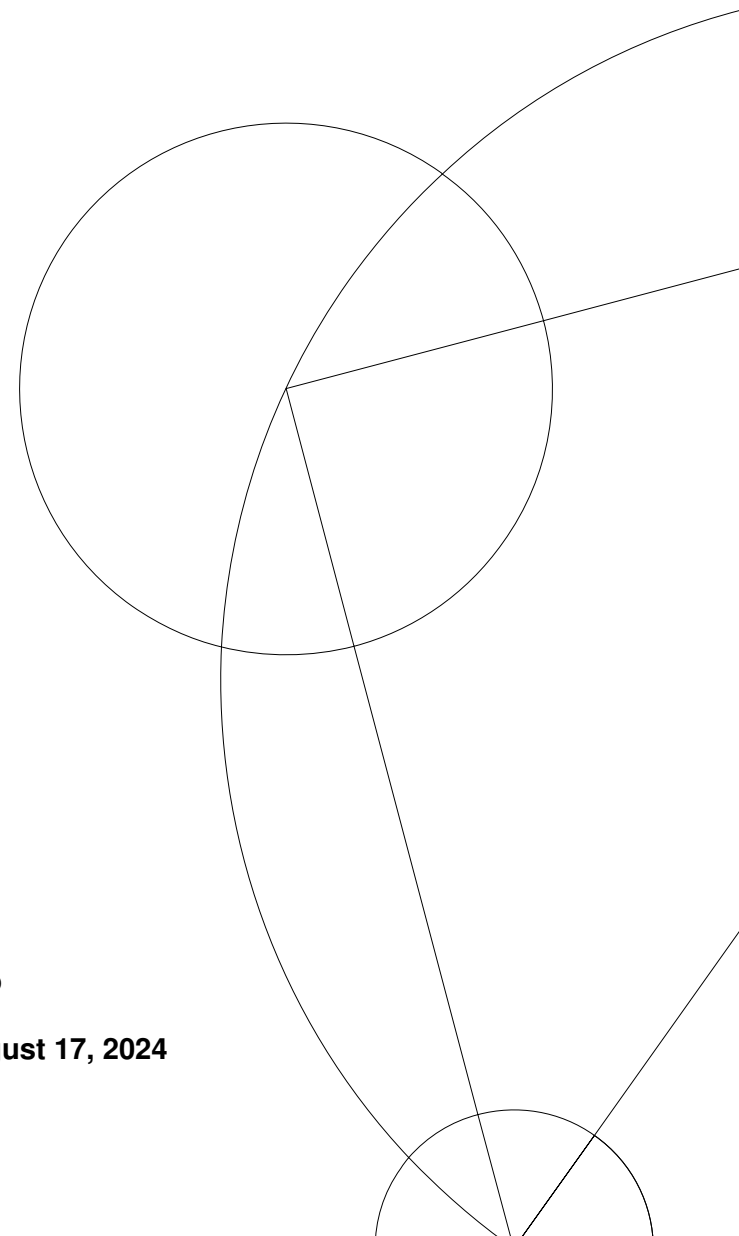
Computational Cognitive Science II

MSc in IT & Cognition, 2023

From Images to Memes: Assessing the Transferability of Cross-Modal Embedders in Meme-Text Scenario



Lecturers: Costanza Navarretta and Patrizia Paggio

Department of Nordic Studies and Linguistics. August 17, 2024

Group members

Name	KUid	Contributions
Shiling Deng	bjc154	zero-shot evaluation (ALIGN and CLIP), case study (coding), report (related work, results, discussion, and conclusion)
Yuwei Shen	hwd369	zero-shot evaluation (ALBEF), case study (analyzing), report (introduction, results, and discussion)
Qing Li	nql538	fine-tune and evaluation (ALIGN), case study (analyzing) report (abstract, methodology, results, and discussion)

1 Abstract

Internet memes, blending images and text through visual metaphors, present a unique challenge for AI systems trained on standard image-text data. This study investigates the generalizability of cross-modal embedding models to the metaphor-rich context of memes. We hypothesize that models pretrained on general image-text pairs underperform on meme-text data due to the abstract nature of meme communication. Using the MEMECAP dataset, which includes 6,384 annotated memes, we evaluate models like ALIGN, CLIP, and ALBEF on tasks of meme-text and text-meme retrieval. Our extensive experiments reveal that these models do not perform exceptionally well on the meme data, validating our hypothesis. These findings underscore the limitations of current multi-modal models in capturing abstract metaphorical associations, positioning meme understanding as a critical test-bed for assessing advanced multi-modal reasoning in AI. This study motivates the development of more robust multi-modal representations to improve metaphor comprehension in AI systems.

Contents

2 Introduction

With the rapid development of the internet, the volume of multimedia data, including images, text, audio, and video, has surged significantly. These diverse encoding types that convey identical semantics are referred to as multimodal data (Barsalou, 2008; Cassell, 2001).

