memes. To address the shortage of labelled memes, we propose to supplement the training of a multimodal meme classifier with unimodal (image-only and text-only) data. In this work, we present a novel variant of supervised intermediate training that uses relatively abundant sentiment-labelled unimodal data. Our results show a statistically significant performance improvement from the incorporation of unimodal text data. Furthermore, we show that the training set of labelled memes can be reduced by 40% without reducing the performance of the downstream model.

1 Introduction

As Internet Memes (or just “memes”) become increasingly popular and commonplace across digital communities worldwide, research interest to extend natural language classification tasks, such as sentiment classification, hate speech detection, and sarcasm detection, to these multimodal units of expression has increased. However, state-of-the-art multimodal meme sentiment classifiers significantly underperform contemporary text sentiment classifiers and image sentiment classifiers. Without accurate and reliable methods to identify the sentiment of multimodal memes, social media sentiment analysis methods must either ignore or inaccurately infer opinions expressed via memes. As memes continue to be a mainstay in online discourse, our ability to infer the meaning they convey becomes increasingly pertinent (Sharma et al., 2020; Mishra et al., 2023).

Achieving similar levels of sentiment classification performance on memes as on unimodal content remains a challenge. In addition to its multi-

modal nature, multimodal meme classifiers must discern sentiment from culturally specific inputs that comprise brief texts, cultural references, and visual symbolism (Nissenbaum and Shifman, 2017). Although various approaches have been used to extract information from each modality (text and image) recent works have highlighted that meme classifiers must also recognise the various forms of interactions between these two modalities (Zhu, 2020; Shang et al., 2021; Hazman et al., 2023).

Current approaches to training meme classifiers are dependent on datasets of labelled memes (Kiela et al., 2020; Sharma et al., 2020; Suryawanshi et al., 2020; Patwa et al., 2022; Mishra et al., 2023) containing sufficient samples to train classifiers to extract relevant features from each modality and relevant cross-modal interactions. Relative to the complexity of the task, the current availability of labelled memes still poses a problem, as many current works call for more data (Zhu, 2020; Kiela et al., 2020; Sharma et al., 2022).

Worse still, memes are hard to label. The complexity and culture dependence of memes (Gal et al., 2016) cause the Subjective Perception Problem (Sharma et al., 2020), where varying familiarity and emotional reaction to the contents of a meme from each annotator causes different ground-truth labels. Second, memes often contain copyright-protected visual elements taken from other popular media (Laineste and Voolaid, 2017), raising concerns when publishing datasets. This required Kiela et al. (2020) to manually reconstruct each meme in their dataset using licenced images, significantly increasing the annotation effort. Furthermore, the visual elements that comprise a given meme often emerge as a sudden trend that rapidly spreads through online communities (Bauckhage, 2011; Shifman, 2014), quickly introducing new semantically rich visual symbols into the common meme parlance, which carried little meaning before

(Segev et al., 2015). Taken together, these characteristics make the labelling of memes particularly challenging and costly.

In seeking more data-efficient methods to train meme sentiment classifiers, our work attempts to leverage the relatively abundant unimodal sentiment-labelled data, i.e. sentiment analysis datasets with image-only and text-only samples. We do so using Phang et al.’s (2019) **Supplementary Training on Intermediate Labeled-data Tasks (STILT)** which addresses the low performance often encountered when finetuning pre-trained text encoders to data-scarce Natural Language Understanding (NLU) tasks. Phang et al.’s STILT approach entails three steps:

1. Load pretrained weights into a classifier model.
2. Finetune the model on a supervised learning task for which data is easily available (the **intermediate task**).
3. Finetune the model on a data-scarce task (the **target task**) that is distinct to the intermediate task.

STILT has been shown to improve the performance of various models in a variety of text-only target tasks (Poth et al., 2021; Wang et al., 2019). Furthermore, Pruksachatkun et al. (2020) observed that STILT is particularly effective in target tasks in NLU with smaller datasets, e.g. *WiC* (Pilehvar and Camacho-Collados, 2019) and *BoolQ* (Clark et al., 2019). However, they also showed that the performance benefits of this approach are inconsistent and depend on choosing *appropriate* intermediate tasks for any given target task. In some cases, intermediate training was found to be detrimental to target task performance; which Pruksachatkun et al. (2020) attributed to differences between the required “**syntactic and semantic skills**” needed for each intermediate and target task pair. However, STILT has not yet been tested in a configuration in which intermediate and target tasks have different input modalities.

Although only considering the text or image of a meme in isolation does not convey its entire meaning (Kiela et al., 2020), we suspect that unimodal sentiment data may help incorporate *skills* relevant to discern the sentiment of memes. By proposing a novel variant of STILT that uses unimodal sentiment analysis data as an intermediate task in

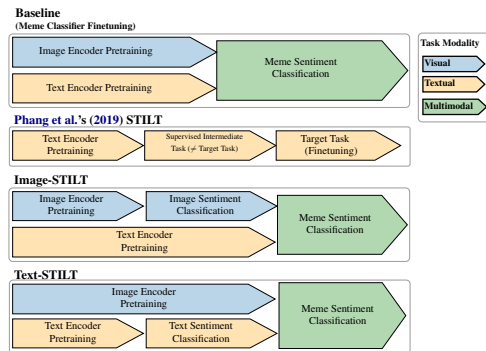


Figure 1: Training tasks in Baseline, Phang et al.’s (2019) STILT, and our proposed Image-STILT and Text-STILT approaches.

training a multimodal meme sentiment classifier, we answer the following questions:

RQ1: Does supplementing the training of a multimodal meme classifier with unimodal sentiment data significantly improve its performance?

