American Journal of Engineering Research (AJER) ***2026***
American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN: 2320-0936
Volume-15, Issue-3, pp-44-55
www.ajer.org
Research Paper Open Access

Cognitive Model of Visual Captioning in Georgian

Davit Bitmalkishev [1], Sergo Tsiramua [2], Sulkhan Sulkhanishvili [3]
[1](*Department of Informatics, University of Georgia,* Tbilisi, Georgia),
[2](*Institute of Informatics, University of Georgia,* Tbilisi, Georgia),
[3](*Department of Informatics, University of Georgia,* Tbilisi, Georgia)
Corresponding Author:Davit Bitmalkishev

Abstract - *When developing "Martha" – our Georgian visual caption model – we aimed to understand how people see and describe the world around them. The idea emerged from several intensive discussions where we discussed how visual perception is connected to language and how language influences what we see. This connection turned out to be particularly interesting in the Georgian language, where word order is flexible and morphology plays a crucial role in conveying meaning.*
For this reason, instead of using predefined and templated captions, we decided to create a system that linked the recognition of visual features to their conceptual and grammatical structures. In other words, we tried to give "Martha" not only a technical perspective, but also a cognitive one – a modeling based on sensations and meaning.
In practice, this means that "Martha" is not limited to naming objects; It tries to convey what a person might naturally say when they see the scene. For example, if an image shows a child rolling a ball, "Martha" will not simply say "child and ball," but will create a contextual description, such as "child rolling a ball in the yard."
The results of the first experiments are quite promising — the captions generated by the model are accurate, but at the same time sound natural and contextual. This shows that cognitively informed approaches can provide substantial progress in languages that are relatively resource-poor for artificial intelligence. We believe that such an approach opens up new prospects for the development of multimodal systems that will be closer to the principles of human perception and understanding.
Keywords: *cognitive modeling, visual captioning, Georgian language, multimodal learning, natural language generation, low-resource NLP*

Date of Submission: 08-03-2026 Date of acceptance: 20-03-2026

I. INTRODUCTION

One of the most important and at the same time most difficult challenges in artificial intelligence research is to train models that can work seamlessly across different languages and cultural environments. Modern systems such as BLIP-2, Flamingo or GPT-4V are undoubtedly technological advances that have taken the interaction of visual perception and language to a whole new level. However, most of them are trained mainly on English-language data. This focus has yielded incredible results on a global scale, but at the same time has clearly revealed inequality — speakers of small languages, including Georgian, practically do not have high-quality tools that correspond to their cultural and linguistic reality.

It is this circumstance that inspired our project — "Martha", which seeks to fill this gap. Our goal was to create a hybrid system that would combine computer vision with natural language generation in such a way that it would be possible to describe images in Georgian — directly, without translation. Simply put, "Martha" combines a BLIP-2 transformer, which extracts high-level visual features from images, and a ByT5 language model, which transforms this data into coherent and natural Georgian sentences. The synergy of these two parts creates a system that can "see" and "speak" Georgian — not as a secondary language, but as a native language.

"Martha" is not just a technological experiment, however. For us, it is also somewhat cultural — an attempt to establish a place for a small language in the field of artificial intelligence. When a system sees and speaks Georgian, it is a small but symbolic step towards linguistic inclusion and cultural self-awareness. It reminds

us that technological progress should not be limited to the "big" languages, and that innovation becomes real when it is shared by everyone—including languages whose sounds are still rarely heard in the world of machines.

For Georgia, this may be one of the first steps toward AI systems that genuinely reflect local expression and context.

A. GENERAL OVERVIEW OF ARTIFICIAL NEURAL NETWORKS

The paper [1] provides a detailed description of the working principle of an Artificial Neural Network (ANN). An ANN is a mathematical model that simulates the structure and functionality of biological neural networks. The basic building block of every ANN is the artificial neuron, which is a simple mathematical function. At the input of the artificial neuron, the inputs $\{x_1, x_2, \dots, x_n\}$ are weighted - each input x_i is multiplied by an individual weight w_i, where $i \in [1, n]$, from the set $\{w_1, w_2, \dots, w_n\}$. In the middle section of the neuron, a summation function adds all the weighted inputs along with a bias term. At the output, this sum passes through an activation function (also called a transfer function):

$$Y = F\left(\sum_{i=1}^{n} w_i x_i + b\right),$$

where, x_i, $i \in [1, n]$, is an input value, w_i, $i \in [1, n]$, is a weight value, b - bias, F is a transfer function, Y - output value.

