AraCLIP: Cross-Lingual Learning for Effective Arabic Image Retrieval

Muhammad Al-Barham[1], Imad Afyouni[2], Khalid Almubarak[3],
Ashraf Elnagar[2], Ayad Turky[2], Ibrahim Abaker Hashem[2]
[1]RISE, University of Sharjah, United Arab Emirates
[2]Computer Science Department, University of Sharjah, United Arab Emirates
[3]Computer Engineering Department, PSAU, Saudi Arabia
malbarham@sharjah.ac.ae, iafyouni@sharjah.ac.ae, k.almubarak@psau.edu.sa,
{ashraf,aturky,ihashem}@sharjah.ac.ae

Abstract

This paper introduces Arabic Contrastive Language-Image Pre-training (AraCLIP), a model designed for Arabic image retrieval tasks, building upon the Contrastive Language-Image Pre-training (CLIP) architecture. AraCLIP leverages Knowledge Distillation to transfer cross-modal knowledge from English to Arabic, enhancing its ability to understand Arabic text and retrieve relevant images. Unlike existing multilingual models, AraCLIP is uniquely positioned to understand the intricacies of the Arabic language, including specific terms, cultural nuances, and contextual constructs. By leveraging the CLIP architecture as our foundation, we introduce a novel approach that seamlessly integrates textual and visual modalities, enabling AraCLIP to effectively retrieve images based on Arabic textual queries. We offer an online demonstration allowing users to input Arabic prompts and compare AraCLIP's performance with state-of-the-art multilingual models. We conduct comprehensive experiments to evaluate AraCLIP's performance across diverse datasets, including Arabic XTD-11, and Arabic Flicker 8k. Our results showcase AraCLIP's superiority in image retrieval accuracy, demonstrating its effectiveness in handling Arabic queries. AraCLIP represents a significant advancement in cross-lingual image retrieval, offering promising applications in Arabic language processing and beyond. See project page[1].

1 Introduction

Text-to-image retrieval models have been extensively studied in the literature, garnering attention due to their applicability and remarkable success across diverse domains such as medical imaging (Shu et al., 2024), automated image captioning (Emami et al., 2022; Afyouni et al., 2021), visual question answering (Aggarwal and Kale, 2020), and text-to-image generation (Li et al., 2022), among others. However, the majority of proposed models focus on English texts, neglecting the vast number of Arabic speakers worldwide. The Arabic language poses unique challenges, including its richness in morphology, complex grammar, and diverse dialects (Farghaly and Shaalan, 2009). Therefore, we aim to push the boundaries by building a dedicated text-to-image retrieval model that caters to the Arabic language.

Several learning approaches have been employed for text-to-image integration, including contrastive learning and teacher learning approaches. Contrastive learning has demonstrated a powerful learning approach to align text-image pairs. (Radford et al., 2021) proposed an efficient method called Contrastive Language-Image Pre-training (CLIP), which demonstrated capability in learning useful representations of text and images. However, the predominant focus has been on English text. Also, contrastive learning was found to be compositionally heavy as it was first introduced in (Radford et al., 2021).

The cross-lingual approach for text-to-image synthesis has gained significant attention in recent years, particularly in the context of cross-lingual models that can effectively capture the nuances and complexities of various languages. (Bianchi et al., 2021) employed an Italian model instead of a multi-lingual model, where cross-lingual models can capture the nuances and complexities of the Italian language, leading to more accurate image-text retrieval results. The teacher-learning approach for text-image integration for multiple languages was introduced in (Carlsson et al., 2022a) as they trained cross-lingual models for the Swedish language and multi-lingual models for many languages (about 68 languages) by utilizing machine translation while starting from a pre-trained text encoder for the needed target language.

To this end, this paper utilizes this technique to

[1] https://arabic-clip.github.io/Arabic-CLIP

Proceedings of The Second Arabic Natural Language Processing Conference, pages 102–110
August 16, 2024 ©2024 Association for Computational Linguistics

introduce the first Arabic CLIP, a model trained using teacher learning to align Arabic text with images. "AraCLIP" is a substantial extension of the CLIP model (Radford et al., 2021) specifically designed for image retrieval tasks in the Arabic language. AraCLIP is tailored to understand the unique terms and contextual constructs that form the Arabic language. Building upon the transformative CLIP architecture, we propose a novel approach that leverages the power of Knowledge Distillation (Gou et al., 2021) to seamlessly transfer the cross-modal knowledge encoded in a pre-trained English textual model to an Arabic counterpart.