We separately tested our proposed approach with image-only and text-only 3-class sentiment data (creating **Image-STILT** and **Text-STILT**, respectively) as illustrated in Figure 1). If either proves effective, we additionally answer:

RQ2: With unimodal STILT, to what extent can we reduce the amount of labelled memes whilst preserving the performance of a meme sentiment classifier?

2 Related Works

2.1 Meme Affective Classifiers

Meme sentiment classifiers fall within the broader category of meme affective classifiers, which can be defined as multimodal deep learning models trained to classify memes by a given affect dimension, including sentiment polarity, offensiveness, motivationality, sarcasm (Sharma et al., 2020; Patwa et al., 2022; Mishra et al., 2023), hate speech (Kiela et al., 2020), and trolling behaviour (Suryawanshi et al., 2020). Based on the majority of state-of-the-art meme classifiers, the current literature suggests that these different tasks do not require architecturally distinct solutions (Hazman et al., 2023). Broadly, two general architectural approaches exist among multimodal meme affective classifiers: first, multi-encoder models that use multiple pretrained unimodal encoders which are then fused prior to classification – numerous examples are summarised by Sharma et al. (2020)



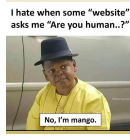

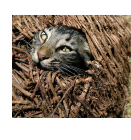
	(a) Meme			(b) Image		(c) Text	
	(i)	(ii)	(iii)	(i)	(ii)	(i)	(ii)
Input Image						—	—
Input Text	they talk about you all the time i know thats why i sent you	when the boss asks how youre doing halfway through the dinner rush	i hate when some "website" asks me "Are you human..?" No, I'm mango.	—	—	I tried a new place. I can't wait to return and try more.	My wife was disappointed.
Label	Positive	Neutral	Negative	Positive	Negative	Positive	Negative

Table 1: Sample (a) multimodal memes (Ramamoorthy et al., 2022), (b) unimodal images (CrowdFlower, 2016), and (c) unimodal text (Potts et al., 2021) from the datasets used. Unimodal images and texts of neutral sentiment not pictured here.

and Patwa et al. (2022). These models use both a text encoder and an image encoder that were each trained in unimodal self-supervised and unsupervised tasks such as BERT or SentenceTransformer for text, and VGG-19 or RESNET50 for images. In contrast, single-encoder models are based on a pretrained multimodal vision-and-language model, most often a transformer that has been pretrained on multimodal tasks and accepts both modalities as a single input. The single-encoder approach (Muenighoff, 2020; Zhu, 2020) reuses models that have been pretrained on multimodal tasks such as VL-BERT, UNITER, ERNIE-ViL, DeVLBERT and VisualBERT. There is little empirical evidence to show that one architectural approach consistently outperforms the other in the various meme classification tasks.

Typically, both multi- and single-encoder architectures use transfer learning by finetuning pretrained models on a dataset of labelled memes. While pretraining is often assumed to yield performance benefits for meme classification tasks, this has not been exhaustively proven, especially when viewed relative to studies in image- and text-only tasks (Jiang et al., 2022). Multimodally pretrained baseline models for the Hateful Memes dataset (Kiela et al., 2020) outperformed their unimodally pretrained counterparts. Suryawanshi et al. (2020) showed that the use of pretrained weights did not consistently provide performance benefits to their image-only classifiers of trolling behaviour in Tamil code-mixed memes. Although the use of pretrained encoders is common amongst meme sentiment classifiers (Sharma et al., 2022; Bucur et al., 2022; Pramanick et al., 2021a; Sharma et al., 2020; Patwa et al., 2022), there is little evidence as

to whether pretrained representations are suitable for the downstream task or if an encoder’s performance in classifying unimodal input transfers to classifying multimodal memes.

Beyond using pretrained image and text encoders, several recent works have attempted to incorporate external knowledge into meme classifiers. Some employed additional encoders to augment the image modality representation such as human faces (Zhu, 2020; Hazman et al., 2023), while others have incorporated image attributes (including entity recognition via a large knowledge base) (Pramanick et al., 2021b), cross-modal positional information (Shang et al., 2021; Hazman et al., 2023), social media interactions (Shang et al., 2021), and image captioning (Blair et al., 2021). To our knowledge, no published attempts have been made to directly incorporate unimodal sentiment analysis data into a multimodal meme classifier.

2.2 Supplementary Training of Meme Classifiers

Several recent works addressed the lack of labelled multimodal memes by incorporating additional non-meme data. Sharma et al. (2022) presents two self-supervised representation learning approaches to learn the “semantically rich cross-modal feature[s]” needed in various meme affective classification tasks. They finetuned an image and a text encoder on image-with-caption tweets before fitting these representations on to several multimodal meme classification tasks including sentiment, sarcasm, humour, offence, motivationality, and hate speech. These approaches showed performance improvement on some, but not all, tasks. In some cases, their approach underperformed in

comparison to the more typical supervised finetuning approaches. Crucially, since the authors did not compare their performance to that of the same architecture without the self-supervised step, isolating performance gains directly attributable to this step is challenging. Furthermore, while the authors reported multiple tasks where their approach performed best while training on only 1% of the available memes, their included training curves imply that these performance figures were selected at the point of maximum performance on the testing set during training. This differs from the typical approach of early stopping based on performance on a separately defined validation set, which hinders direct comparisons to competing solutions.