Interconnected artificial neurons form an artificial neural network. Different types of artificial neural network topologies are suited to solving different types of problems. Once the type of the problem is determined, we must select an appropriate neural network topology and then fine-tune it. This involves adjusting both the structure of the topology and its parameters. A fine-tuned neural network topology does not mean the network is ready for use - it is only a precondition. Before the network can be applied to solve a specific problem, it must be trained to do so. There are three major learning paradigms: supervised learning, unsupervised learning, and reinforcement learning. The choice of learning paradigm, like the choice of network topology, depends on the type of problem being solved. Although these learning paradigms differ in their underlying principles, they share a common goal: using training data and learning rules (typically defined by a cost function) to enable the neural network to produce appropriate output responses based on given input signals [1].

It is important to note that neurons perform more efficiently when working with smaller values, as this simplifies the computations. Therefore, each neuron aims to pass a smaller value to the next one, which in turn passes an even smaller value to the following neuron, and so on throughout the network. In this process, the value of Y is determined by activation functions (F). An activation function is a mathematical function that compresses the output into a specific range - from 0 to 1, from -1 to 1, or from 0 to Y (depending on the neural network architecture and the type of task). Such functions are Sigmoid, Hyperbolic Tanh, and ReLU.

In the supervised learning process, where we have a predefined dataset, the loss, that is, the neuron's classification error, is recorded. This loss is calculated using the following formula:

$$L = \frac{1}{n}\sum_{i=1}^{n}\left(Ypred_i - Ytrue_i\right)^2,$$

where n is the number of rows in the dataset, $Ypred$ is the value predicted by the neuron, and $Ytrue$ is the actual value in the given dataset. The described process in machine learning is called **feedforward**.

The goal of machine learning is to adjust the neuron's weights (w_i) and bias (b) in such a way that the predicted output is close to the actual output, and the loss value approaches zero. In order to determine how each weight and bias should be changed, the loss function must be differentiated with respect to w_1, w_2, and b, and the resulting function must be optimized, i.e., its minimum must be found. This process is called gradient descent. To ensure that the minimum point is not missed, a parameter called the learning rate is used, which reduces the step size. The values obtained precisely indicate how to calculate the new weights and bias.

After completing this process, the training cycle must be repeated, taking into account the newly calculated weights and bias, to re-evaluate the value of the loss function. If it is still high, the entire process starts again and continues until the loss L approaches zero.

There are different types of neural networks, for example, Convolutional Neural Networks (CNN), which are discussed in papers [1], [2], [3], and [4]. CNNs are used for processing graphical data (images) .

Recurrent Neural Networks (RNN) and their improved version - Long Short-Term Memory (LSTM) [8], are discussed in papers [1], [5], [6], and [7]. These models are designed for processing time-dependent data such as text, speech, time series, music.

In paper [9], the BLIP AI model is discussed. Unlike other models, it does not use CNNs or LSTMs; instead, it employs models based on the transformer architecture, specifically, a Vision Transformer and a BERT-like transformer. This architecture uses a self-attention mechanism, which enables parallelism and the identification of contextual dependencies, making the process faster, more accurate, and higher in quality.

The BLIP model is currently capable of generating image descriptions only in English. The goal of our research is to replace the language transformer with a version fine-tuned for the Georgian language so that it can generate image descriptions in Georgian.

B. captioning with Cnn and lstm

The model consists of three main components:

- **Encoder**: encodes input $x$into hidden states $H = \{h_t\}$and context vector c. Supports e.g. CNN encoders (for images) or RNN encoders (for sequences).
- **Reviewer**: performs $T_r$review steps over encoder hidden states H, producing a sequence of *thought vectors* $\{f_1, f_2, \dots, f_{T_r}\}$. Each review step uses attention over $H$and an LSTM to update.
    - Two variants: *Attentive Input Reviewer* (attention input → LSTM) and *Attentive Output Reviewer* (LSTM input zero vector + attention output).
    - Weight-tying variants: either the reviewer LSTM units share weights across steps or each has separate weights (untied).
- **Decoder**: an LSTM decoder that attends over the thought-vector sequence $F = \{f_t\}$. At each decoding step the hidden state attends to Fthen generates a token.

Encoder: Given input x, produce hidden states $H = \{h_1, \dots, h_{T_x}\}$and context vector c. (For example, CNN: use conv features as H; RNN: hidden states).