Our key contributions in this paper are as follows:

- Releasing a translated and cleaned version of the **(CC3M+CC12M+SBU, Filtered synthetic caption by ViT-L)** dataset in Arabic with approximately 12.5 million caption-image pairs. Also, releasing a translated version of MS COCO **(Microsoft Common Objects in Context)** with approximately 123,280 caption-image pairs, and a testing **(Arabic XTD-11 dataset)**[2] dataset in Arabic with about 1,000 caption-image pairs.
- Proposing a new model for the text-image retrieval task (AraCLIP), marking the first instance of Arabic-focused models utilizing the original CLIP model and Knowledge Distillation.
- Surpassing the current multi-lingual models across two main evaluation metrics, including Mean Reciprocal Rank (MRR) and Recall, by about 10%.
- Releasing an online tool on Hugging Face for testing AraCLIP, along with the code for translation, dataset processing, training, and evaluation[3].

2 Methodology

In our research, we employed the method of Knowledge Distillation (teacher learning), as described by (Hinton et al., 2015) which is a technique that has become increasingly popular in the field of artificial intelligence (Gou et al., 2021). Knowledge distillation is a process where a smaller model, known as the student, is trained to mimic the behavior of a larger, pre-trained model, known as the teacher. This approach is particularly useful when dealing with large models that are computationally expensive to train and fine-tune. By distilling the knowledge from the teacher model into the student model, we can achieve similar performance with reduced computational costs.

There are many approaches to train CLIP models, such as contrastive learning and teacher learning for the text encoder. Contrastive learning is a type of self-supervised learning that involves training a model to differentiate between similar and dissimilar inputs. In the image-text retrieval systems, contrastive learning can be used to train a model to learn a joint representation of images and texts by maximizing the similarity between matching pairs and minimizing the similarity between non-matching pairs. However, we found that contrastive learning is computationally expensive to train and fine-tune a CLIP model using it.

Therefore, we have chosen the teacher learning approach, which depends on fine-tuning the text encoder for the Arabic language with a text encoder from an English CLIP model. In particular, we applied the teacher learning method by treating an English text model as the "teacher" and a pre-trained Arabic model as the "student". This approach allows us to leverage the knowledge learned by the English model and adapt it to the Arabic language, achieving better performance with reduced computational costs.

By using knowledge distillation and teacher learning, we were able to develop an efficient and effective image-text retrieval system that can retrieve relevant images from a large database based on a given text query. Our approach has shown promising results, and we believe it has the potential to be applied in various real-world applications, such as image search engines and multimedia retrieval systems.

2.1 AraCLIP Framework Overview

Our goal was to transfer the knowledge from the English model to the Arabic model, allowing the latter to effectively learn the associations with the visual domain. This was based on the assumption that the English model exhibits a strong correlation with the image model in the embedding space for matching text-image pairs. Consequently, the

[2] https://huggingface.co/datasets/khalidalt/xtd_11

[3] https://huggingface.co/spaces/Arabic-Clip/AraCLIP

text encoder of the Arabic language will be aligned with the image encoder of the CLIP model. By adopting this strategy, we eliminated the need to directly involve images in the training process. This method significntly reduces the computational resources required compared to contrastive learning (Carlsson et al., 2022a).

Through the use of data generated by neural machine translation, we trained the student model to produce embeddings that closely match those generated by the teacher model. During the training phase, we kept the teacher's CLIP text encoder unchanged, while only updating the parameters of the student language encoder. Our focus was on minimizing the embedding distance between the two models for text pairs, which consisted of an English caption and its corresponding Arabic translation.

