



School of Information Technology and
Engineering at the ADA University



School of Engineering and Applied Science
at the George Washington University

**INVESTIGATION OF MULTIMODAL VISION-LANGUAGE TASKS IN
LOW-RESOURCE LANGUAGES**

A Thesis

Presented to the Graduate Program of Computer Science and Data Analytics of the
School of Information Technology and Engineering
ADA University

In Partial Fulfillment of the
Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Ali Asgarov

Supervisor Dr. Samir Rustamov

April 2024

THESIS ACCEPTANCE

This Thesis by: Ali Asgarov

Entitled: *Investigation of Multimodal Vision-Language Tasks in Low-Resource Languages* has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

(Adviser) (Date)

(Program Director) (Date)

(Dean) (Date)

ACADEMIC INTEGRITY STATEMENT

“I affirm that this is my own work, I attributed where I used the work of others, I did not facilitate academic dishonesty for myself or others, and I used only authorized resources for my Thesis, per the ADA University Academic Integrity requirements. If I failed to comply with this statement, I understand consequences will follow my actions. Consequences may range from failing the course to expulsion from the program/university and may include a transcript notation.”

Ali Asgarov

01.05.2024

(Full Name)

(Signature)

(Date: DD.MM.YY)

ABSTRACT

In this thesis, we have investigated the challenges of building multimodal vision-language models for image retrieval tasks in a low-resource languages, particularly in Azerbaijani. Due to the multiple reasons this task is challenging for the low resource languages. First reason is the limitations of large-scale vision-language models, such as CLIP, which does not support approximately 90% of low-resource languages. Another reason is present computational challenges, even when there are Parameter Efficient Fine-Tuning (PEFT) methods. So, we have explored the integration of a Multilingual BERT with the base image encoder models to build custom models from the ground up for those languages. Our investigations include a variety of model architectures, including ResNet50, EfficientNet0, Vision Transformer (ViT), Tiny Swin Transformer alongside the multilingual BERT model, to evaluate performance across different datasets. Our findings shows significant variations in model performance, influenced by data quality and annotation richness. For instance, models generally show better in-domain performance on the MSCOCO dataset compared to Flickr datasets, due to the MSCOCO's comprehensive annotations and diverse image content which is more than 300K images in total. To solve these challenges, our study includes the generation of synthetic datasets through machine translation for Azerbaijani and image augmentation, along with a comparative analysis of various encoder models to establish efficient, cost-effective training strategies for low resource languages. Augmented image data boosted model performance, with EfficientNet0 achieving 0.87 MAP on Flickr30k, while almost all the models struggled with out-domain generalization. Tiny Swin Transformer exhibited adaptability across datasets with consistent 0.80 MAP scores. Our approach not only enhances model adaptability across different domains but also contributes to the broader application of vision-language retrieval systems in low-resource languages. By sharing our configurations and results, we aim to facilitate further research and technological adaptation across diverse linguistic landscapes. We release our code and pre-trained model weights at <https://github.com/aliasgerovs/azclip>

Keywords: Multimodal vision-language models, Image retrieval, Low-resource languages, Synthetic datasets, Machine translation, Encoder models, Computational efficiency, Azerbaijani language

Contents

1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement.....	1
1.3 Objective	2
2 Literature Review	6
2.1 Introduction to Multimodal Retrieval	6
2.2 History of Multimodal Tasks - Image Retrieval	7
2.3 Datasets Used in Multimodal Tasks	9
2.4 Training Techniques from Different Studies	10

