
Cultural-Aware AI Model for Emotion Recognition

Leveraging Multimodal Insights for Culturally
Sensitive Emotion Analysis

Meet the Research Team

Mehrdad Baradaran

Department of Computer and Data
Sciences, Shahid Beheshti University,
Tehran, Iran
meh.baradaran@mail.sbu.ac.ir

Payam Zohari

Department of Computer
Engineering, Khaje Nasir University
Of Technology,
Tehran, Iran
m.zohari@email.kntu.ac.ir

Abtin Mahyar

High Performance Computing
Laboratory, School of Computer
Science, Institute for Research in
Fundamental Sciences,
Tehran, Iran
abtinmahyar@gmail.com

Hossein Motamednia

High Performance Computing
Laboratory, School of Computer
Science, Institute for Research in
Fundamental Sciences,
Tehran, Iran
h.motamednia@ipm.ir

Payam Zohari

Faculty of Computer Science and
Engineering, Shahid Beheshti
University,
Tehran, Iran
rahmati@sbu.ac.ir

Abtin Mahyar

Department of Computer Engineering,
Chosun University, Gwangju, 61452,
South Korea
gorgin@irost.ir

Table of contents

01

Introduction

Setting the Stage for Culturally-Aware
Emotion Recognition

02

Related Work

Review of Current Approaches and
Methodologies

03

Proposed Method

Our Innovative Approach to Multimodal
Emotion Detection

04

Experiments

Testing and Validating Our Model

05

Results & Discussion

Insights and Implications from Our
Findings

06

Conclusion

Summary and Future Directions



Introduction

The Importance and Challenges of
Emotion Recognition

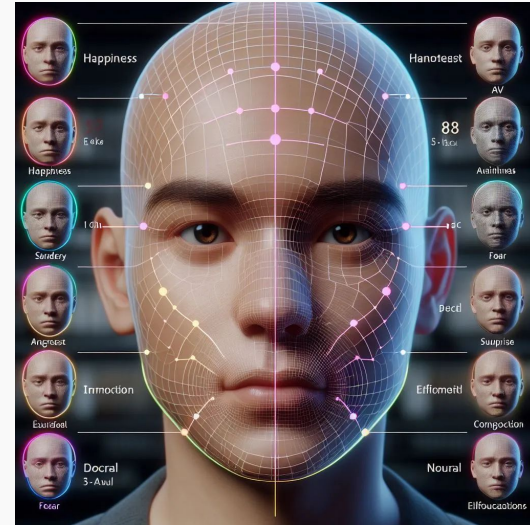
The recognition of emotions is crucial for improving human-computer interaction.

Challenge

Emotions vary across cultures and languages

Objective

Develop a model that integrates cultural awareness into emotion recognition





Related Works

Exploring Existing Research and
Techniques

Related works - Overview

Current Research Landscape: Emotion detection research encompasses various methodologies to understand human emotions from textual and visual data.

Focus Areas:

- Text-Based Approaches: Utilize Natural Language Processing (NLP) techniques for sentiment analysis and emotion classification in textual data.
- Image-Based Approaches: Employ computer vision algorithms, such as Convolutional Neural Networks (CNNs), to analyze facial expressions and other visual cues for emotion recognition.
- Multimodal Approaches: Combine text and image data to achieve a more comprehensive understanding of emotions, leveraging the strengths of both modalities.

Significance of Multilingual Datasets: The use of multilingual datasets is crucial for capturing diverse emotional expressions across different languages and cultures, enhancing the robustness and applicability of emotion recognition systems.

Related works - Backgrounds



Multimodal Approaches

Current studies integrate text-based emotion detection employing NLP techniques and image-based methods using CNNs in computer vision. Some research focuses on intra-modality, analyzing spatial and contextual relationships within images or grammatical and semantic connections within text. Many explore inter-modality, establishing connections between visual elements in images and semantic elements in sentences to deepen emotional comprehension across diverse data modalities. Our proposed multimodal approach innovatively combines both intra-modality and inter-modality techniques, aiming to enhance the comprehensive understanding of emotions.



Multimodal Approaches

Addressing language diversity is crucial in emotion detection, employing techniques such as machine translation and cross-lingual transfer learning. The use of balanced datasets that encompass diverse cultures enhances the accuracy and applicability of emotion recognition systems.