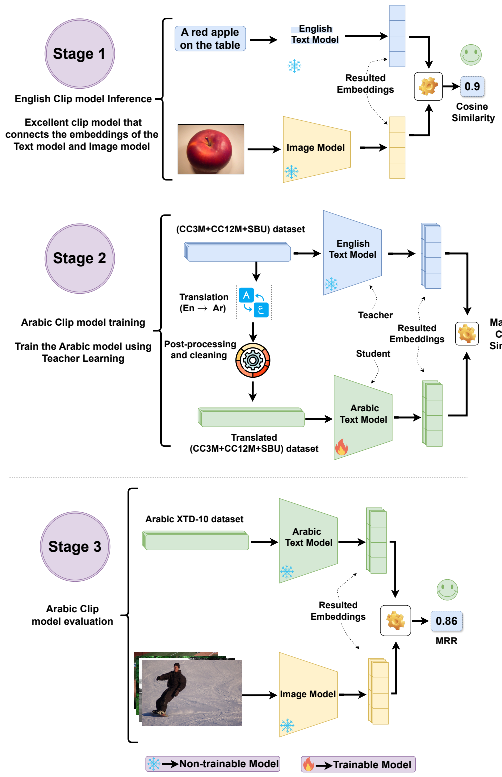

Figure 1: Cross-Lingual Learning Framework for Arabic Image Retrieval: Illustrating the three-stage process of training an Arabic text model to learn from an English text model, with the goal of understanding text-image relationships.

The overall view of the cross-lingual Arabic framework involves three main stages as shown in Figure 1.

Figure 1-stage 1 illustrates the English text model and the image model on contrastive learning. These models are strongly connected in the embedding space since this is our assumption from the beginning. Therefore, if we give the model an image and its caption, their embedding should show a high cosine similarity. Also, these models are already trained using contrastive learning to be able to capture the similarity between image-text pairs based on their embedding. Note that the English text model and image model can be separated.

In stage 2, we conduct the model training where we get the English text model (teacher model) from the first stage and a pre-trained Arabic model (student model), Arabert model (Antoun et al., 2020) in this work. The input for the teacher model will be the English caption and the input for the student model will be the translated and pre-processed Arabic caption. Then, the Arabic text model (student model) will produce an embedding vector of the Arabic caption that is similar to the embedding vector of the English caption that was produced from English model. Then, we minimize the Mean Squared Error (MSE) between the resulting output embeddings of these models.

It is important to highlight that this approach is different from the original training objective of CLIP, which focuses on training the model to establish correlations through cosine similarity calculations between pairs of images and texts. Nevertheless, it is feasible to directly employ cosine similarity in the context of teacher learning; however, prior research has demonstrated that the minimization of MSE yields a more informative signal for learning (Carlsson et al., 2020). On this stage, we are using the teacher-learning (Knowledge Distillation) training, as the student model will learn the features from the teacher, so it could be connected to the image model features.

Stage 3 shows the the evaluation of the Arabic text trained model (student model) with the image model, to check the performance in understanding text images. In our work, we evaluated it using different datasets (Arabic XTD-11 and Arabic flicker 8k) based on different metrics.

Dataset Splitting. All datasets, including those used for training, validation, and evaluation, have been translated, cleaned, and meticulously detailed in the accompanying Table 1.

Dataset Name	Splits	Size	Sampled from
Datasets processed by Vit-B-16-plus-240 model			

Arabic_3M_5M_ViT-B-16-plus-240	Training
	Validation
		Training=2M
		Validation=5000
			(CC3M+CC12M+SBU)

Arabic_MSCOCO_1st_ViT-B-16-plus-240	Training	Training=113281	MSCOCO dataset
Datasets processed by ViT-B-16-SigLIP-512 model			

Arabic_3M_5M_ViT-B-16-SigLIP-512	Training
	Validation
		Training=2M
		Validation=5000
			(CC3M+CC12M+SBU)

Arabic_MSCOCO_1st_ViT-B-16-SigLIP-512	Training	Training=113281	MSCOCO dataset
Evaluation Datasets			
Arabic_XTD-11	Evaluation	1000	MSCOCO dataset
Arabic_Flicker_8k	Evaluation	8000	Flicker 8k

Table 1: Dataset used on training, validation and evaluation.

2.2 Dataset Translation, Cleaning and Splitting

The integrity and quality of datasets play an important role in enhancing the performance outcomes of tasks predicated on machine learning algorithms, as shown in the comprehensive surveys and studies within the domain (Lee et al., 2021; Ilyas and Rekatsinas, 2022). Furthermore, the significance of dataset collection extends to multi-modal datasets, which often comprise web-crawled data. The aforementioned research underscores the necessity of dataset improvements to support the efficacy of models across a diverse array of tasks (Nguyen et al., 2023; Betker et al., 2023). This body of work collectively affirms that systematic enhancements to dataset quality directly contribute to the amplification of model capabilities, underscoring the criticality of clean data in the advancement of machine learning technologies.