Bucur et al. (2022) proposed a multitask learning approach that simultaneously trained a classifier on different meme classification tasks – sentiment, sarcasm, humour, offence, motivationality – for the same meme inputs. Their results showed that multitask learning underperformed in the binary detection of humour, sarcasm, and offensiveness. This approach was found to be only effective in predicting the intensity of sarcasm and offensiveness of a meme. However, in sentiment classification, this multitask approach showed inconsistent results. Although multitask learning did not improve the performance of their text-only classifier, their multitask multimodal classifier offers the best reported results on the Memotion 2.0 sentiment classification task to date.

To the best of our knowledge, only one previous work used unimodal inputs to supplement training of multimodal meme classifiers. Suryawanshi et al.’s (2020) initial benchmarking of the TamilMemes dataset showed that the inclusion of unimodal images improved the performance of their ResNet-based image-only model in detecting trolling behaviour in Tamil memes. The authors augmented their dataset of memes with images collected from Flickr; by assigning these images as not containing trolling language. They found that this augmentation with 1,000 non-meme images decreased the performance of their classifier. With 30,000 images, their classifier performed identically to one that only used pretrained weights and supervised training on memes; both were outperformed by their model that did not use either pretrained weights or data augmentations.

Existing supplementary approaches to improve meme classification performance have shown

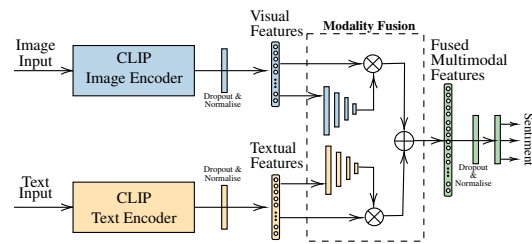


Figure 2: Our model architecture. Source: Adapted from (Hazman et al., 2023).

mixed results. Notably, the observations made in these works were measured only once and were not accompanied by statistical significance tests, necessitating caution when drawing conclusions on their effectiveness.

3 Methodology

To address our research questions, we chose the 3-class sentiment polarity of multimodal memes as our target task as defined by Ramamoorthy et al. (2022) for our chosen dataset. Our experimental approach revolves around comparing the performance of a multimodal classifier trained only on memes (our **Baseline**) and those trained first on unimodal image or text data (our **Image-STILT** and **Text-STILT** models, respectively) before being trained on memes. These models are architecturally identical to each other, all trained in the Memotion 2.0 training set and tested against the Memotion 2.0 testing set to isolate the effect of unimodal intermediate training on meme sentiment classification performance. The results of the performance of the model are measured using the weighted F1-score, as defined by the authors of the selected meme dataset (Sharma et al., 2022). A detailed description of this metric is available in Appendix B.

3.1 Model Architecture

As this work does not seek to propose a new meme classifier architecture, we heavily base our model on one found in literature: the **Baseline** model proposed by Hazman et al. (2023). Per this previous work, we also use the image and text encoders from OpenAI CLIP (Radford et al., 2021) to represent each modality, respectively, and the same modality fusion weighting mechanism they had used. However, we added dropout and batch normalisation after encoding each modality and the fusion of these encodings, which were helpful in preventing overfitting. Figure 2 illustrates our architecture and a detailed description is presented in

Appendix C.

3.2 Datasets

Multimodal Memes: This work uses sentiment-labelled multimodal memes from the Memotion 2.0 (Ramamoorthy et al., 2022) benchmark dataset as our target task. We did not use the earlier (Sharma et al., 2020) and later (Mishra et al., 2023) iterations of this dataset as the former did not provide a validation set and the latter focused on code-mixed languages. Each sample in this meme dataset comprises a meme collected from the web that was then labelled by multiple annotators as conveying either a Positive, Negative or Neutral sentiment. For each meme sample, the dataset presents an image file and a string of the text that was extracted using OCR with manual validation.

To assess the effectiveness of our approach on various amounts of labelled memes available for training, that is, to answer RQ2, we defined fractional training datasets by randomly sampling the memes training set at the following fractions: 5, 10, 20, 30, 40, 50, 60, 70, and 80%. For each random restart, we repeat this sampling to account for variance in model performance attributable to training data selection. Where matched pairs are needed for hypothesis testing (see Section 3.4.RQ2 below), we do not resample between training Baseline, Image-STILT and Text-STILT models. To prevent the models from converging into a model that predicts only the most prevalent class in the training set, we balance the classes in these fractional datasets by applying weights inverse of the class distribution during sampling without replacement.

Unimodal Images and Texts: For unimodal intermediate training, we used two unimodal datasets: Crowdflower (CrowdFlower, 2016) for unimodal images, and DynaSent (Potts et al., 2021) for unimodal text. Both datasets comprise crowdsourced samples collected from social networking sites,

Dataset		Samples			Total
		Pos	Neu	Neg	
Memotion 2.0	Train	1,517	584	172	7,000
	Val	325	975	200	1,500
	Test	78	971	451	1,500
Crowdflower		5,313	1,259	1,227	7,799
DynaSent		6,038	5,782	4,579	16,399

Table 2: Meme, Image and Text sample counts in the Memotion 2.0 (Ramamoorthy et al., 2022), Crowdflower (CrowdFlower, 2016), DynaSent (Potts et al., 2021), respectively.

and both contain crowd-annotated 3-class sentiment labels¹. We included all images from the CrowdFlower dataset that we were able to fetch via the provided URLs; not all samples were retrievable. The summaries of, and examples from, these datasets are presented in Tables 2 and 1, respectively.