Related Work - Text-based & Multi-source Emotion Detection

Multi-source

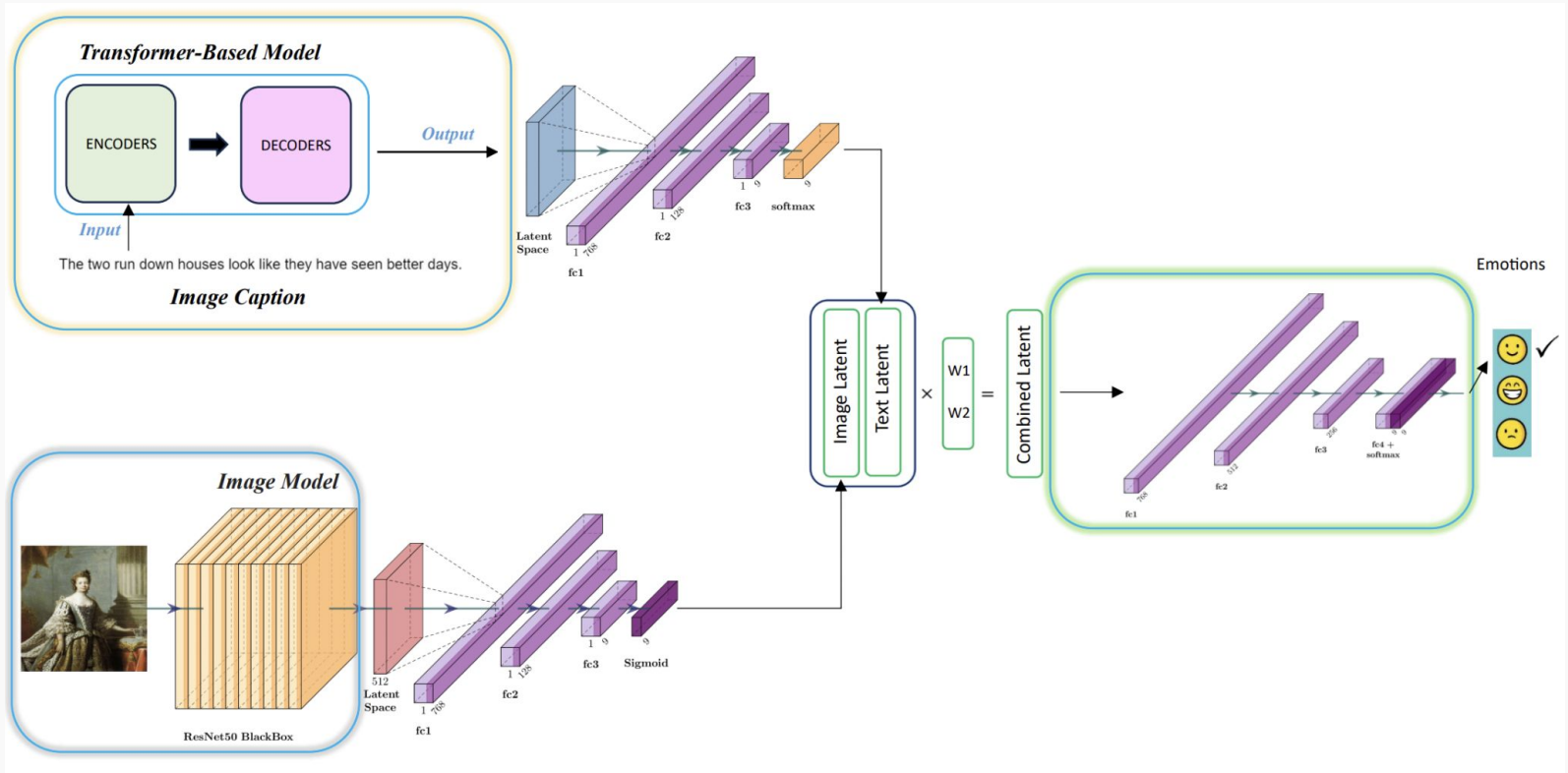
combines text and images for more accurate emotion detection, as demonstrated by models such as CLIP and ArtEmis. Techniques include vision-language transformers and contrastive learning methods to leverage both modalities effectively .