Unfortunately, there is currently a lack of a large, clean dataset that includes Arabic captions and corresponding images for image retrieval tasks (Mohamed et al., 2023; Hejazi and Shaalan, 2021). To address this gap, we have adopted the approach of translating a comprehensive English datasets into Arabic.

Dataset Translation. We used synthetic Conceptual Captions (CC3M+CC12M+SBU) dataset, as previously utilized in (Li et al., 2022; Carlsson et al., 2022b). Our selection consisted specifically of (CC3M + CC12M + SBU), which was filtrated by the ViT-L model[4] and comprises approximately 12,556,500 samples (Li et al., 2022). In addition, we used an open source neural machine translation model to translate English captions into Arabic. Specifically utilizing the (Helsinki-NLP/opus-mt-en-ar) model[5] that has been published by University of Helsinki (Tiedemann, 2020). Our selection for this model based on a human manual testing of open source models. The translation process that we used in our study was derived from the original code[6]. We also used an edited version of this translation code for our processes that is suitable for Arabic.

Dataset Cleaning. The dataset was subjected to two primary cleaning operations. Firstly, captions with duplicate unrelated subtext (more than five times where we chose it based on the analysis and testing) were removed. In addition, captions that contain any English text or symbols were removed, including instances where the « or » symbols were present, which indicates the justification of the text, as well as captions that were empty. Table 2 shows these cases with examples. For data cleaning processes, we used the cleaning functions

[4] https://github.com/salesforce/BLIP#pre-training-datasets-download

[5] https://huggingface.co/Helsinki-NLP/opus-mt-en-ar

[6] https://github.com/FreddeFrallan/Multilingual-CLIP/tree/main/translation

from Maha library[7] (Al-Fetyani, 2022).

Case	Example
Duplicated text more than 5 times in the caption	اللغـة العربيـة للغـة الرامـا الرامـا الرامـا الرامـا الرامـا الرامـا، اللغـة الاسلاميـة، الرامـا، الرامـا الرامـا
It has « or » which means a justification of of the text	«به الكثير من الأشجار» أي ماء نقي « ماء صافي ».
If it contains any English text or symbols	برج المياه في منطقة NC, الشاطئ في جزيرة المحيط.

Table 2: Examples of captions were removed from the dataset.

The second major cleaning operation involved the removal of Harakat, Tatweel, extra spaces, and other symbols from the captions. Detailed examples with corresponding references can be found in Table 3.

Case	Reference
Tatweel	–
Harakat	Fathatan, Dammatan, Kasratan, Fatha, Damma, Kasra, Shadda, Sukun
Remove extra spaces	Extra spaces on the text
Other symbols	” ’ ( ) * ♪ :: : - [ ] « « »

Table 3: Symbols Removal from the dataset.

3 Model Training and Experimental Setup

We conducted extensive experiments using the teacher-learning approach, building upon the code provided by (Carlsson et al., 2022a) and incorporating our own enhancements and integrations. The datasets used for training, validation, and evaluation are summarized in Table 1.

This table is organized into three sections, each corresponding to a different experimental setup. The first section presents the datasets processed by the ViT-B-16-plus-240 model, where the English encoder served as the teacher model. We pre-computed the embeddings of the English captions to reduce computation time and memory requirements during training and validation. The datasets used in this setup were (CC3M+CC12M+SBU) and MSCOCO, with training samples from both datasets and validation samples only from (CC3M+CC12M+SBU).

The second section details the datasets used when the English encoder of the ViT-B-16-SigLib-512 model was the teacher model. Again, the training sets consisted of both (CC3M+CC12M+SBU) and MSCOCO datasets, while the validation set was only from (CC3M+CC12M+SBU).

The third section lists the evaluation datasets, which comprised Arabic_XTD-11 and Arabic_flicker_8k. These datasets were used with various metrics, as discussed in Section 4. By using these diverse datasets, we aimed to comprehensively evaluate the performance of our model across different problem domains and input types.

During the training, the main goal was to maximize the cosine similarity between the embeddings of the English and Arabic encoders using the teacher-learning approach.

4 Models Evaluation

Evaluating the performance of our approach relative to existing methods is crucial for understanding its strengths and limitations. In this section, we present a comparison of our work to a state-of-the-art multi-lingual model, using both quantitative and qualitative evaluation metrics. Quantitatively, we compare our approach to others using metrics such as MRR and recall, while qualitatively, we assess the model's understaning of the text queries based on two categories such as complex sentence category performance under noise category. Our results show that our approach outperforms the existing multi-lingual model in several key tasks.