Unlike simple images or texts, memes blend visual and textual elements to create context-specific meanings, making them a vital digital communication tool (Hwang & Shwartz, 2023). This characteristic has rendered memes ubiquitous in online communication, significantly shaping social media interactions since the late 2000s. On average, users encounter numerous memes daily, employing them to convey complex ideas and emotions succinctly (Vyalla & Udandarao, 2020).

Typically, a meme is a photo or video that humorously portrays a concept or idea, often accompanied by text that uses metaphor, sarcasm, irony, or absurdity (Mishra et al., 2023). For example, as shown in Figure 1, a meme might compare teammates to garbage cans, suggesting they are poor players despite having fancy skins. This meme conveys complex emotions such as anger, love, and conflict feelings, which are difficult to express accurately through words alone.

Figure 1: A video game related meme

Over the last decade, significant research has been conducted on memes. In 2020, Facebook launched a meme hateful speech detection competition (Kiela et al., 2020). Additionally, the Memotion task was introduced, requiring researchers to analyze the sentiment, emotions, and semantic classes of memes. Subsequently, Memotion 2 and Memotion 3 tasks were released, focusing on similar analyses but utilizing different datasets (Bucur, Cosma, & Iordache, 2022; C. Sharma et al., 2020; Mishra et al., 2023).

Furthermore, studies such as (S. Sharma, Ramaneswaran, Akhtar, & Chakraborty, 2024; Zhou, Jurgens, & Bamman, 2023) have examined the emotions and sociolinguistic variations in memes, revealing underlying patterns in meme usage and classification.

With the availability of millions of meme instances and templates from Imgflip[1] and Know Your Meme[2], meme generation and captioning have become feasible, as demonstrated by the work of (Peirson V & Tolunay, 2018; Hwang & Shwartz, 2023).

Most of these studies have utilized multimodal embedders, such as CLIP (Radford et al., 2021), to generate representations of memes and related text. However, none have specifically examined

[1] https://imgflip.com/
[2] https://knowyourmeme.com/

how well a pre-trained cross-modal embedder performs in a meme-text scenario. These models are typically pre-trained on general image-text datasets, whereas memes often convey messages through metaphor, humor, and satire. This complexity requires cross-modal embedding algorithms to not only understand visual content and text but also to comprehend how visual and textual cues are combined to express specific social and cultural meanings. Therefore, studying the generalizability of these models is both intriguing and significant. Despite achieving substantial success on traditional datasets, their utility for memes remains challenging.

To address this gap in research, we will apply recently developed deep learning models, specifically ALIGN (A Large-scale ImaGe and Noisy-text embedding) (Jia et al., 2021), CLIP (Contrastive Language-Image Pre-training), and ALBEF (ALign the image and text representations BEfore Fusing) (Li et al., 2021), to meme-text retrieval tasks. We will evaluate their effectiveness and performance using the Recall@K metric. Among various meme datasets, we have chosen MemeCap (Hwang & Shwartz, 2023), a dataset designed for captioning and explaining internet memes. The meme captions in this dataset describe the actual meanings of the memes, providing potential perspectives for understanding them.

Due to the complexity of understanding visual metaphors and the lack of contextual information in memes, we hypothesize that retrieval results on the meme-text dataset will be inferior to those on traditional image-text datasets. Our code is available on GitHub [3].

In Section 3, we review related research on image-text retrieval and memes in greater detail. Section 4 describes the data and models used in our experiments. Section 5 presents the results and case studies of meme-text retrieval. Finally, we conclude our findings in Section 6.

Our results indicate that ALIGN and CLIP can generalize well to meme-text data, whereas ALBEF experiences a significant performance drop compared to its performance on general image-text data. Additionally, while ALIGN is effective at extracting information about visual objects and text, it struggles with understanding visual metaphors and sarcasm. Our fine-tuned model shows no performance improvement, likely due to non-optimal hyperparameters or model structure constraints.

3 Related Work

3.1 Meme analysis

In study (Kiela et al., 2020), they launched the meme hateful speech detection task, formulated as a classification problem. Memes were labeled with a hateful degree on a scale from 1 to 3, where 1 indicates definitely hateful, 2 indicates uncertain, and 3 indicates definitely not hateful. The baseline models included unimodal approaches, such as BERT (Devlin, Chang, Lee, & Toutanova, 2018), and multi-modal approaches, such as ViLBERT (Lu, Batra, Parikh, & Lee, 2019). Their baseline models reveal that even with state-of-the-art models, understanding hatefulness remains challenging. However, this work sheds light on handling memes with deep-learning models, paving the way for future research.

The series of Memotion tasks (Bucur et al., 2022; C. Sharma et al., 2020; Mishra et al., 2023) even took a step further to study the emotions and semantical meanings conveyed by memes. Three sub-tasks are released:

[3] https://github.com/Seefreem/meme_text_retrieval

- **Task A: Sentiment Analysis:** Classify a meme into positive, negative or neutral.
- **Task B: Emotion Classification:** Classify a meme into humor, sarcasm, offensive, etc.
- **Task C: Scales/Intensity of Emotion Classes:** Quantify the magnitude of emotions.