2.5	Multimodal Image Retrieval Model Architectures	11
2.6	CLIP Model and Attention Network	12
2.7	Challenges with Low-Resource Languages	12
2.8	Review of Image and Text Encoding Methods	14
2.9	Adaptation Techniques for Low-Resource Settings	14
3	Research Methodology	16
3.1	Method	16
3.2	Dataset and Preprocessing	16
3.2.1	Generated Dataset for Azerbaijani Language	18
3.3	Model Architecture	19
3.3.1	Image Encoders	19
3.3.2	ResNet-50	20
3.3.3	EfficientNet-B0	20
3.3.4	Vision Transformer (ViT)	22
3.3.5	Swin Transformer	23
3.3.6	ConvNeXt	24
3.3.7	MobileNetV3	25
3.3.8	Text Encoder	25
3.3.9	Projection Head	26
3.3.10	Contrastive Learning Loss	28
3.4	Enhancing Image Feature Extraction	29
3.5	Enhancing Text Feature Extraction	30
3.6	Training Procedure	31
3.7	Image Retrieval Process	33
3.8	Evaluation Metrics	34
3.8.1	Mean Average Precision (MAP)	34
3.8.2	Mean Average Recall (MAR)	34
3.8.3	Mean Average F1 Score (MAF1)	34
3.8.4	Mean Average Accuracy (Top-k Accuracy)	35
3.9	Implementation	35

4	Results	37
4.1	Experiment Design	37
4.2	Controlled Variables and Their Impact on Model Performance	38
4.3	Analysis of Results	39
5	Limitations	44
6	Conclusion and Future Work	46
7	Appendix	52
7.1	Model Inference Results	52
7.2	Model Implementation with Enhanced Contrastive Loss	54

List of Figures

FIGURE 1. Model Inference Example 1. The query posed is "Velosiped sür@n g@nc cütlük."	2
FIGURE 2. Clip Model Structure. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples [35].	13
FIGURE 3. COCO Dataset. Example images and captions from the Microsoft COCO Caption dataset [25].	17
FIGURE 4. Translation of Captions from Datasets. Translation of Captions from COCO, Flickr30k and ROCO datasets.	18
FIGURE 5. Resnet-50 Model Architecture. Comprising an input layer followed by convolutional layers, batch normalization, ReLU activation, and max pooling. It progresses through stages with convolutional and identity blocks, ending with average pooling and a fully connected layer to produce the output. [13]	21
FIGURE 6. Architecture of EfficientNet-B0. Architecture with MBConv as basic building blocks. [46]	21
FIGURE 7. Overview of the ViT Base Model architecture. The Vision Transformer (ViT) model utilizes transformer architecture for image classification tasks. Each transformer block processes image patches, enabling the model to handle images as sequences of tokens. [9]	22
FIGURE 8. Overview of the Swin Transformer architecture and Swin Transformer blocks. The Swin Transformer is a hierarchical transformer model designed for visual recognition tasks. It utilizes a hierarchical attention mechanism to capture information at different scales efficiently. [27]	23
FIGURE 9. Overview of the ConvNeXt-Tiny overall network structure and ConvNeXt Block structure. The ConvNeXt-Tiny architecture is designed for efficient and lightweight convolutional neural networks. It employs ConvNeXt Blocks, which are composed of parallel convolutional pathways with different kernel sizes and depths, allowing the network to capture multi-scale features effectively. [28]	24
FIGURE 10. MobileNetV3 Model Architecture. The flow begins with a 1x1 convolutional layer with non-linearity, followed by a 3x3 depthwise convolution also with non-linearity. A pooling layer reduces the spatial dimensions, feeding into a sequence of fully connected layers with ReLU and hard-sigmoid activations. [14]	25
FIGURE 11. Multilingual BERT training approach using masked language	

modeling. It shows tokenized text input with some tokens replaced by [MASK] for the model to predict, facilitating the model’s learning of language context. The process utilizes multilingual data from Common

Crawl, enabling mBERT’s application across different languages. [7] . . . 26

FIGURE 12.**Transformer architecture.** Featuring parallel encoder layers that process input sequences and decoder layers that generate output sequences. The architecture employs self-attention to relate different positions within a single sequence for encoding and decoding, enabling the model to translate between languages effectively. [48] 27

FIGURE 13. **Projection Head:** This diagram represents a dual-encoder framework where separate text and image encoders generate respective embeddings. These embeddings are projected into a shared space to compute compatibility scores for image-text pairs, illustrated by the matrix of dot products ($I_n * T_n$). This enables the matching of text to the most relevant images and for image retrieval based on textual queries. [34] 27