Reviewer:

$$f_t = g_t(H, f_{t-1})$$

where $g_t$is a modified LSTM with attention over H.

Decoder: At each time step t, hidden state $s_t$is computed via:

$$\tilde{s}_t = \text{att}(F, s_{t-1}), s_t = f''([\tilde{s}_t;\ y_{t-1}], s_{t-1}),$$
$$y_t = \arg\ \max\ \ \text{softmax}(W_y\, s_t)$$

where F= thought vectors, y_{t-1}previous token embedding.

Loss function: Combination of generative (negative log-likelihood) and discriminative supervision:

$$L(x, y) = \frac{1}{T_y}\sum_{t=1}^{T_y} \ -\log\ P(y_t \mid y_{<t}, x)\ +\ \lambda\, L_d$$

where $L_d$is multi-label margin loss for word-occurrence prediction from thought vectors. [10]

II. HOW THE RESEARCH TRANSLATES INTO PRACTICAL INNOVATION

1. **Model Integration.**
   This work represents one of the first attempts to connect **BLIP-2's** semantic tensor outputs with the **ByT5** language channels specifically adapted for Georgian. This synthesis forms a bridge between visual and linguistic representations that had not previously been explored for the language.[11]
2. **Byte-level language processing**
   The ByT5 transformer uses an unusual architecture that bypasses traditional tokenization.
   This allows it to generate fluent Georgian text directly at the byte level, which avoids many of the segmentation problems that plague other systems. [12]

3. **Dataset Creation**
   To support the training and subsequent evaluation of the models, a specialized image caption database was created, covering a wide thematic and stylistic spectrum — from everyday scenes to culturally

specific contexts. To our knowledge, this is the first large-scale resource that has been purposefully created for Georgian visual captions. The development of this database involves preserving both visual and linguistic diversity so that the model has the ability to remember and correctly perceive the natural nuances of Georgian syntax and semantics. The database is constantly updated and is open to researchers interested in developing multimodal AI systems in low-resource languages.

4. **Practical applications**
   The "Martha" technology has significant potential in a context much broader than research purposes. It can be used both in the public sector, where image recognition and text generation systems will contribute to the efficiency of data processing, and in the private sector, for example, in the media and design industries, where visual descriptions created in the Georgian language will open the way to new types of services. In addition, this technology can be integrated into digital platforms for public services, which will allow citizens to benefit from access to more accurate, context-sensitive and multilingual artificial intelligence tools.
5. **Social impact**
   The integration of high-level technologies into the Georgian language environment is not just a technical achievement - it is an important step towards social inclusion. Such initiatives contribute to the broader idea that linguistic diversity should be reflected in technological development. Georgian, as a language with relatively few resources, is thus gaining a voice in the digital space. At the same time, the project contributes to improving accessibility for people with disabilities, especially in the direction of visual perception and textual information transformation. Ultimately, "Martha" shows that the development of artificial intelligence can become not only a technological, but also a culturally and socially unifying force.

A. Why the Results Matter

The results of this research are expected to have an impact both within Georgia and beyond its borders.

1. National significance

The main goal of the project is to support the establishment and development of an artificial intelligence ecosystem for the Georgian language in Georgia. This involves uniting universities, research centers, and startups around a common goal — the creation of an infrastructure that is based on local realities and Georgian linguistic data. The "Martha" initiative contributes to the deepening of a culture of collaboration in academic and technological circles, while at the same time creating a basis for the development of new Georgian databases and applications that are tailored to the educational, cultural, and technological needs of the country.

2. Regional significance

In the context of the South Caucasus, "Martha" can become an innovative example for communities whose languages face similar challenges — a small number of linguistic resources, limited access to data, and a lack of technological infrastructure. The project demonstrates that, with targeted investment and interdisciplinary collaboration, even small languages can find their place in the global AI ecosystem. Thus, MARTHA can become a regional platform for sharing experiences and encouraging joint research — not only for Georgia, but also for the scientific and technological communities of neighboring countries.

3. Global significance

At a global level, this research shows that the development of AI systems that perceive and produce underrepresented languages is not only possible, but also necessary — both for promoting linguistic equality and for protecting and preserving cultural diversity. MARTHA demonstrates in this context that technological progress can serve not only efficiency and innovation, but also cultural pluralism.