For text-image retrieval systems, two crucial metrics are Mean Reciprocal Rank (MRR) and Recall. MRR measures the reciprocal of the rank at which the first relevant item appears in the search results, averaged over all queries. It emphasizes the importance of ranking relevant items high in the search results. Recall, on the other hand, measures the proportion of relevant items that are retrieved out of all relevant items in the dataset. It is often calculated at a specific cutoff (e.g., Recall@K), which means it only considers the top K retrieved items. Both metrics are important for evaluating the performance of text-image retrieval

[7] https://maha.readthedocs.io/en/latest/overview.html#cleaners

Model Name	Arabic XTD-11			Arabic Flickr8k		
	MRR@1	MRR@5	MRR@10	MRR@1	MRR@5	MRR@10
Others work (Carlsson et al., 2022a)						
XLM-Roberta-Large-Vit-B-16Plus	0.578	0.682	0.693	0.258	0.358	0.372
Our work						
Arabert-v2-base-ViT-B-16-SigLIP-512-2M	**0.673**	0.752	0.762	0.380	0.474	0.487
Arabert-v2-base-ViT-B-16-SigLIP-512-2M-mscoco	0.669	**0.755**	**0.764**	**0.383**	**0.474**	**0.488**
Arabert-v2-large-ViT-B-16-plus-240-2M	0.554	0.663	0.675	0.293	0.380	0.393

Table 4: MRR results of the text-to-image retrieval models based Arabic XTD-11 and Arabic Flickr8k datasets.

Model Name	Arabic XTD-11			Arabic Flickr8k		
	R@1	R@5	R@10	R@1	R@5	R@10
Others work (Carlsson et al., 2022a)						
XLM-Roberta-Large-Vit-B-16Plus	0.578	0.851	0.934	0.258	0.526	0.633
Our work						
Arabert-v2-base-ViT-B-16-SigLIP-512-2M	**0.673**	0.878	0.948	0.380	**0.630**	0.724
Arabert-v2-base-ViT-B-16-SigLIP-512-2M-mscoco	0.669	**0.887**	**0.954**	**0.383**	0.625	**0.900**
Arabert-v2-large-ViT-B-16-plus-240-2M	0.554	0.832	0.919	0.293	0.528	0.626

Table 5: Recall results of the text-to-image retrieval models based Arabic XTD-11 and Arabic Flickr8k datasets.

systems, with MRR focusing on the ranking of the first relevant item and Recall focusing on the overall proportion of relevant items retrieved.

Our proposed models for text-image retrieval are designed to leverage the strengths of both Arabic language understanding and visual representation learning. The models are, the first mdoel is Arabert-v2-base-ViT-B-16-SigLIP-512-2M: This model combines the Arabert-v2-base language model with the ViT-B-16-SigLIP-512 visual transformer, trained on 2 million samples. The second model is Arabert-v2-base-ViT-B-16-SigLIP-512-2M-mscoco: This model is similar to the previous one but fine-tuned on the MSCOCO dataset, which provides additional training data for the model to learn from. The last model is Arabert-v2-large-ViT-B-16-plus-240-2M: This model uses the larger Arabert-v2-large language model and the ViT-B-16-plus-240 visual transformer, trained on 2 million samples.

4.1 Quantitative Evaluation

We evaluated the performance of our models on many datasets based on different metrics. We used the text-image pairs from Arabic XTD-11 dataset, Arabic Flickr8k dataset to test the models' performance based on MRR and Recall metrics.

Table 4 describes the overall performance with respect to the MRR metric based on the Arabic XTD-11 dataset and the Arabic Flickr8k dataset. Our proposed AraCLIP models generally outperform the other work model, XLM-Roberta-Large-Vit-B-16Plus (Carlsson et al., 2022a), which we refer to as 'mCLIP' (Multilingual CLIP) across most metrics. This is evident in higher MRR scores at MRR@1, MRR@5, and MRR@10 for both datasets.

On the Arabic XTD-11 dataset, AraCLIP models, particularly Arabert-v2-base-ViT-B-16-SigLIP-512-2M and its variant with mscoco, show superior performance compared to the mCLIP model, with MRR scores above 0.66 at MRR@1, indicating a higher likelihood of returning the most relevant result in the first position. There is a noticeable decline in performance with the Arabert-v2-large-ViT-B-16-plus-240-2M models, yet they still maintain an edge over mCLIP model.

