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

# List of Figures

# List of Tables

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 "Velosiped sür@n g@nc 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 lowresource 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?