At its core, "Martha" seeks to combine visual perception and natural language generation so that machines can describe scenes in Georgian — accurately, naturally, and in a way that mimics human intuition. Its broader goal is to integrate the Georgian language into the growing global space of artificial intelligence research and technology. In doing so, the project fosters language-based innovation, develops the country's digital autonomy, and establishes Georgia's place on the map of modern technological research.

B. Research Methodology and Design

The "Martha" project is based on an integrated framework that combines computer vision and natural language processing (NLP) approaches to enable the generation of descriptive texts in Georgian. The overall architecture of the system follows a classical encoder-decoder design, where a visual module extracts high-level semantic representations from images, while a linguistic module transforms these representations into naturally formed Georgian sentences.

The visual encoder is based on the BLIP-2 model, which is distinguished by its ability to "analyze" image content efficiently and at a general level. The linguistic decoder — ByT5 — functions as a text generator that transforms these visual signals into grammatically correct and contextually relevant Georgian sentences. The system architecture aims not only to describe objects, but also to convey semantic depth and natural linguistic intuition.

The research methodology was organized into several sequential phases: Data preparation; Model adaptation; Component integration; System evaluation.

The goal of each phase was to consistently refine and synchronize the interaction between the visual and linguistic modules.

1. Data collection and pre-processing

To create a solid training base, the team created a dataset of Georgian image captions that combines existing multilingual corpora with newly collected, locally sourced visual and textual material. Each text underwent a multi-step pre-processing process:

Orthographic and morphological corrections were performed. The data was cleaned of semantic inconsistencies, and the images were standardized in resolution and format to maintain data consistency.

This process is critically important for low-resource languages, where the quality of the data often determines the overall result of the model.

2. Model adaptation

At this stage, the BLIP-2 qFormer , qFormer projection and query tokens were retrained on multilingual image-text data to better understand the specifics of the Georgian environment. In parallel, the ByT5 decoder was specially adapted to the morphological and syntactic features of the Georgian language.

ByT5's byte-level architecture has proven particularly effective in languages like Georgian, where rich suffixation and flexible word order often pose challenges to traditional tokenization. This approach allowed the system to avoid the complexities of tokenization and generate text more naturally, with dramatically reduced formal errors.

3. System Synchronization

After both modules were independently refined, they were combined into a single encoder-decoder pipeline. This stage required precise alignment of the tensor representations so that the semantic information extracted by BLIP-2 could be smoothly and efficiently transferred to the ByT5 decoder.

Additional adjustment mechanisms enabled the generation of text that not only accurately reflects the content of the image, but also is contextually logical and stylistically sounds natural in the Georgian language. The final model demonstrated that the synthesis of visual perception and language generation is possible even when dealing with a language with few resources — which represents an important step towards Georgian artificial intelligence research and technological independence.

C. Alignment of Methodology with Research Objectives

This methodological framework provides a solid foundation for the "Martha" project and serves its main goal — to develop a model that can describe visual scenes in Georgian accurately, contextually, and naturally. The multi-stage design combines both scientific rigor and technical soundness with direct practical applicability.

The approach is based on the integration of basic models, quantitative evaluation, and systematic experiments aimed not only at testing theoretical hypotheses important for research, but also at creating a tool that can be used in practice. Such a programmatic focus emphasizes the applied nature of the project — the goal of creating a system that users can truly rely on in both research and practical contexts.

As a result, this approach combines theory and practice, connecting academic research with engineering implementation, and creates the prerequisites for "Martha" to become not only a technically sound, but also a socially and linguistically significant initiative. Thus, the project demonstrates that the development of artificial intelligence in small languages can serve not only innovation, but also cultural preservation and linguistic consolidation.

D. Research Limitations and Key Advantages

Limitations

1. Limited Availability of Georgian Data

Currently available Georgian image caption datasets are significantly smaller than corpora of languages considered to have high resources, such as English or Chinese. This circumstance limits the model's access to a variety of linguistic and contextual patterns, which directly affects the system's generalization ability. The lack of relevant data for the Georgian language complicates the model's training process and reduces its accuracy in tasks that require a deep understanding of the context.

2. High computational requirements

Training the BLIP-2 and ByT5 models requires significant computational resources, especially GPU power and RAM. Such requirements create a practical barrier for many research organizations and independent researchers who do not have high-end infrastructure. As a result, the system becomes difficult to replicate or modify, which hinders the wider dissemination and further development of the technology.