3.3 Training

Baseline: For each run, the model is initialised by loading pretrained weights for the encoders and randomly initialising the weights in the fusion mechanism. For our Baseline approach, the model is trained on the Memotion 2.0 training set, with early stopping at the point of minimum loss on the validation set, and evaluated against the testing set. We maintain the dataset splits defined by Ramamoorthy et al. (2022).

Unimodal STILTs – Image-STILT and Text-STILT:

In our proposed approaches, the initialisation of the model is the same as for Baseline and is followed by training the model on a selected unimodal dataset while freezing the encoder of the other modality, that is, the text encoder is frozen while training on unimodal images in Image-STILT and vice versa. Unimodal training ends with early stopping based on the model’s performance on the Memotion 2.0 validation set. This model is then trained and tested on the Memotion 2.0 training and testing sets, respectively, as was done in the Baseline approach. Hyperparameters used when training on the Memotion 2.0 dataset are kept constant across all models (see Appendix A).

3.4 Experimental Approach

RQ1: To establish whether Image-STILT or Text-STILT offers a statistically significant performance improvement over Baseline, we employ the Wilcoxon Signed-Rank test. The null hypothesis in each case is that there is no significant performance difference between our Baseline approach and Image-STILT or Text-STILT, respectively. We ran 10 random-restarts for each approach: Baseline, Image-STILT, and Text-STILT. All models were trained on all memes from the Memotion 2.0 (Ramamoorthy et al., 2022) training set. Separate tests were conducted for (1) Baseline vs. Image-STILT and (2) Baseline vs. Text-STILT; resulting in a total of 10 pairs each for hypothesis testing.

¹CrowdFlower’s Highly Negative and Highly Positive are treated as Negative and Positive.





	Image		Meme	
	(a)	(b)	(c)	(d)
				
Sentiment	Positive	Negative	Positive	Negative

Table 3: Example unimodal images and multimodal memes showing distinct visual symbols.

RQ2: To characterise the performance benefits of Image-STILT or Text-STILT with limited availability of labelled memes, we train Baseline, Image-STILT and Text-STILT on varying amounts of training memes. For each approach and at each of the training set sizes, we ran five random-restarts, resulting in 45 observations for each Baseline vs. Image-STILT and Baseline vs. Text-STILT, separately. For each random restart, we resample the training set, we define a matched pair (as required by Wilcoxon Signed-Rank test assumptions) as the performance of two models having been trained on the same set of memes. We performed a Wilcoxon Signed-Rank test across the entire range of labelled meme availability, but separately for Baseline vs. Image-STILT and Baseline vs. Text-STILT.

4 Results

4.1 RQ1: Performance Improvement

Text-STILT was found to outperform Baseline, at a level of statistical significance. Figure 3 and Table 4 show the performance distribution of each approach, with 10 random restarts each. The Wilcoxon Signed-Rank test resulted in p-values of 0.193 and 0.0273 for Baseline vs. Image-STILT and Baseline vs. Text-STILT, respectively.

To our knowledge, Text-STILT is the first ap-

Approach	F1	Prec	Rec	p-value vs. Baseline
Baseline	51.19 (0.00393)	54.86 (0.0112)	56.37 (0.00662)	-
Image-STILT	51.45 (0.00485)	54.96 (0.0149)	58.78 (0.0142)	0.193
Text-STILT	51.78 (0.00659)	56.58 (0.0131)	57.66 (0.00950)	0.0273

Table 4: Mean of Weighted F1-score, Precision and Recall and their standard deviation (in parantheses) for Baseline, Image-STILT & Text-STILT, across 10 runs each.

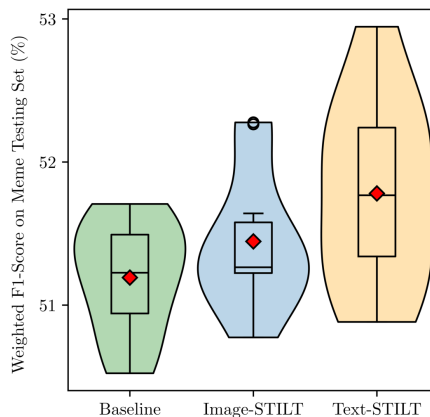
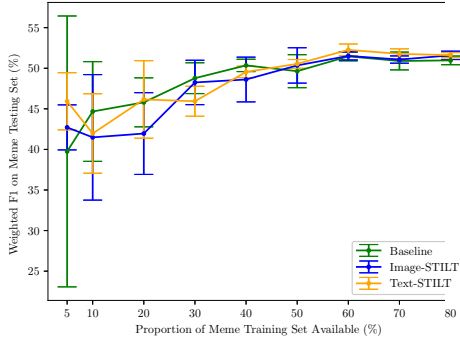


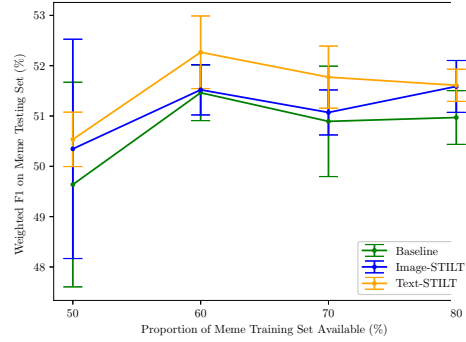
Figure 3: Performance of the Baseline, Image-STILT, and Text-STILT. Box-plots indicate the 2nd - 3rd quartile range and \blacklozenge indicates mean performance.