They proposed a Multi-Modal-Multi-Task Transformer (MMMT) for these tasks, achieving an F1-score of 0.8111 for humor detection and 0.8191 for sarcasm detection. These promising results suggest that, although it is challenging to fully comprehend the exact meaning of a meme, it is still feasible to classify memes into basic emotional categories. This enables models to align and understand memes and text through the labels of these categories. Consequently, cross-modal embedders are expected to align memes and related text based on these basic emotions or semantics.

3.2 Meme generation

Earlier, (Peirson V & Tolunay, 2018) introduced a novel meme caption-generating system. Their approach first utilizes a CNN model to generate image representations. These embeddings are then inputted into various LSTM models for meme caption generation. They suggested that captioning could be a method for understanding meme content. Their results show that synthetic meme captions have a 70% differentiation rate from human-written captions, indicating that the LSTM models do not perform excellently on this task. Nevertheless, their research demonstrated the potential for deep learning models to understand memes and encode them into embeddings.

3.3 Image-text cross embedding

Although there is no direct meme-text embedder, numerous image-text embedding models are available (Cao, Li, Li, Nie, & Zhang, 2022; Xia, Yang, Ge, & Yin, 2024). Memes, being a type of image, share subsets of features with images, indicating that cross-modal embedders trained on general datasets are applicable to our task (Saakyan, Kulkarni, Chakrabarty, & Muresan, 2024).

Instead of training the image and text embedders on supervised tasks with selected labels, CLIP is trained on image-text pairs using a contrastive learning objective. In the original paper, the authors primarily experimented with two image embedders: ResNet-50 (He, Zhang, Ren, & Sun, 2016) and Vision Transformer (ViT) (Dosovitskiy et al., 2020). For text, they also used a transformer architecture. Ultimately, their model achieved excellent zero-shot accuracy on many downstream tasks, suggesting it may also perform well on the meme-text dataset.

Contrastive learning is a self-supervised learning method where data samples are categorized into positive and negative samples. The objective function minimizes the distances between positive samples and maximizes the distances between negative samples (Balestriero & LeCun, 2022). In the case of image data augmentation, each raw image is considered a category, with its flipped, cropped, and scaled versions taken as positive samples, and all other images as negative samples. In CLIP, given an image, its paired text is the positive sample, and the objective function maximizes the similarity between the image embedding and the text embedding, while treating all other texts in the batch as negative samples.

ALIGN is a dual-encoder architecture designed to learn aligned visual and language representations from large-scale, noisy image-text pair data using a contrastive loss. As shown in Figure 2, text and image are embedded separately. For text, they applied a WordPiece BERT with an input length of 64 tokens, determined by their training data, although our dataset may require longer inputs. For images, they used EfficientNet (Tan & Le, 2019), a model that is smaller and faster in inference compared to previous state-of-the-art CNNs, achieving 91.7% accuracy on CIFAR-100. Ultimately,