Text-based

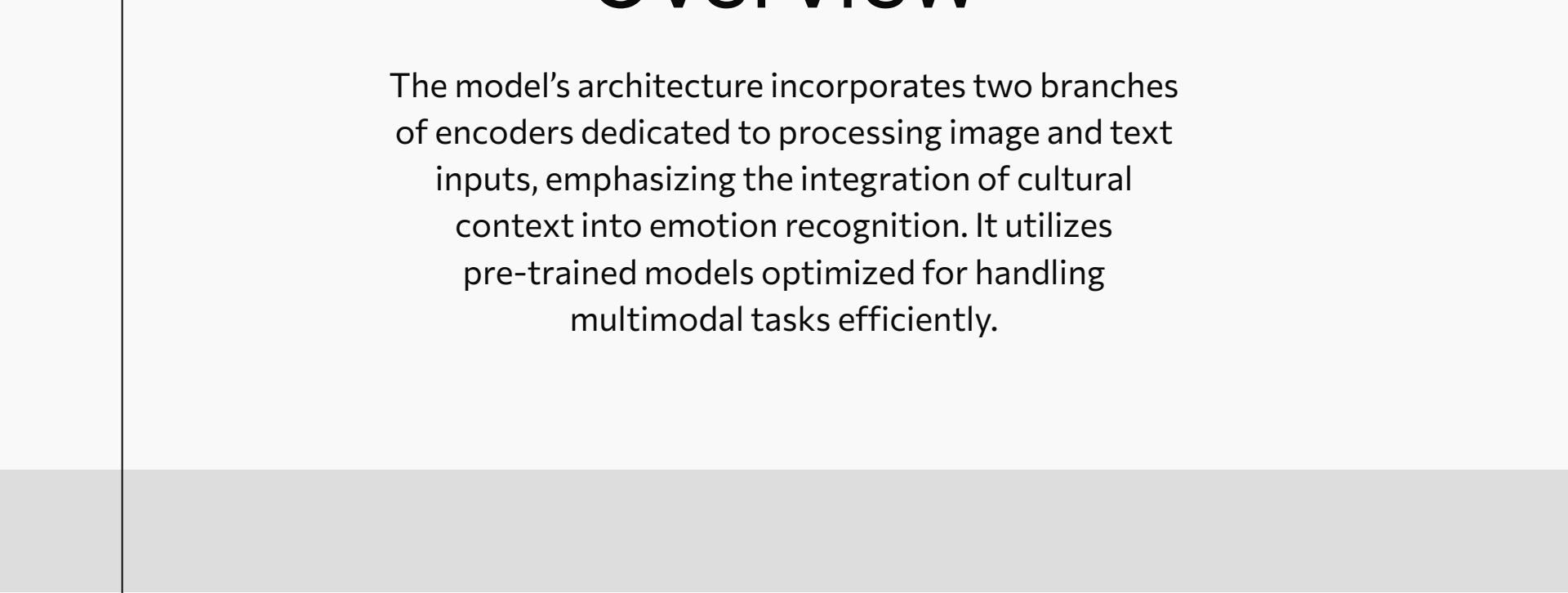
involve semantic analysis of texts, utilizing datasets like SemEval2018 for emotion classification, and employing various machine learning algorithms and deep learning models to enhance accuracy .

Proposed Method



Overview

The model's architecture incorporates two branches of encoders dedicated to processing image and text inputs, emphasizing the integration of cultural context into emotion recognition. It utilizes pre-trained models optimized for handling multimodal tasks efficiently.

A thin vertical black line is positioned on the left side of the slide, extending from the top of the text area to the bottom. A solid gray horizontal bar spans the entire width of the slide at the bottom.

Text Model

- Transformer-based text model for emotion recognition
- Extraction of latent features representing contextual information
- Training specifically for emotion classification