proach to successfully incorporate supplementary unimodal data into the training of multimodal meme classifiers showing a statistically significant performance improvement. However, our results do not indicate why Text-STILT was effective. We posit that while each meme’s semantics rely on both the image and text modalities, memes that contain longer texts and/or a textual structure that hints at the meme’s overall sentiment are more accurately classified by Text-STILT (see examples in Table 5). Consider the meme in Table 5(b): While the negative component is represented visually, that is, the bottom image segment, the structure of the text “what people think... what it really is like...” strongly suggests a negative inversion of something normally considered to be positive. Thus, its negative sentiment could be inferred largely from text alone. More rigorous investigation of the relationship between text and meme sentiment analysis is warranted.

Although Text-STILT significantly outper-



(a) 5% to 80% of Memes available.



(b) 50% to 80% of Memes available.

Figure 4: Baseline, Image-STILT, and Text-STILT performance across varying amount of memes available; 5 random restarts each.

formed Baseline, Image-STILT did not. Although Image-STILT shows higher mean, maximum, and minimum performance than Baseline, the distribution (see the violin plot in Figure 3) indicates that the two performed similarly. This could be attributed to the distinct role of visual symbols in memes, which derive their meaning from popular usage rather than literal connotations. Consider the memes in Table 3, each made using highly popular **meme templates**: *Success Kid*² and *Bad Luck Brian*³, respectively. These have come to symbolise specific meanings through online usage, which is distinct from what is literally shown. In the case of *Bad Luck Brian*, see Table 3(d), a teenage boy smiling in a portrait does not inherently convey tragedy, or misfortune, but this connotation stemmed from the template’s usage in online discourse.

In contrast, the unimodal images in Table 3 show a visual language that is less culturally specific, i.e. a serene beach has positive connotations and a disfigured *zombie-esque* head conveys negative ones. The cultural specificity of visual symbols in memes likely contributed to Image-STILT’s lack of significant performance improvement. These may explain similar observations by Suryawanshi et al. (2020), as discussed in Section 2.2, and would suggest that the transfer of visual sentiment *skills* from unimodal images to multimodal meme classifiers may be inherently limited.

²<https://knowyourmeme.com/memes/success-kid-i-hate-sandcastles>. Accessed: 11 Jun 2023.

³<https://knowyourmeme.com/memes/bad-luck-brian>. Accessed: 11 Jun 2023.

4.2 RQ2: Limited Labelled Memes

We found that Text-STILT significantly improves performance over Baseline across varying amounts of labelled meme availability between 50% and 80% (shown in Figure 4b). Within this range, while both intermediate training approaches consistently showed higher mean performance than Baseline, only Text-STILT showed a significant performance improvement and Image-STILT did not; p-values were 0.000109 for Baseline vs. Text-STILT and 0.0723 for Baseline vs. Image-STILT, respectively.

Based on these measurements, we found that Text-STILT was still able to outperform Baseline while using only 60% of the available labelled memes. Figure 5 shows the performance distribution of Baseline with 100% memes available and Text-STILT with 50% and 60% memes available.

We also noted that neither Image-STILT nor Text-STILT was found to significantly improve per-

	(a)	(b)
Sentiment Predicted	Negative	Negative
– Baseline	Positive	Neutral
– Text-STILT	Negative	Negative

Table 5: Example memes which were correctly labelled by Text-STILT but not by Baseline.

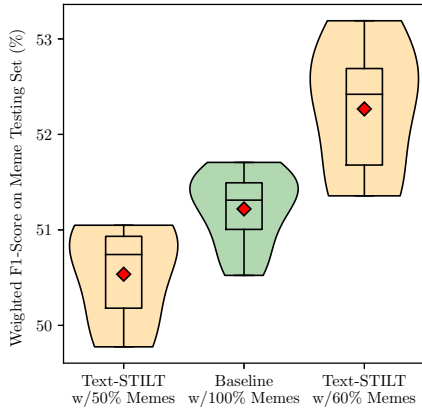


Figure 5: Performance of Baseline trained on 100% of memes available and Text-STILT trained on [50%, 60%] of memes available.

formance over Baseline across the entire range of availability of labelled memes from 5% to 80%. Figure 4 shows the mean performance and standard deviation of Baseline, Image-STILT, and Text-STILT across this range. When hypothesis testing is applied across the entire range, neither Image-STILT nor Text-STILT showed statistically significant improvements over Baseline, with p-values of 0.667 and 0.985, respectively.

Although Text-STILT performed better than Baseline, the difference is small. Contemporary approaches show similar small differences in performance (see Appendix D). Furthermore, 41% of memes in the testing set were not correctly classified by either Text-STILT and Baseline (see Appendix E). This suggests that a significant portion of memes remain a challenge to classify. This challenge might be addressed by combining Text-STILT with other supplementary training steps.

5 Limitations and Future Works

To generate comparable results between Baseline and Text-STILT, we kept many hyperparameters constant. Additional work would be required to determine the maximum achievable performance of Text-STILT on the chosen task.