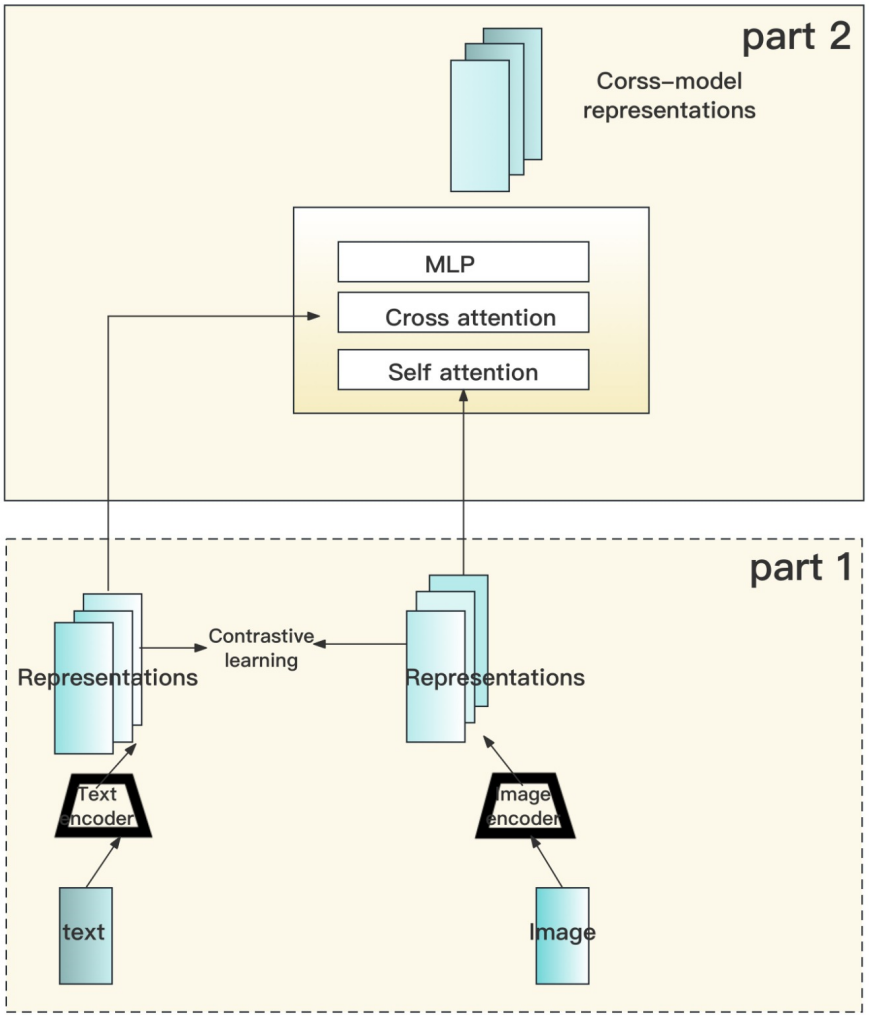


Figure 2: Part 1 is a common multi-modal embedder architecture, e.g. ALIGN and CLIP; Part 2 is a cross-modal attention block, e.g. ALBEF.

they utilized two loss functions to combine the two submodules for contrastive learning:
Image to text classification loss:

$$L_{i2t} = -\frac{1}{N}\sum_{i}^{N}\log\frac{exp(x_i^T y_i/\sigma)}{\sum_{j=1}^{N} exp(x_i^T y_j/\sigma)}$$

Text to image classification loss:

$$L_{t2i} = -\frac{1}{N}\sum_{i}^{N}\log\frac{exp(y_i^T x_i/\sigma)}{\sum_{j=1}^{N} exp(y_i^T x_j/\sigma)}$$

Here, x_i and y_j are the L2-normalized image embedding and text embedding, in the ith pair and the jth pair, respectively. N represents the batch size. A larger N allows the model to compare a positive sample to a wider range of negative samples, enhancing the model's learning capability. The σ here is the temperature that adjusts the classification distribution for better performance.

They wrapped the similarities between images and text into log probabilities. By minimizing the classification losses, they formulated contrastive learning into two classification tasks.

While both CLIP and ALIGN achieve state-of-the-art performance on their respective datasets and are widely used in downstream tasks, they may struggle to capture the nuanced semantics and contextual factors present in internet memes. Both models utilize global alignment methods. Successful image-text retrieval with a global alignment approach typically requires embedding vectors to contain comprehensive fine-grained features, which is challenging to achieve. In contrast, local alignment methods can align image patches to words in a text, enabling the cross-embedding model to focus more effectively on the most important elements of both modalities (Cao et al., 2022). Therefore, we also evaluated a model that incorporates local alignment, namely **ALBEF** (Li et al., 2021), illustrated in Figure 2.

In addition to the separate text and image encoders, ALBEF includes a cross-modal attention block, which consists of the upper six layers of the BERT model. This additional block aligns words with image patches. To optimize the system, three objective functions are applied: contrastive loss, masked language loss, and image-text matching loss. For the image-text matching task, the representation of [CLS] is used as the joint representation of the image-text pair and is input into a classification layer to predict whether the image and text match. The local alignment serves as a key component matching method, allowing the model to focus on information shared by both modalities while ignoring irrelevant content. This mirrors how humans often understand memes—by focusing on the central elements, such as facial expressions, and disregarding the background.

Given the significant success of these models, we would expect them to perform well on meme-text data. However, they may fall short due to the abstract and nuanced meanings often present in memes.

4 Methodology

Similar to the image-text retrieval task, we use the meme-text retrieval task to evaluate these cross-modal embedders. If the distance between the representations of a meme-text pair is small, the corresponding meme or text should rank high in the resulting list during retrieval. Therefore, a higher Recall@K indicates better embedding performance.

4.1 Memecap dataset

In the image-text retrieval task, samples are formulated into image-text pairs, allowing the model to match them. Similarly, we chose a meme dataset that includes both memes and their related captions. By meme caption, we refer to the text that interprets the meaning of the meme, not a direct description of the meme's content, but rather what the meme poster intends to convey.

This dataset contains approximately 6.3K memes [4], including titles, meme captions, textual image descriptions, and visual metaphors. Statistical information is shown in Table 1. As illustrated in Figure 3, the meme caption and the image caption can be entirely different. The only common elements in this example are the words "student" and the "graduation cap." Understanding the meme also requires interpreting the embedded text on the meme image.

[4] https://github.com/eujhwang/meme-cap

Splits	#Memes	#M-Cap	#I-Cap
Train+Val	5,828	5,828	5,828
Test	559	2,036	599

Table 1: The number of memes; meme captions and image captions.