3. Morphological complexity of the Georgian language

The Georgian language is distinguished by its rich morphological system — plural inflections, agglutinative forms, and flexible word order often complicate accurate text generation. For this reason, generated sentences may be semantically correct, but grammatically or stylistically imperfect. Ensuring linguistic accuracy requires additional refinement, special evaluation mechanisms, and the involvement of human experts in the model evaluation phase.

Advantages and strengths

1. ByT5 token-free processing

The ByT5 model architecture is based on the principle of byte-level processing, which allows text analysis and generation without tokenization. This is especially important for Georgian, where tokenization often makes it difficult to accurately recognize morphological forms. This approach allowed the system to naturally process the Georgian alphabet and produce dynamic, idiomatic, and context-sensitive text, which is rare in low-resource languages.

2. BLIP-2's generic visual encoding

BLIP-2 provides semantically rich and multi-layered representations of images, which makes it possible to use the model not only for caption generation, but also for other multimodal tasks — for example, semantic search, visual question comprehension, or image-based translation. Such generic encoding creates a solid foundation for interdisciplinary experiments and further research. [13]

3. Multi-stage evaluation mechanism

The project implements a step-by-step testing framework that includes both automated metrics (BLEU, METEOR, CIDEr) and human evaluations. This mechanism ensures empirical validation of the results, increases the reproducibility of experiments, and strengthens the reliability of the system in real environments.

4. Social and inclusive benefits

The development of "Martha" goes beyond technological innovation — the project significantly contributes to digital equality and linguistic inclusion. The adaptation of the system to different needs, including those of people

with disabilities, emphasizes the social value of the research. Thus, "Martha" is an example of how artificial intelligence can lead not only to technological but also to social change.

E. Anticipated Risks, Challenges, and Mitigation Strategies

1. Data Scarcity and Quality Control

Risk: The limited amount of high-quality Georgian image caption data can significantly hinder the model's ability to effectively generalize across different visual and linguistic contexts. Data scarcity increases the likelihood of overfitting or overtraining and limits the accuracy of the model in real-world settings.

Mitigation Strategy: The project will leverage existing multilingual databases such as COCO Captions, Flickr30k, and other open-source resources, with additional integration of Georgian captions. New material will be generated from local sources, and texts will be verified by experienced linguists to ensure orthographic, semantic, and stylistic quality. In this way, the database will gradually become more diverse and sustainable.

2. Computational Constraints

Risk: Training BLIP-2 and ByT5 models requires significant computational power, which increases both financial and technical costs. Resource constraints may hinder timely implementation of experiments and full-scale testing of the system.

Mitigation Strategy: To increase the efficiency of model training, cloud infrastructure (e.g., Google Cloud, AWS, or Azure ML) will be used to evenly distribute the computational load. To optimize performance, LoRA (Low-Rank Adaptation) and quantization methods will be implemented, which reduce the size of model parameters and computational costs without losing accuracy [14].

3. Linguistic and architectural challenges

Risk: The complex morphological structure and diverse grammatical forms of the Georgian language may affect the clarity and grammatical accuracy of the generated text, which in some cases will lead to semantic ambiguity or loss of context.

Mitigation strategy: Targeted linguistic tuning will be implemented — additional training on specially selected subsections of the Georgian corpus (high-frequency suffix forms, neologisms, and semantically complex constructions). A qualitative assessment by human experts is also planned, which will help maintain grammatical and semantic accuracy.

4. Ensuring inclusive functionality

Risk: Testing the system's accessibility for people with disabilities involves both logistical and ethical challenges. Inadequate informed consent or technical disruptions may compromise the credibility of the study.

Implementation Strategy: The project will collaborate with inclusion centers and relevant NGOs to ensure informed consent of participants and full compliance with ethical standards. Adapted testing protocols and simple user interfaces will be developed, designed for people with sensory or physical disabilities.

5. Financial and time pressures

Risk: High computational requirements increase costs and create a risk that the project will not be able to complete within the planned timeframe. Financial constraints also limit the ability to obtain new data sources and technical support.

Mitigation Strategy: A phased budgeting approach will be implemented, allowing the project to flexibly allocate resources to each phase. In parallel, external funding will be sought (e.g., research grants, international foundations, academic-corporate partnerships). Parallel tasks will be planned to optimize the workflow, which will help maintain progress within deadlines and continuously produce results.