On the Arabic Flickr8k dataset, AraCLIP models also outperform mCLIP model indicating a consistent ability to rank relevant results higher. The Arabert-v2-base-ViT-B-16-SigLIP-512-2M models (both standard and mscoco) show the best performance, suggesting that the base variant might be more effective for these specific

ID	Category	Description	Fig.	Example
1	Complex sentence	Sentence with some complexity “mention things on the input sentence”	2	رجل يتزلج على الماء بلوح شراعي (Man surfing on a windsurf)
2	Performance under noise	Sentence with some objects that not clear on the image	3	كلب يهاجم قطة بينما القطة تحت مقعد خشبي (A dog attacks a cat while the cat is under a wooden bench)

Table 6: Qualitative categories for testing the models.

datasets. As for model variants, the use of mscoco in AraCLIP models appears to slightly improve performance, which could indicate the benefit of additional training data or diversity in training material. However, the ‘large’ variants of AraCLIP models, while performing better than the mCLIP model, do not outperform our ‘base’ variants. This might suggest that the additional complexity in the large models does not translate to a significant performance gain for these specific tasks.

The analysis of Table 4 indicates a strong performance of AraCLIP models, particularly the base variants, over the mCLIP model across both datasets.

Table 5 shows a comparison of Recall results (R@1, R@5, R@10) of the various models based on the Arabic XTD-11 dataset and Arabic Flickr8k dataset. The mCLIP model is compared against our AraCLIP models. The overall performance of AraCLIP models generally outperform mCLIP model in most metrics across both datasets. This indicates a potentially more effective approach in our models for these specific tasks.

As for consistency across datasets, the performance improvement of AraCLIP models is consistent across both Arabic XTD-11 dataset and Arabic Flickr8k dataset, suggesting robustness in diverse data scenarios. Different variants of AraCLIP models show slightly varying performance. For instance, models trained with mscoco dataset tend to show higher Recall, especially at R@5 and R@10. While all AraCLIP models outperform the CLIP model in the Arabic XTD-11 dataset, the margin of improvement varies, indicating possible areas for further optimization.

Our observations suggest that AraCLIP models are more adept at capturing relevant information in the tested datasets compared to mCLIP model, especially in terms of Recall. The variations among AraCLIP variants also provide insights into how different training approaches and architectures impact performance.

Based on the above analysis of Tables 4, and 5, it is evident that the performance of the AraCLIP models generally surpasses that of the mCLIP model in various metrics across different datasets. The AraCLIP models demonstrate higher efficiency in MRR and Recall rates, indicating a more robust capability in understanding and processing Arabic language text. Overall, the AraCLIP models represent a significant advancement in Arabic language processing field, with potential areas for further improvement identified through these comparisons.

4.2 Qualitative Evaluation

In this part, we evaluated on model for qualitative perspectives. We focused on two categories of assessments, complex sentence understanding and performance under noise. Table 6 shows sample sentences that we used for the testing. Also, we added the equivalent translated English sentence for clarification. For the experiments, we used the Arabic flicker 8k dataset (ElJundi. et al., 2020) for image retrieval along with input sentences based on the two categories as shown in Table 6.

Our evaluation is based on our AraCLIP model(Arabert-v2-base-ViT-B-16-SigLIP-512-2M) and mCLIP model (XLM-Roberta-Large-Vit-B-16Plus). For each query, we chose the top three images retrieved by both our model and mCLIP model. The upper set of images displays the results generated by our model, AraCLIP model, while the lower set showcases the results produced by the mCLIP model.

Figure 2 demonstrates a result of the complex category which has some complexity on the sentence as رجل يتزلج على الماء بلوح شراعي (Man surfing on a windsurf). For this example, we see that AraCLIP was better than mCLIP on retriving the images related to this query. AraCLIP retrieved images that have sail while mCLIP fails on this. Both model retrieve images that have some objects mentioned on the input text.

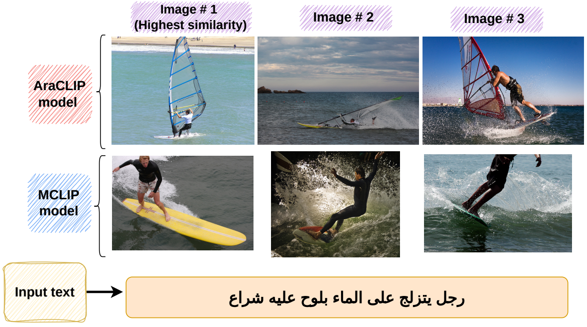

Figure 2: Complex sentence category, (English translation of the input: Man surfing on a windsurf).