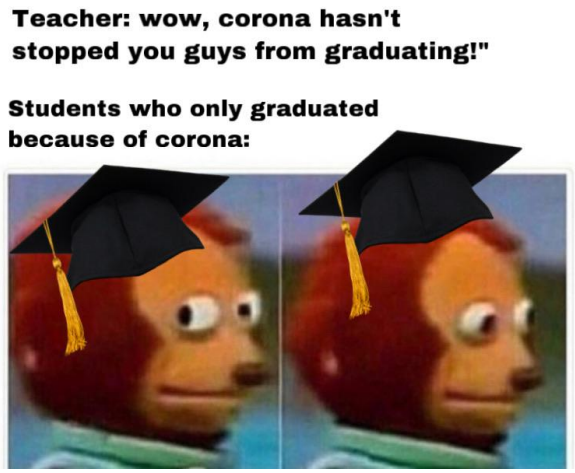


Figure 3: **Title**: Thanks to covid!! **Meme caption**: Meme poster thinks students are aware they are unprepared because of covid; **Image caption**: A muppet is looking side-eyed in a graduation cap.

4.2 Recall@K

Originally, in a retrieval system, Recall-at-K (R@K) measures the ratio of relevant items in the top-K results to the overall number of relevant items. In our experiments, each image has only one related text, and vice versa. Consequently, R@K will be either 1 or 0. Therefore, we measure the average R@K across the entire test set.

4.3 Experiment setup

We used the open-source pre-trained ALIGN and CLIP models from Huggingface [5] [6], and downloaded and integrated the pre-trained ALBEF model from a GitHub repository [7]. For each experiment, we input all the memes and corresponding captions into the models to obtain image and text embeddings. We then calculated their cosine similarity matrix. By applying softmax, we obtained a normalized matrix, referred to as the classification matrix (consistent with contrastive learning), where each element represents a classification score.

First, we evaluated them in a zero-shot setting on the Memecap test set to verify the effectiveness of these models. We then calculated the R@1, R@5, R@10, and the mean of these three metrics for both meme-text retrieval and text-meme retrieval sub-tasks.

Furthermore, we conducted case studies to investigate why the models perform excellently on some samples but poorly on others. This study aims to verify our hypothesis that due to the complexity of understanding metaphors, sarcasm, irony, or absurdity, these cross-modal embedders struggle to achieve high R@K scores.

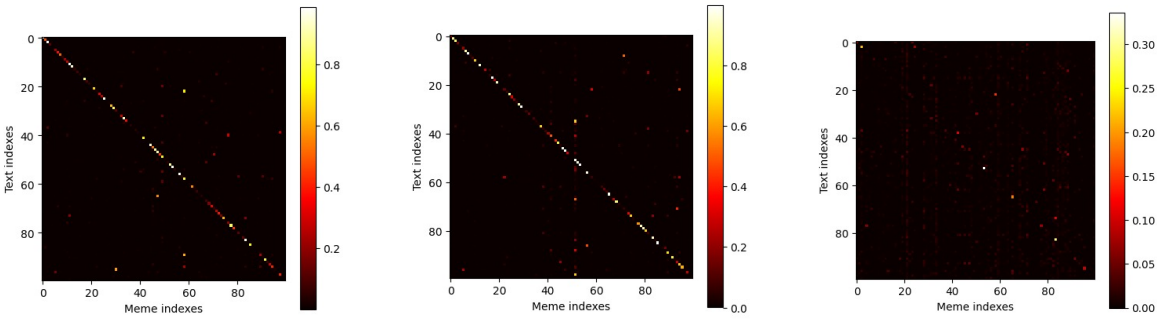
Lastly, using the data-streaming method, we fine-tuned ALIGN on the training set, which includes

[5] https://huggingface.co/docs/transformers/en/model_doc/align
[6] https://huggingface.co/docs/transformers/en/model_doc/clip
[7] https://github.com/salesforce/ALBEF

4,658 meme-text pairs, with an additional 1,165 samples in the development set. We conducted a hyper-parameter sweep in a full fine-tuning setup, with epochs ranging from 1 to 2. The learning rate was sampled from a uniform distribution in the range of 10^{-6} to $4 * 10^{-5}$. The goal of sweeping is minimizing the loss. Due to limited GPU memory, we set the batch size to 6, which may be too small to enable the model to fully learn the essence of memes.

5 Results and Discussion

5.1 Zero-shot evaluation



(a) (b) (c)

Figure 4: The hot map of text-meme classification matrix of ALIGN (a), CLIP (b), and ALBEF (c)

For zero-shot evaluation, we first visualized the text-meme classification matrix of the three models on the first 100 samples. As shown in Figure 4, ALIGN and CLIP display mostly bright diagonal pixels, suggesting that most of the text-meme pairs have similar embeddings. In contrast, the ALBEF model struggles to match texts with the corresponding memes.

Model	T2M R@1	T2M R@5	T2M R@10	T2M mean	M2T R@1	M2T R@5	M2T R@10	M2T mean	Overall mean
ALIGN	**0.539**	**0.726**	**0.774**	**0.680**	**0.568**	**0.742**	**0.789**	**0.700**	**0.690**
CLIP	0.457	0.665	0.733	0.618	0.518	0.720	0.772	0.670	0.644
ALBEF	0.120	0.200	0.254	0.191	0.093	0.157	0.211	0.154	0.173
fine-tuned ALIGN	0.512	0.680	0.758	0.650	0.490	0.687	0.739	0.639	0.644

Table 2: The R@k score of meme-to-text retrieval and text-to-meme retrieval on Memecap on a scale of $0 - 1$.