FIGURE 14.**Model’s Architecture and training process.** Using separate encoders for text and images to learn embeddings, aiming to maximize the cosine similarity between correct text-image pairs and minimize it for incorrect ones during training. [34] 31

FIGURE 15.**Optimizers.** Number of times ArXiv titles and abstracts mention specific optimizer per year. [19] 32

FIGURE 16.**Base Model (AzClip) Training Loss** 32

FIGURE 17.**Base Model (AzClip) Epochs** 32

FIGURE 18.**Model Inference Example 2.** The query posed is "Qız küç@d@yem@k yeyir" FIGURE 19.**Model Inference** 45

Example 3. The query posed is "M@kt@b uşaqları futbol oynayır" FIGURE 20.**Model** 45

Inference Example 4. The query posed is "Trafik zamanı qırmızı işıqda dayanan avtomobill@r" FIGURE 21.**Model** 52

Inference Example 5. The query posed is "Ofisd@kompüter qarşısında işl@y@n proqramçılar" FIGURE 22.**Model** 52

Inference Example 6. The query posed is "Ata v@oğul göl k@narında balıq tuturlar" FIGURE 23.**Model** 53

Inference Example 7. The query posed is "Avtomobil sür@n qadın" 53

FIGURE 24.**High Level Training Logic.** 56

List of Tables

TABLE 1. Performance and architectural specifications of the ResNet-50 model. 20

TABLE 2. Performance and architectural specifications of the EfficientNet-B0 model.	TABLE 3.21	
Technical specifications and performance metrics of the ViT Base model		
	(ViT _{B(16)}),	22
TABLE 4. Technical specifications and performance metrics of Swin-Tiny model.	. . .	23
TABLE 5. Performance and architectural specifications of the ConvNeXt tiny model.		24
TABLE 6. Technical specifications and performance metrics of the MobileNetV3 model.		25
TABLE 7. Summary of Base Model Training (Resnet50 + mBERT) Parameters . . .	TABLE 8.33	
Average Evaluation Metrics for Base ResNet50 + Base Multilingual Bert		
Across MSCOCO Dataset (Base Loss).	TABLE 9. Average40	
Evaluation Metrics for Base ResNet50 + Base Multilingual Bert		
Across MSCOCO Dataset (Enhanced Loss).		40
TABLE 10. Average Evaluation Metrics for Augmented Image Data (Technique - 1) with ResNet-		
50 + Multilingual Bert Across Flickr30k Dataset (Enhanced		
Loss).		41
TABLE 11. Average Evaluation Metrics for Base EfficientNet0 + Base Multilingual		
Bert Across Flickr30k Dataset.		41
TABLE 12. Average Evaluation Metrics for Augmented Image Data (Technique - 1) with		
EfficientNet0 + Enhanced Base Multilingual Bert (Technique - 1)		
Across Flickr30k Dataset. (Enhanced Loss)	TABLE 13. Average41	
Evaluation Metrics for Base ViT + Base Multilingual Bert Across		
Flickr8k Dataset. (Base Loss)	TABLE 14. Average42	
Evaluation Metrics for Tiny Swin Transformer + Base Multilin-		
gual Bert Across Flickr8k Dataset. (Base Loss)		42
TABLE 15. Part of the model (Table 14) results showing detailed evaluation metrics for various		
scenarios (in Azerbaijani). Each metric is calculated based on		
15 retrieved images per sample.		43

LIST OF ABBREVIATIONS

Abbreviation Explanation

CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Networks
BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
CV	Computer Vision
ML	Machine Learning
AI	Artificial Intelligence
API	Application Programming Interface
GPU	Graphics Processing Unit
IoU	Intersection over Union
TPU	Tensor Processing Unit

SOTA State of the Art
PEFT Parameter Efficient Fine-Tuning
MAP Mean Average Precision MAR Mean
Average Recall
MAA Mean Average Accuracy
MAF1 Mean Average F1 Score