III. MATHEMATICAL MODELING OF THE MARTHA MODEL

In recent years, artificial intelligence (AI) has developed at an exceptionally fast pace, creating new opportunities in many fields of science, technology and industry. One of the most dynamic areas has become the interaction of computer vision (CV) and natural language processing (NLP) — two disciplines that, through their integration, allow systems to analyze visual information and based on it generate text in human language. Such hybrid systems are already widely used today in areas such as autonomous transportation, assistive robotics, security technologies and inclusive communication platforms, which are gaining particular importance in the context of digital accessibility and social integration.

However, progress in this rapidly growing field is mainly concentrated on the world's leading languages, especially English, which effectively leaves smaller language communities and cultures in the shade. The lack of interest in low-resource languages has led to a linguistic disparity in technological innovation, with high-accuracy systems available only for languages with large corpora and rich databases.

In this context, the Georgian language represents one of the most significant challenges. Its complex morphological structure, limited digital resources, irregular orthography, and rich lexical diversity make it difficult to directly adapt standard AI models. As a result, Georgian is rarely included in the list of languages for which natural, contextually accurate descriptions of images or visual scenes are possible.

It is in response to these challenges that the "Martha" project was created — an initiative aimed at creating a visual description system specifically tailored for the Georgian language. The system is based on an encoder-decoder architecture that combines two modern models:

BLIP-2 — for visual coding, which extracts semantically rich representations from images;

ByT5 — to generate text that transforms these representations into coherent, naturally formed Georgian descriptions.

The innovation of "Martha" lies precisely in this interconnection: BLIP-2 acts as an image "cognitive" module that identifies objects and their relationships, while ByT5 functions as a linguistic "narrator" that transforms these visual signals into grammatically correct and contextually consistent Georgian text.

In this way, "Martha" not only solves a technological problem, but also represents an important step towards linguistic inclusion and cultural self-representation — an attempt to give the Georgian language a worthy place in the modern artificial intelligence ecosystem.

$$z = f_\theta(x), y = g_\phi(z)$$

where:

- $x \in \mathbb{R}^{H\times W\times 3}$: input RGB image
- f_θ: BLIP2 Encoder
- $z \in \mathbb{R}^d$: semantic (visual) vector
- g_ϕ: ByT5 Decoder
- $y = (y_1, y_2, \dots, y_T)$: generated Georgian text

BLIP2 uses a **Vision Transformer (ViT)** and a **Q-Former**:

- The image $x$is divided into patches $p_i \in \mathbb{R}^{P\times P\times 3}$.
- ViT encodes each patch into a vector $h_i = E_{ViT}(p_i)$.
- Q-Former applies attention over patch embeddings to produce semantic tokens:

$$z_j = \text{Attention}(q_j, \{h_1, \dots, h_N\})$$

producing a tensor $Z \in \mathbb{R}^{M\times d}$.

Because BLIP2's output dimension (d) differs from ByT5's input dimension (d'), a projection is applied:

$$h_j = Wz_j + b, W \in R^{d'\times d},\ b \in R^d$$

Resulting in:

$$H = (h_1, h_2, \dots, h_M), H \in \mathbb{R}^{M\times d'}$$

ByT5 is an **autoregressive Transformer decoder** that models text generation as:

$$P(y_t \mid y_{<t}, H) = \text{softmax}(W_o s_t)$$

where $s_t$is the decoder hidden state and $H$the encoder output.

Text is generated autoregressively:

$$y_t = \arg\max P(y_t \mid y_{<t}, H)$$

The objective is to maximize the probability of generating the correct caption:

$$L(\theta, \phi, W, b) = -\sum_{t=1}^{T} log\ P(y_t \mid y_{<t}, H)$$

This is a **cross-entropy loss**, ignoring padding tokens and optionally using **label smoothing**:

$$L_{smooth} = (1 - \epsilon)L_{CE} + \epsilon U$$

Regularization Techniques

- **Dropout:** randomly deactivates neurons

$$h' = (m \odot h)/p, m \sim \text{Bernoulli}(p)$$

- **Weight Decay:** penalizes large parameter norms

$$L_{final} = L + \lambda \parallel \theta \parallel_2^2$$

- **Gradient Clipping:** prevents gradient explosion

$$g' = g/\max(1, \parallel g \parallel/\tau)$$

Optimization
AdamW optimizer is used:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\ \theta_{t+1} &= \theta_t - \eta \frac{m_t}{\sqrt{v_t} + \epsilon} - \lambda \eta \theta_t \end{aligned}$$