Despite the efficacy of Text-STILT over Image-STILT, these results do not suggest that only the text modality is significant in classifying multimodal memes. Previous works have performed modality ablation studies in this problem space (Bucur et al., 2022; Pramanick et al., 2021b; Keswani et al., 2020) with multimodal architectures remain-

ing the apparent state of the practise. All models in this work are similarly multimodal. In the future, we plan to reformulate Image-STILT with respect to the approach and data used to isolate the cause of its non-performance on the downstream task. Furthermore, we did not test Text-STILT on classifiers that represent the image modality of a meme in textual forms, as others did (Singh et al., 2022; Pramanick et al., 2021b).

Notwithstanding our results, Text-STILT may not benefit all multimodal meme classifiers. Phang et al. (2019) showed that STILT offers varying degrees of benefit depending on the encoders chosen. Future work is needed to verify if these observations hold across the wide range of pretrained encoders commonly used in meme classifiers. In particular, some modifications to unimodal STILTs are needed to be applied to single-stream multimodal encoders, as those used in other works.

Furthermore, Pruksachatkun et al. (2020) showed that intermediate training benefits various text-only tasks differently. We have yet to identify other meme classification tasks that would benefit from unimodal STILTs. Thus, we plan to conduct more extensive experimentation to validate the effectiveness of Text-STILT on other meme classification tasks, e.g. pairing hate-speech detection in text (Toraman et al., 2022) as an intermediate task for hateful meme detection (Kiela et al., 2020).

6 Conclusion

In this work, we addressed the challenge of training multimodal meme sentiment classifiers on a limited number of labelled memes by incorporating unimodal sentiment analysis data. We did so by proposing the first instance of STILT that applies unimodal intermediate tasks to a multimodal target task. Specifically, we tested image-only and text-only sentiment classification as intermediate tasks in training a meme sentiment classifier. We showed that this approach worked – unimodal text improved meme classification performance to a statistically significant degree. This novel approach allowed us to train a meme classifier that outperforms meme-only finetuning with only 60% as many labelled meme samples. As possible explanations for our observations, we discuss apparent similarities and differences in the roles of image and text modalities between unimodal and multimodal sentiment analysis tasks.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224; and the provision of computational facilities and support from the Irish Centre for High-End Computing (ICHEC).

References

- Christian Bauckhage. 2011. [Insights into internet memes](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):42–49.
- Efrat Blaier, Itzik Malkiel, and Lior Wolf. 2021. [Caption enriched samples for improving hateful memes detection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9350–9358, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ana-Maria Bucur, Adrian Cosma, and Ioana Iordache. 2022. [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.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the surprising difficulty of natural yes/no questions](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.
- CrowdFlower. 2016. [Image sentiment polarity - dataset](#). Available at <https://data.world/crowdflower/image-sentiment-polarity>. Accessed: 2023-01-15.
- Baishan Duan and Yuesheng Zhu. 2022. [Browallia at memotion 2.0 2022 : Multimodal memotion analysis with modified ogb strategies \(short paper\)](#). In *De-Factify @ AAAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR Workshop Proceedings. AAAI.
- Noam Gal, Limor Shifman, and Zohar Kampf. 2016. [“it gets better”: Internet memes and the construction of collective identity](#). *New Media & Society*, 18(8):1698–1714.
- Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. [Hybrid Attention based Multimodal Network for Spoken Language Classification](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2379–2390, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Muzhaffar Hazman, Susan McKeever, and Josephine Griffith. 2023. [Meme sentiment analysis enhanced with multimodal spatial encoding and face embedding](#). In *Artificial Intelligence and Cognitive Science*, pages 318–331, Cham. Springer Nature Switzerland.
- Junguang Jiang, Yang Shu, Jianmin Wang, and Mingsheng Long. 2022. [Transferability in deep learning: A survey](#). *arXiv preprint*, arXiv:2201.05867.
- Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. 2020. [IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of Internet memes](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1135–1140, Barcelona (online). International Committee for Computational Linguistics.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, et al. 2020. [The hateful memes challenge: Detecting hate speech in multimodal memes](#). *Advances in Neural Information Processing Systems*, 33:2611–2624.
- Liisi Laineste and Piret Voolaid. 2017. [Laughing across borders: Intertextuality of internet memes](#). *The European Journal of Humour Research*, 4(4):26–49.
- Gwang Gook Lee and Mingwei Shen. 2022. [Amazon pars at memotion 2.0 2022: Multi-modal multi-task learning for memotion 2.0 challenge](#). In *De-Factify @ AAAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR Workshop Proceedings. AAAI.
- Quinn McNemar. 1947. [Note on the sampling error of the difference between correlated proportions or percentages](#). *Psychometrika*, 12(2):153–157.
- Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinakotla, Asif Ekbal, and Srijan Kumar. 2023. [Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes](#). *arXiv preprint*, arXiv:2303.09892.
- Niklas Muennighoff. 2020. [Vilio: state-of-the-art visio-linguistic models applied to hateful memes](#). *arXiv preprint arXiv:2012.07788*.
- Thanh Van Nguyen, Nhat Truong Pham, Ngoc Duy Nguyen, Hai Nguyen, Long H. Nguyen, and Yong-Guk Kim. 2022. [Hcilab at memotion 2.0 2022: Analysis of sentiment, emotion and intensity of emotion classes from meme images using single and multi modalities \(short paper\)](#). In *De-Factify @ AAAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR Workshop Proceedings. AAAI.