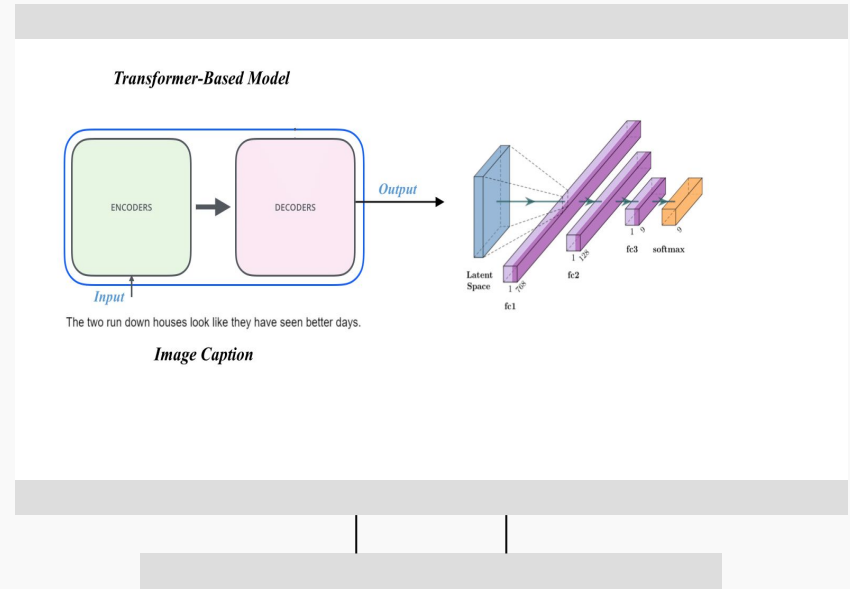
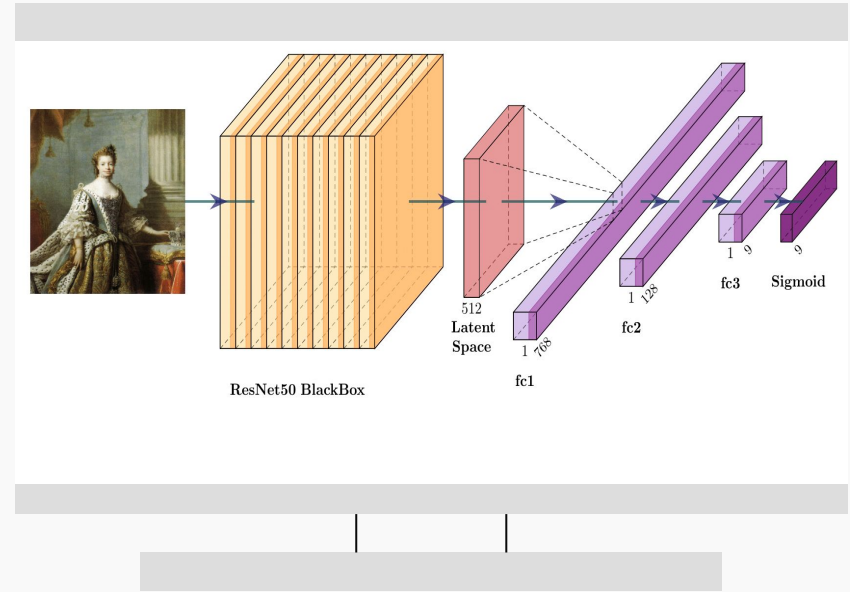


Image Model

- CNN-based image model for emotion recognition
- Feature extraction from visual data
- Adjustment of latent spaces for accurate predictions



Proposed Method - Fusion Approach

- Linear combination technique to merge text and image features
- Final latent representation capturing content from both modalities
- Diagram of the model architecture (refer to Figure 1)



Experiments

Datasets and Training

WIKIART



Construction of an emotional image captioning model, requires the task of image processing as well as emotion recognition. The “WIKIART” dataset, supplying thousands of classified paintings helped us through this journey.