Learning Rate Scheduling
A **ReduceLROnPlateau** scheduler is employed:

$$\eta_{t+1} = \begin{cases} \eta_t \cdot \gamma, & \text{if no improvement for } N \text{ epochs} \\ \eta_t, & \text{otherwise} \end{cases}$$

Interpretation

- Through this modeling approach:
- The encoder learns a compact representation of the visual world.
- The projection layer aligns visual and linguistic latent spaces.
- The decoder generates linguistically accurate and contextually relevant Georgian descriptions

A. Experimental Setup and Evaluation Framework
Dataset Structure

To support the training of the **Martha** model, a dedicated **Georgian image–caption dataset** of approximately **300,000 images** was compiled. The collection spans a wide variety of visual contexts, including everyday life scenes, educational and scientific imagery, cultural and historical artifacts, tourist and natural landscapes, and diverse forms of social interaction.

This range of subject matter was designed to strengthen the model's ability to generalize across real-world scenarios and linguistic domains. [15][16][17]

Data Splitting

The dataset was divided into three subsets to facilitate systematic model development and evaluation:

- **Training set (90%, ~270,000 images):** used for model learning and parameter optimization.
- **Validation set (10%, ~30,000 images):** employed to monitor performance and mitigate overfitting during training.

Data Preprocessing

The image data was processed to a uniform resolution of 224×224 pixels to ensure compatibility with the structural requirements of the architecture (BLIP-2) and the computational operations policy. The color channels (RGB) were normalized using mean values and standard deviations, which ensures a consistent distribution of colors and reduces the sensitivity of the model to changes in illumination.

The grammatical annotations underwent orthographic and grammatical normalization to maintain linguistic consistency and accuracy across the entire dataset. These were controlled by key linguists, morphological constraints, and the exclusion of semantically inaccurate or incomplete sentences. The final dataset contained only those annotations that satisfied a clear grammatical structure and natural syntax.

Data Augmentation

To increase the robustness of the model and overfitting, this approach aimed to develop the model's flexibility in visual perception and generalization ability. The operational tools used were:

- random resizing in the range of 0.8–1.0 to give the model different distances and framing characteristics;
- horizontal flipping with a probability of 0.5, which helped to develop robustness to the positional relationships of objects;
- color jittering, which changed brightness, contrast, and hue, thereby providing adaptation to the conditions.

Only replacement and center reduction are performed on the validation and testing subsets (which was intended to protect the evaluation process and maintain repeatability). As a result, the evaluation phase was carried out in a controlled manner, ensuring comparability across experiments.

IV. CONCLUSION

The Martha project is the first initiative in Georgia,
which aims to create a visual environment description system based on a modern encoder-decoder architecture specifically for the Georgian language.

“Martha” is the first initiative in Georgia to create a visual description system based on a modern encoder-decoder architecture, specifically for the Georgian language. The main idea of the project is to combine computer vision and natural language generation using artificial intelligence so that the machine can describe visual scenes in its native Georgian language — accurately, naturally, and contextually.

Experimental Framework

The methodological design of the project combines several interconnected components that consistently create a full-fledged multimodal system:

1. Database Creation

A collection of approximately 300,000 images was created, each of which is paired with a descriptive text created in the Georgian language. This corpus is the main basis for training and further evaluating the model. Both open international corpora and locally collected visual material were used in the data selection process, ensuring linguistic and cultural diversity.

2. Visual encoding

The BLIP-2 encoder is used for image processing, which extracts high-level semantic representations from the image. The model learns not only to identify objects, but also to understand the contextual and functional relationships between them, which forms the basis for moving on to deep description generation.

3. Language decoding

The ByT5 decoder transforms these semantic representations into natural Georgian text, generating fluent, grammatically correct, and semantically accurate sentences. Its byte-based architecture allows the system to avoid tokenization problems and effectively work with the complex morphology of the Georgian alphabet.

4. Model alignment

A special projection layer has been implemented for module integration, which synchronizes the visual and linguistic embedding spaces. In this way, the semantic information obtained by BLIP-2 is accurately transferred to ByT5, ensuring the content and stylistic relevance of the text.

5. Optimization and Training

Regularization techniques and additional phases of parameter refinement were used in the training process, which ensured stable convergence and maximized accuracy. The model was trained in the Google Colab Pro+ environment, with NVIDIA A100 GPU (80 GB) of memory. Experiments have shown that the system can describe visual scenes in Georgian — semantically accurate, stylistically natural, and while preserving human dynamics.