Next, we evaluated the three models on the entire test set. As shown in Table 2, ALIGN achieves the highest scores on all metrics, reaching an R@10 of 0.774 on the text-meme retrieval task and 0.789 on the meme-text retrieval task. Interestingly, the fine-tuned version of ALIGN performs slightly worse than the zero-shot setting, possibly due to the small batch size of 6 used during fine-tuning. Despite this, the fine-tuned ALIGN still outperforms CLIP. CLIP ranks third, scoring 0.733 R@10 on the text-meme retrieval task and 0.67 R@10 on the meme-text retrieval task. ALBEF, however, shows significantly lower scores on these metrics.

This discrepancy could be attributed to the model size and the size of its training datasets. ALBEF has 12 layers for image encoding, 6 layers for text encoding, and 6 layers for cross-modal attention.

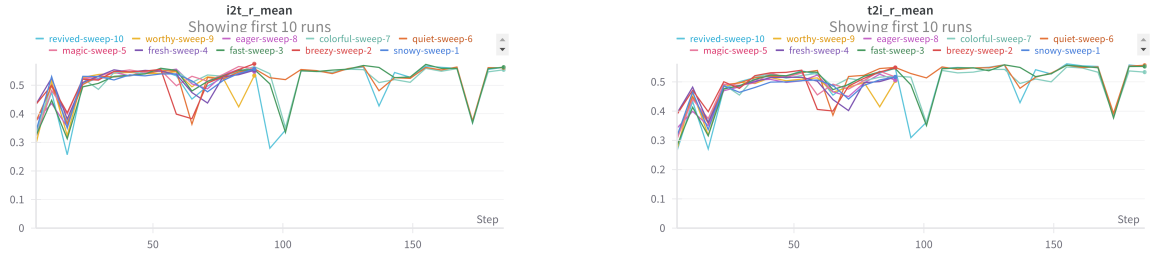
In contrast, ALIGN and CLIP are deeper and larger. ALBEF was trained on 14.1 million samples, whereas CLIP was trained on 400 million image-text pairs, and ALIGN on 1.8 billion image-text pairs. These results suggest that both ALIGN and CLIP perform well on the meme-text dataset.

Dataset	T2I R@1	T2I R@5	T2I R@10	I2T R@1	I2T R@5	I2T R@10
Flickr30K (1K test set)	0.757	0.938	0.968	0.886	0.987	0.997
Memecap (599 test set)	0.539	0.726	0.774	0.568	0.742	0.789
MSCOCO (5K test set)	0.456	0.698	0.786	0.586	0.830	0.897

Table 3: The zero-shot R@k score of ALIGN on three different datasets on a scale of $0-1$.

Table 3 presents the zero-shot R@K scores of ALIGN on three datasets. For each dataset, only the test sets were used. The model performs best on Flickr30K (Plummer et al., 2015), followed by Memecap, and performs weakest on MSCOCO (Chen et al., 2015; Lin et al., 2014). The performance difference between Flickr30K and MSCOCO may be due to the greater number of categories in MSCOCO, suggesting that MSCOCO is a more challenging benchmark. Compared to MSCOCO, ALIGN achieves relatively higher scores on Memecap.

5.2 Fine-tuning



(a) (b)

Figure 5: The sweep results

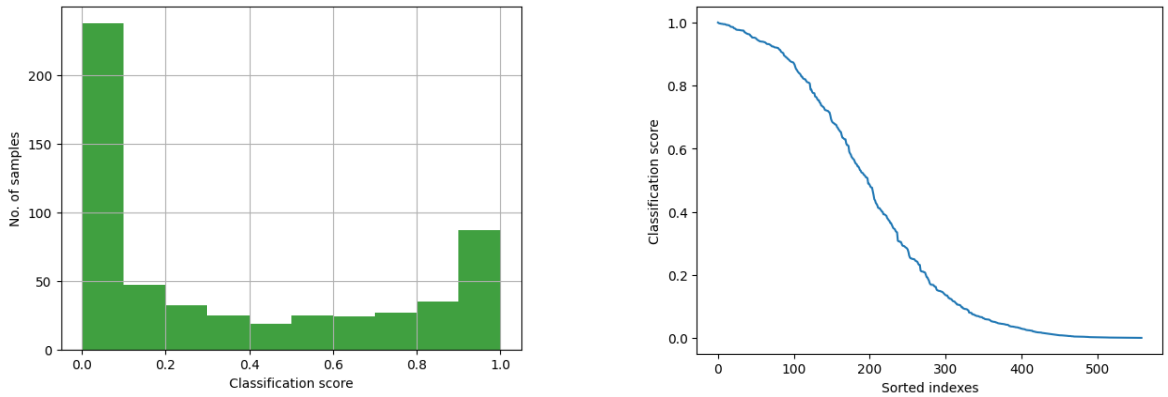
Finally, we conducted a hyper-parameter sweep experiment to find an optimal model for the Memecap dataset. As shown in Figure 5, regardless of changes in the learning rate and the number of epochs, the average R@K scores for both tasks remained below 0.6, indicating an upper limit.