In Figure 3, we provided an example of a sentence with some objects that are shown clearly as كلب يهاجم قطة بينما القطة تحت مقعد خشبي (A dog attacks a cat while the cat is under a wooden bench) on the retrieved images. Our model was able to capture the objects even though they are not clear in the image, it got the related image as the highest image while the other images are related by containing some related objects to the input sentence on them. While mCLIP struggles to capture the correct image as the first option, it can retrieve related images later which has the correct image.

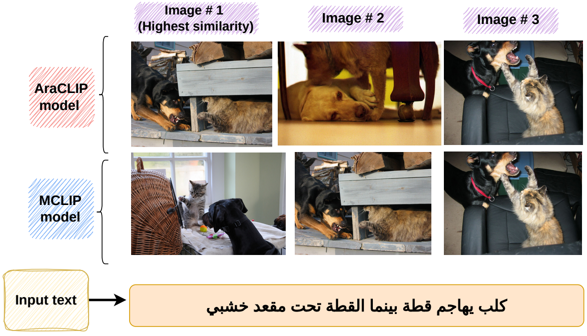

Figure 3: Performance Under Noise category, (English translation of the input: A dog attacks a cat while the cat is under a wooden bench).

Overall, we presented a comparison between our model (AraCLIP) and the best CLIP model trained on the multilingual in (Carlsson et al., 2022b) which is mCLIP. However, our model still needs many improvements across many metrics which open the door for future work in many aspects including the translation, model selection and the evaluation aspects that we should assets based on them.

Conclusion

In this study, we have successfully extended the CLIP model to support the Arabic language using AraCLIP framework. Given that CLIP serves as a fundamental backbone for numerous image-text applications, the introduction of an Arabic version represents a significant advance. Our model outperformed current multilingual state-of-the-art models by using different metrics and retrieval tasks. We believe that our work represents a significant step toward developing more robust and efficient models for Arabic language vision tasks. This research lays the groundwork for many research opportunities and questions that need to be addressed in future work.

References

Imad Afyouni, Imtinan Azhar, and Ashraf Elnagar. 2021. Aracap: A hybrid deep learning architecture for arabic image captioning. *Procedia Computer Science*, 189:382–389.

Pranav Aggarwal and Ajinkya Kale. 2020. Towards zero-shot cross-lingual image retrieval. *arXiv preprint arXiv:2012.05107*.

Mohammad Al-Fetyani. 2022. Maha Processing Library.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3).

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022a. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022b. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International conference on learning representations*.

Obeida ElJundi., Mohamad Dhaybi., Kotaiba Mokadam., Hazem Hajj., and Daniel Asmar. 2020. Resources and end-to-end neural network models for arabic image captioning. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 233–241. INSTICC, SciTePress.

Jonathan Emami, Pierre Nugues, Ashraf Elnagar, and Imad Afyouni. 2022. Arabic image captioning using pre-training of deep bidirectional transformers. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 40–51.

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):1–22.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Hani Hejazi and Khaled Shaalan. 2021. Deep learning for arabic image captioning: A comparative study of main factors and preprocessing recommendations. *International Journal of Advanced Computer Science and Applications*, 12(11).

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ihab F. Ilyas and Theodoros Rekatsinas. 2022. Machine learning and data cleaning: Which serves the other? *J. Data and Information Quality*, 14(3).

Ga Young Lee, Lubna Alzamil, Bakhtiyar Doskenov, and Arash Termehchy. 2021. A survey on data cleaning methods for improved machine learning model performance. *arXiv preprint arXiv:2109.07127*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Alcoba Inciarte, and Muhammad Abdul-Mageed. 2023. Violet: A vision-language model for arabic image captioning with gemini decoder. *Preprint*, arXiv:2311.08844.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Chang Shu, Yi Zhu, Xiaochu Tang, Jing Xiao, Youxin Chen, Xiu Li, Qian Zhang, and Zheng Lu. 2024. Miter: Medical image–text joint adaptive pretraining with multi-level contrastive learning. *Expert Systems with Applications*, 238:121526.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.