Extended Impact

The significance of the project goes beyond technical innovation and encompasses social, cultural, and educational aspects:

- Linguistic Inclusivity — For the first time, the Georgian language is integrated into a high-level visual-linguistic model, which is an important step towards linguistic equality and the development of multilingual artificial intelligence.
- Inclusive Technology — The system creates new opportunities for increasing accessibility, including for people with visual impairments, through the automatic generation of audio descriptions. This step contributes to digital equality and the reduction of social barriers.
- Education and Research — "Martha" can become a valuable resource for academic institutions — especially in the fields of linguistics, computer vision, AI ethics, and technological humanities.
- Cultural Heritage — The system contributes to strengthening the digital cultural identity of Georgia, as it can generate descriptive texts for museums, archives and visual heritage collections. This capability is particularly important in the context of cultural heritage protection and tourism promotion.

Future Directions

Based on the project results, several strategic directions for further development were identified:

- Multimodal Extension: Expanding the functionality of the system to video and audio data, which will allow the model to analyze dynamic scenes and create real-time subtitles or audio descriptions.
- Multilingual Adaptation: Expanding the model training to neighboring languages — Armenian, Azerbaijani, Ossetian — and experiments with Georgian-English code switching, which will contribute to deepening regional integration.
- Real-time description technology: Developing a low-latency version for assistive devices (e.g., smart glasses) to provide visually impaired people with live scene descriptions.
- Retrieval-Augmented Generation (RAG): Enriching the database with cultural-historical context—for example, integrating information about historical monuments and landmarks, which will enhance the depth and relevance of text knowledge [18].

Performance optimization

Further truncation of the model and implementation of lightweight training strategies to enable it to be deployed on edge devices—increasing practical usability and energy efficiency.

REFERENCES

[1] Andrej Krenker, Janez Bešter and Andrej Kos. Introduction to the Artificial Neural Networks. Artificial Neural Networks - Methodological Advances and Biomedical Applications, edited by Prof. Kenji Suzuki, ISBN 978-953-307-243-2 Hard cover, 362 pages (2011).

[2] Anirudha Ghosh, Abu Sufian, Farhana Sultana, Amlan Chakrabarti & Debashis De. Fundamental Concepts of Convolutional Neural Network. Journal Springer Nature "Recent Trends and Advances in Artificial Intelligence and Internet of Things" (2020), DOI: 10.1007/978-3-030-32644-9_36.

[3] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In:Computer Vision–ECCV 2014, pp. 818–833. Springer (2014).

[4] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Largescale video classification with convolutional neural networks. In: Computer Visionand Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1725–1732. IEEE (2014).
[5] Zheng, Shuai, et al. "Conditional random fields as recurrent neural networks." Proceedings of the IEEE International Conference on Computer Vision (2015).
[6] Chung, Junyoung, et al. "Gated feedback recurrent neural networks." CoRR, abs/1502.02367 (2015).
[7] Lipton, Zachary C., John Berkowitz, and Charles Elkan. "A Critical Review of Recurrent Neural Networks for Sequence Learning."arXiv:1506.00019 [cs], May 29, 2015. http://arxiv.org/abs/1506.00019.
[8] Sepp Hochreiter, Jürgen Schmidhuber. Long Short-Term Memory. Neural Computation 9(8):1735-1780 (1997). DOI: 10.1162/neco.1997.9.8.1735.
[9] Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Computer Vision and Pattern Recognition. DOI: 10.48550/arXiv.2201.12086.
[10] Zhilin Yang, Ye Yuan, Yuexin Wu, Ruslan Salakhutdinov, and William W. Cohen. Review networks for caption generation. In NIPS, 2016.
[11] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.
[12] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, Colin Raffel . ByT5: Towards a token-free future with pre-trained byte-to-byte models. arXiv:2105.13626 https://arxiv.org/abs/2105.13626, 2021
[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.
[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.
[15] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021.
[16] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems, 24, 2011.
[17] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. arXiv:1504.00325, 2015.
[18] Sergo Tsiramua, Hamlet Meladze , Tinatin Davitashvili, Davit Bitmalkishev , Tatia Elbakidze. AI and NLP Models for Q&A in Georgian . CSIT Conference 2025, Yerevan, Armenia. https://doi.org/10.51408/csit2025_11