Overall, the cross-modal embedding models trained on general image-text datasets can generalize to the meme-text dataset and achieve good R@K scores. The final R@K scores vary according to the dataset. However, the potential for improvement through fine-tuning was not observed in this study.

5.3 Case study

In this section, we manually examined samples that ALIGN can perfectly match (with text-to-meme classification scores greater than 0.9 and nearly 1.0) and samples where the model performs poorly.

To better select examples, we first analyzed the statistical features of the classification scores on the test set, as shown in Figure 6. Over 200 memes have near-zero scores for their corresponding captions, while approximately 90 memes are highly aligned with their captions. Based on this, we



(a) (b)

Figure 6: (a) The histogram of classification scores; (b) The descending sorted classification scores.

selected samples from both the top and bottom of the sorted scores to investigate why the ALIGN model performs excellently on some samples and poorly on others.

The factors that influence the performance of the ALIGN model in the meme-text matching task can be summarized as follows based on these meme samples:

1. Visual object recognition capability - the model can effectively align textual keywords with specific visual objects in the image.

2. Embedded text recognition ability - the model can extract textual information from images and align it with captions.

3. Metaphor understanding capability - some examples suggest that the model may possess a certain level of metaphor understanding. However, there are also counter-examples.

4. Difficulty in understanding humor and sarcasm - the model performs poorly in comprehending memes with humor or sarcasm.

5. Insufficient understanding of complex semantics - it has limitations in understanding more abstract and complex semantics such as metaphors and humor.

As shown in Figure 7, the model matches the memes and their captions with classification scores higher than 0.99, indicating that the model identifies identical information in both textual and visual modalities. For example, the keyword ”apple” is verified by the four apples on the meme, and the keywords ”feet” and ”fire” are verified by the foot and fire on the meme. This result meets our basic expectation that ALIGN can align textual words with visual objects, suggesting its feasibility for application to the meme-text dataset.

For memes that have no direct visual counterpart to the keywords in their captions, ALIGN also appears capable of aligning them. As shown in Figure 8, with roads referring to political actions, the car referring to NATO, and Patrick Star as the poster, these pairs have classification scores higher than 0.99. Both memes belong to the metaphor category, suggesting that ALIGN might have some understanding of metaphors.

However, the samples in Figure 9, which are also in the metaphor category, have classification scores lower than 0.0001, with ALIGN mismatching the memes to incorrect captions. In Figure 9 (a), the train represents the poster, and the two trains form a metaphorical construction between the poster's expression and mental state. While ALIGN may recognize the word ”song,” it fails to comprehend the entire embedded text and thus cannot connect the meme to its caption. In Figure 9

(a) (b)

Figure 7: Well-matched meme-text examples: (a) meme caption: Eye doctors expect you to see the difference between identical pictures of the apple; (b) meme caption: Meme poster is trying to sleep but his feet feeling like they are on fire keeps him away.

(b), the poster refers to their childhood as a burning swing, suggesting a traumatic childhood. ALIGN does not connect the burning swing to the concept of a horrible childhood, indicating a limitation in understanding visual metaphors.

We observed that the key difference between Figure 8 and Figure 9 is the presence of embedded text on the memes. In Figure 8, the words from the captions are clearly visible on the memes, which is not the case in Figure 9. Therefore, we suggest that ALIGN can recognize visual objects and embedded texts on an image, allowing it to match memes that lack visual clues but contain relevant words in the embedded text. However, ALIGN struggles to understand visual metaphors.

The samples in Figure 10 further support our conclusion. The caption mentions "Elon Musk," but there is no visual clue about Elon Musk in either meme. The key difference is that the meme on the right includes the text "Elon Musk." ALIGN assigned a classification score of 0.0000125 to the ground truth meme and 0.298 to the false positive meme, incorrectly matching the caption with the wrong meme.

We also found that the model struggles to understand humor and sarcasm. For example, in Figure 11 (a), the poster refers to their dog as a fake barking deer. The dog is mistakenly taken for a deer because of the twig on its head, but the twist is that the "deer" barks like a dog, showcasing ridiculous humor about mistaken identity. However, the model does not recognize this twist or the fake deer, resulting in a low classification score of 0.0000501. In Figure 11 (b), the poster uses sarcasm to express their frustration with Netflix's commercial strategy by using a coffin to symbolize their reaction to Netflix. The classification score for this meme is 0.000255. In both examples, ALIGN failed to correctly match the memes to their captions.

(a)

(b)

Figure 8: Well-matched meme-text examples: (a) meme caption: The meme poster is mocking NATO for imposing new sanctions rather than Article 5, after the missile attack; (b)meme caption: Meme poster can handle small decisions like choosing shower music but screws up the important ones.