1 Introduction

1.1 Background and Motivation

The digital world is overflowing with vast amount of information. Text, images, and videos are produced at a very high rate. Traditional search systems, designed for textual queries, struggle to keep pace with this rate. For example, in a keyword based search, the results tend to be extensive, not capturing the intent of the user or the richness of multimedia data. These challenges present a barrier to accessibility to the searched data. Ideally, information retrieval systems should given an opportunity to the users to find what they need, regardless of their native language or preferred mode of interaction with the model. This is where multimodal image retrieval comes in, it allows the search to use not only a text, but also pictures, spoken words, or a combination of different input modalities. This approach goes beyond the improvement of search, so that it makes information far more accessible to people, regardless of the language they speak or how they prefer to search for the information. For instance, in image-to-image search, one of the previously suggested approach, you could point your camera at a building to search for its architectural style, or use a drawing to find clothes online. These are just a couple of ways that multimodal data search can help people search for what they need more efficiently. However, a significant challenge exists. Most of the multimodal data retrieval systems are dependent on on large, complex models trained on vast amount of datasets. These models are often resource intensive and require access to extensive training data in specific languages. This creates a challenge in the fact that languages with limited digital resources that are often spoken by a significant part of the global world are left behind. This research tries to close that gap. We believe that the benefits that such models can bring should be available to all, irrespective of their language barriers. By making such systems more memory and computation efficient, as well as ways to leverage the vast amount of available image data in all languages, we can unlock the true potential of multimodal retrieval for all languages, whether rich resources or low resources.

1.2 Problem Statement

Most of the recent multimodal systems suffer main deficiencies, especially those vision centric or vision and language systems in low resource linguistic settings are less efficient in terms of scalability with high volume and high dimensional data. Mostly, this is the imbalanced availability of data types, often the case in image enabled settings is the low availability of textual sources. These issues make it very difficult to combine such data for researching or training multimodal systems in languages that have not many digital products available. Data, for example, text, image, audio, or code is not connected in most cases and one available dataset would not provide the utility desired for diversity of the applications. In addition to that, computational requirements of advanced models, like CLIP models, are very high and cannot be adjusted for the limitations of computational resources. Therefore, the systems built on CLIP or even models of the same language that have scale of billions of parameters will not be applicable widely for low resource languages. Even current models are not easily fine tuned since even parameter efficient fine tuning methods (PEFT) require computation resources which makes it very costly.



Figure 1: **Model Inference Example 1.** The query posed is "Vəlosiped sürən gənc cütlük."

1.3 Objective

The primary objective of this thesis is to develop a multimodal vision language retrieval system adapted for the Azerbaijani language, balancing computational efficiency and scalability in low-resource settings. An important part of this research is to analyse ways of training developed model's performance across different domains focusing on the constrained availability of data. Main concern in the this study is whether synthetic data could be generated and utilized for improving the training dataset and where it could boost the performance and effectiveness of the model. The main aim is to create a model that can not only show a high performance within the Azerbaijani language but also represent a scalable skeleton of the model to be applied in the context of other similar low-resource languages, for example Kazakh, Ozbekh, or any other one. This approach is aimed at expanding the use of powerful AI technologies into various linguistically diverse and low-resource settings to make computationally challenging problems accessible to a wider range of people across the world. With doing that, this process aims to democratize access to progressive machine learning towards addressing the existing technological gap by overcoming the barriers of implementation and adaptation of advanced AI technologies in underrepresented regions.

Research Questions

The purpose of this thesis is to assess and improve multimodal vision-language retrieval systems for low-resource languages, with a particular focus on the Azerbaijani language. The research is based on the following questions:

1. How multimodal retrieval models be built for low-digital resource languages, and how to build a multimodal vision-language model for Azerbaijani?

2. What techniques are applicable to balance the trade off between computational resource use and model strength?
3. How to build such a multimodal solution for Azerbaijani language, that could be replicated for other low resource languages as well.
4. How does the process of synthetic data generation enhance model performance and efficiency ?
5. How far do the generated models extend across the low-resource languages in general, and what is shown by the analysis of the visual encoder and text decoder performance on in-domain and out-of-domain data in this respect?