- Asaf Nissenbaum and Limor Shifman. 2017. [Internet memes as contested cultural capital: The case of 4chan’s /b/ board](#). *New Media & Society*, 19(4):483–501.
- Parth Patwa, Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, et al. 2022. Findings of memotion 2: Sentiment and emotion analysis of memes. In *De-Factify @ AAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR Workshop Proceedings. AAI.
- Kim Ngan Phan, Guesang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. 2022. Little flower at memotion 2.0 2022 : Ensemble of multi-modal model using attention mechanism in memotion analysis (short paper). In *De-Factify @ AAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR Workshop Proceedings. AAI.
- Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. [Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks](#). *arXiv preprint*, arXiv:1811.01088.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. [WiC: the word-in-context dataset for evaluating context-sensitive meaning representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.
- Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. [What to pre-train on? Efficient intermediate task selection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. [DynaSent: A dynamic benchmark for sentiment analysis](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online. Association for Computational Linguistics.
- Shraman Pramanick, Md Shad Akhtar, and Tanmoy Chakraborty. 2021a. Exercise? i thought you said ‘extra fries’: Leveraging sentence demarcations and multi-hop attention for meme affect analysis. *Proceedings of the International AAI Conference on Web and Social Media*, 15(1):513–524.
- Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. [MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. [Intermediate-task transfer learning with pretrained language models: When and why does it work?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *De-Factify @ AAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR Workshop Proceedings. AAI.
- Elad Segev, Asaf Nissenbaum, Nathan Stolero, and Limor Shifman. 2015. [Families and Networks of Internet Memes: The Relationship Between Cohesiveness, Uniqueness, and Quiddity Concrete-ness](#). *Journal of Computer-Mediated Communication*, 20(4):417–433.
- Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021. [AOMD: An Analogy-Aware Approach to Offensive Meme Detection on Social Media](#). *Inf. Process. Manage.*, 58(5).
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, et al. 2020. [SemEval-2020 task 8: Memotion analysis-the visuo-lingual metaphor!](#) In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.
- Shivam Sharma, Mohd Khizir Siddiqui, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. [Domain-aware self-supervised pre-training for label-efficient meme analysis](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 792–805, Online only. Association for Computational Linguistics.
- Limor Shifman. 2014. [The cultural logic of photo-based meme genres](#). *Journal of Visual Culture*, 13(3):340–358.

- Pranaydeep Singh, Aaron Maladry, and Els Lefever. 2022. [Combining language models and linguistic information to label entities in memes](#). In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 35–42, Dublin, Ireland. Association for Computational Linguistics.
- Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. [A dataset for troll classification of TamilMemes](#). In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).
- Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. [Large-scale hate speech detection with cross-domain transfer](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.
- Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019. [Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.
- Ron Zhu. 2020. [Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution](#). *arXiv preprint*, arXiv:2012.08290.
- Yan Zhuang and Yanru Zhang. 2022. Yet at memotion 2.0 2022 : Hate speech detection combining bilstm and fully connected layers (short paper). In *De-Factify @ AAI 2022. First Workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR Workshop Proceedings. AAI.

A Hyperparameters and Settings

	Input	
	Memes	Unimodal
LR Scheduling	Cosine Annealing	
Loss	Negative Log-Likelihood	
Learning Rate	1.5e−5 to 5e−5	5e−6 to 1e−5
Max Epochs	40	60
Optimizer	AdamW	
Betas	[0.5, 0.9]	
Weight Decay	0.9	
AMSGrad	False	
Dropout Rate	0.2	
Early-Stopping (per Meme Validation set)	Min Loss	Max Wei. F1

Table 6: Hyperparameter values and settings used during model training by input type.

B Metric: Weighted F1-Score

The performance of our models are measured by Weighted F1-Score, inline with the reporting set by the authors of the Memotion 2.0 dataset (Patwa et al., 2022). The F1-Score is the harmonic mean of precision and recall, equally representing both. “Weighted” here denotes that the F1-score is first computed per-class and then averaged while weighted by the proportion of occurrences of each class in the ground truth labels. We compute this using PyTorch’s implementation `multiclass_f1_score`. Class-wise F1-scores, $F1_c$ where $c \in [1, 2, 3]$, are computed as:

$$\begin{aligned}
 precision_c &= \frac{TP_c}{TP_c + FP_c} \\
 recall_c &= \frac{TP_c}{TP_c + FN_c} \\
 F1_c &= 2 \times \frac{(precision_c \times recall_c)}{(precision_c + recall_c)}
 \end{aligned} \tag{1}$$

Where TP_c, FP_c, FN_c are the count of true positives, false positives and false negatives, respectively. The Weighted F1-score is computed as the weighted average of $F1_c$:

$$\begin{aligned}
 w_c &= \frac{N_c}{\sum_{c=0}^C N_c} \\
 F1 &= \frac{\sum_{c=0}^C w_c F1_c}{C}
 \end{aligned} \tag{2}$$

Where N_c is the number of samples with the ground truth label c in the testing set. The Weighted F1 is often used when the classes are imbalanced – the training, validation and testing sets of Memotion 2.0 show significant and varying class imbalance – as it takes into account the relative importance of each class. Note that this weighted averaging could result in an F1-score that is not between the Precision and Recall scores.