(a)

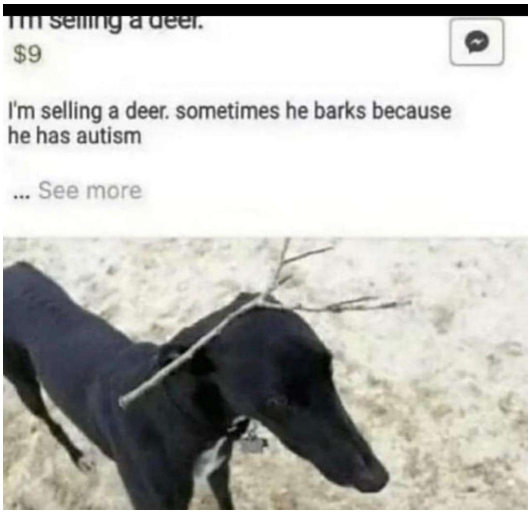
(b)

Figure 9: Mismatched meme-text examples: (a) meme caption: The meme poster is expressing its craze about a song and how it is hiding its reaction from people; (b) meme caption: Meme poster is conveying how horrible their childhood was.

(a)

(b)

Figure 10: Mismatched meme-text example: meme caption: Meme poster doesn't understand how Elon Musk is the way he is. (a) the target meme (classified as False Negative); (b) the retrieved meme (classified as False Positive).



(a)

(b)

Figure 11: (a) meme caption: Meme poster is trying to convey that false thing should happen in different ways; (b) meme caption: After Netflix announces it'll have commercials, users will cancel it.

6 Conclusion

Our findings indicate that ALIGN, a leading cross-modal embedder, can recognize objects and text within images and match them to their captions. However, it lacks the ability to grasp deeper semantic meaning, including metaphors, sarcasm, and humor. Furthermore, ALIGN doesn't seem to capture emotions from text and align them with visuals. This aligns with our hypothesis: general-purpose image-text datasets hinder cross-modal embedders in accurately matching memes to captions, resulting in lower retrieval task performance.

Beyond model limitations, understanding memes goes beyond simple text-image matching. It necessitates comprehending the meme's context and underlying information about the image itself. In some instances, images solely convey emotions, not visual objects, and can be adapted to various contexts. Therefore, accurately identifying the corresponding meme for a given caption remains a significant challenge.

These results pave the way for future research directions: 1) Visual Reasoning: Develop models capable of understanding and reasoning about visual semantics, including metaphors, sarcasm, and humor. 2) Textual Understanding: Create models that not only extract text but also reason based on the extracted content. 3) Enhanced Alignment: Expand alignment beyond object or word recognition to encompass metaphors, sarcasm, and humor.

References

Balestriero, R., & LeCun, Y. (2022). Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, *35*, 26671–26685.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.

Bucur, A.-M., Cosma, A., & Iordache, I.-B. (2022). Blue at memotion 2.0 2022: You have my image, my text and my transformer. *arXiv preprint arXiv:2202.07543*.

Cao, M., Li, S., Li, J., Nie, L., & Zhang, M. (2022). Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

Cassell, J. (2001). Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine*, *22*(4), 67–67.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hwang, E., & Shwartz, V. (2023). Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., ... Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (pp. 4904–4916).

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, *33*, 2611–2624.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, *34*, 9694–9705.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13* (pp. 740–755).

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*.

Mishra, S., Suryavardan, S., Patwa, P., Chakraborty, M., Rani, A., Reganti, A., ... others (2023). Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. *arXiv preprint arXiv:2303.09892*.

Peirson V, A. L., & Tolunay, E. M. (2018). Dank learning: Generating memes using deep neural networks. *arXiv preprint arXiv:1806.04510*.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the ieee international conference on computer vision* (pp. 2641–2649).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Saakyan, A., Kulkarni, S., Chakrabarty, T., & Muresan, S. (2024). V-flute: Visual figurative language understanding with textual explanations. *arXiv preprint arXiv:2405.01474*.

Sharma, C., Bhageria, D., Scott, W., Pykl, S., Das, A., Chakraborty, T., ... Gamback, B. (2020). Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*.

Sharma, S., Ramaneswaran, S., Akhtar, M. S., & Chakraborty, T. (2024). Emotion-aware multimodal fusion for meme emotion detection. *IEEE Transactions on Affective Computing*.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).

Vyalla, S. R., & Udandarao, V. (2020). Memeify: A large-scale meme generation system. In *Proceedings of the 7th acm ikdd cods and 25th comad* (pp. 307–311).

Xia, B., Yang, R., Ge, Y., & Yin, J. (2024). A review of cross-modal retrieval for image-text. In *Fifteenth international conference on graphics and image processing (icgip 2023)* (Vol. 13089, pp. 389–400).

Zhou, N., Jurgens, D., & Bamman, D. (2023). Social meme-ing: Measuring linguistic variation in memes. *arXiv preprint arXiv:2311.09130*.