C Architectural Details

Our models are based on the Baseline model proposed by Hazman et al. (2023) and we similarly utilise the Image and Text Encoders from the pre-trained ViT-B/16 CLIP model to generate representations of each modality.

$$\begin{aligned}
 F_I &= ImageEncoder(Image) \\
 F_T &= TextEncoder(Text)
 \end{aligned} \tag{3}$$

Where each F_I and F_T is a 512-digit embedding of the image and text modalities, respectively, from CLIP’s embedding space that aligns images with their corresponding text captions (Radford et al., 2021).

For unimodal inputs, the encoder for the missing modality is fed a blank input, i.e. when finetuning on unimodal images, the text input is defined as a string containing no characters i.e. “”:

$$\begin{aligned}
 F_I &= ImageEncoder(Image) \\
 F_T &= TextEncoder(“”)
 \end{aligned} \tag{4}$$

Conversely, when finetuning on unimodal texts, the image input is defined as a $3 \times 224 \times 224$ matrix of zeros, or equivalently, JPEG file with all pixels set to black.

$$\begin{aligned}
 F_I &= ImageEncoder(O_{3 \times 224 \times 224}) \\
 F_T &= TextEncoder(Text)
 \end{aligned} \tag{5}$$

For each modality, we added dropout and normalisation:

$$\begin{aligned}
 f_I &= Norm(Dropout(F_I)) \\
 f_T &= Norm(Dropout(F_T))
 \end{aligned} \tag{6}$$

where $Norm()$ is PyTorch’s BatchNorm1D and $Dropout$ has a rate of 0.2. These modality representations f_I and f_T are then placed into an attentive fusion mechanism proposed by Gu et al. (2018) and used by Pramanick et al. (2021a; 2021b) and Hazman et al. (2023). The embedding representation for each modality is passed through four dense

layers of reducing sizes [256, 64, 8, 1], $Dense_i$ and $Dense_t$ for the image and text modalities, respectively. Then, softmax is applied on the output of each stack is to generate a weighted score for each modality. Per (Gu et al., 2018):

$$\begin{aligned} D_i &= Dense_i(f_I) \\ D_t &= Dense_t(f_T) \\ [s_i, s_t] &= softmax(W_f[D_i, D_t] + b_f) \\ S_i &= (1 + s_i) \\ S_t &= (1 + s_t) \\ F_{MM} &= tanh(W_r[S_i(f_I), S_t(f_T)] + b_r) \end{aligned} \quad (7)$$

We added a dropout and normalisation step onto the fused multimodal representation, F_{MM} :

$$f_{MM} = Norm(Dropout(F_{MM})) \quad (8)$$

The predicted logits of each class is given by passing f_{MM} a dense network of GeLU-activated layers with sizes [1024, 256, 3]:

$$\begin{aligned} X_{MM} &= tanh(W_{mm}(W_x(f_{MM}) + b_x) + b_{mm}) \\ logits &= W_l(X_{MM}) + b_l \end{aligned} \quad (9)$$

The model is fitted by minimising the mean multiclass Cross Entropy Loss per PyTorch’s definition:

$$\begin{aligned} l_n &= -w_{y_n} \log \frac{exp(x_{n,y_n})}{\sum_{i=1}^C exp(x_{n,c})} y_n \\ L &= \sum_{n=1}^N \frac{1}{\sum_{n=1}^N w_{y_n}} l_n \end{aligned} \quad (10)$$

Where x_{n,y_n} is the logits for each class and y_n is the target label of a given sample n of total N samples in a minibatch; c is the class in [Negative, Neutral and Positive] and C is the number of classes. The loss for each sample is weighted by:

$$w_{y_n} = 1 - \frac{N_{y_n}}{\sum_{y_n=0}^C N_{y_n}} \quad (11)$$

Where N_0, N_1, N_2 are the number of training samples labelled with Negative, Neutral and Positive sentiment, respectively.

D Performance Benchmarking

Current competing approaches show a small spread of Weighted F1-scores (see Table 7) and the performance improvement offered by Text-STILT is similarly small. This small range of performances in contemporary approaches suggests that there is still a significant portion of memes that remain a challenge to classify.

Solution	Weighted F1 (%)
Bucur et al. (2022)	53.18
Text-STILT w/ 60% (Max)	53.15
Duan and Zhu (2022)	52.55
Text-STILT w/ 60% (Mean)	52.45
Our Baseline (Max)	51.70
Our Baseline (Mean)	51.19
Zhuang and Zhang (2022)	50.88
Phan et al. (2022)	50.81
Greeny (via Patwa et al., 2022)	50.37
Hazman et al. (2023)	50.35
Lee and Shen (2022)	50.25
Nguyen et al. (2022)	49.95

Table 7: The mean and maximum Weighted F1-scores from our Baseline and Text-STILT approaches against various SOTA solutions.

E Contingency Table: Baseline vs. Text-STILT

		Baseline	
		Correct	Wrong
Text-STILT	Correct	610	146
	Wrong	136	608

Table 8: Contingency Table between similarly performing Text-STILT (trained with 60% memes) and Baseline (trained with 100% memes).

Table 8 shows the contingency table – as one would prepare for a McNemar’s Test between two classifiers (McNemar, 1947) – between the model trained with Text-STILT on 60% Memes and Baseline trained on 100% Memes available which had the most similar performance. While the two models performed similarly in terms of Weighted F1-scores, Text-STILT correctly classified a notable number of memes that Baseline did not and vice versa. Examples of such memes are discussed in Section 4.1. Furthermore, approximately 40% of memes in the testing set were incorrectly classified by both models. This suggests that these memes convey sentiment in a way that cannot be reliably predicted by either approach.