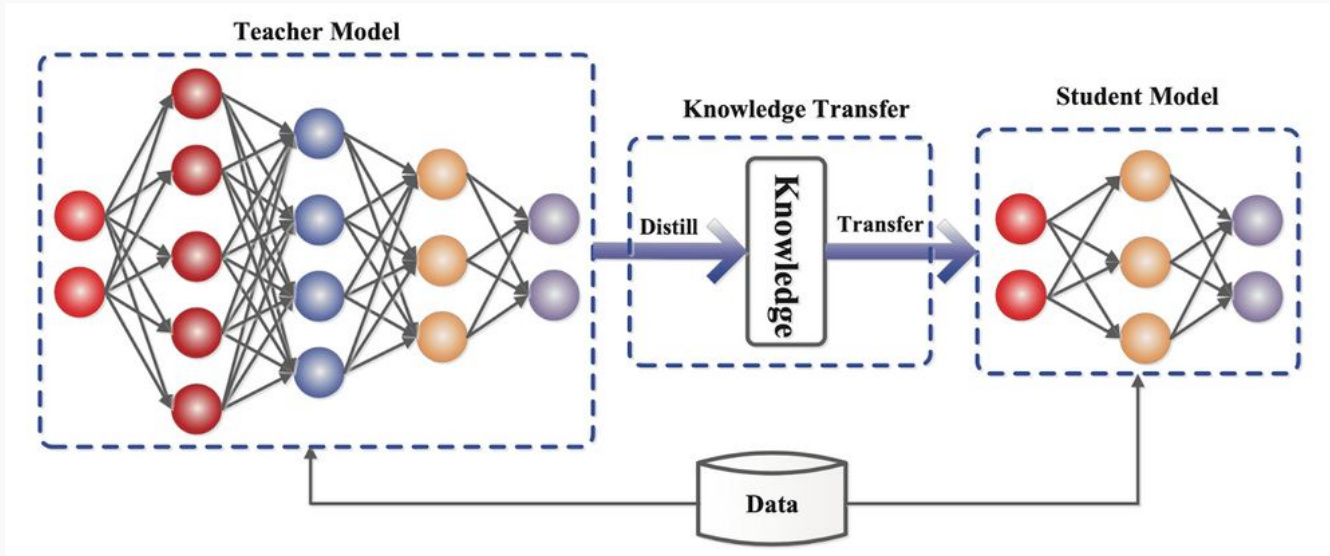
ArtELingo

	COCO	ArtEmis	ArtELingo
Image Source	Photos	WikiArt	WikiArt
#Images	328k	80k	80k
#Annotations	2.5M	0.45M	1.2M
#Annot/Image	7.6	5.68	15.3
Emotions	0	9	9
Languages	E	E	ACES

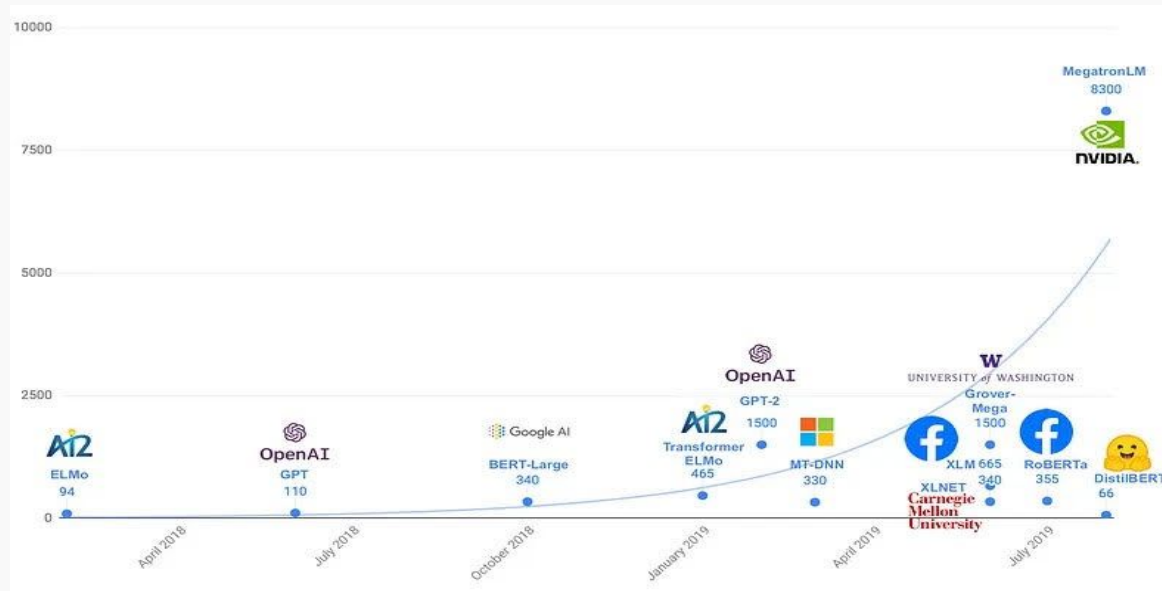
To approach cultural-awareness model, we needed to train our model on a dataset rich in multilingual captions and “ArtELingo” provided us with a collection of 80k annotated artworks in 3 languages, namely English, Arabic, and Chinese.

But why ArtELingo?

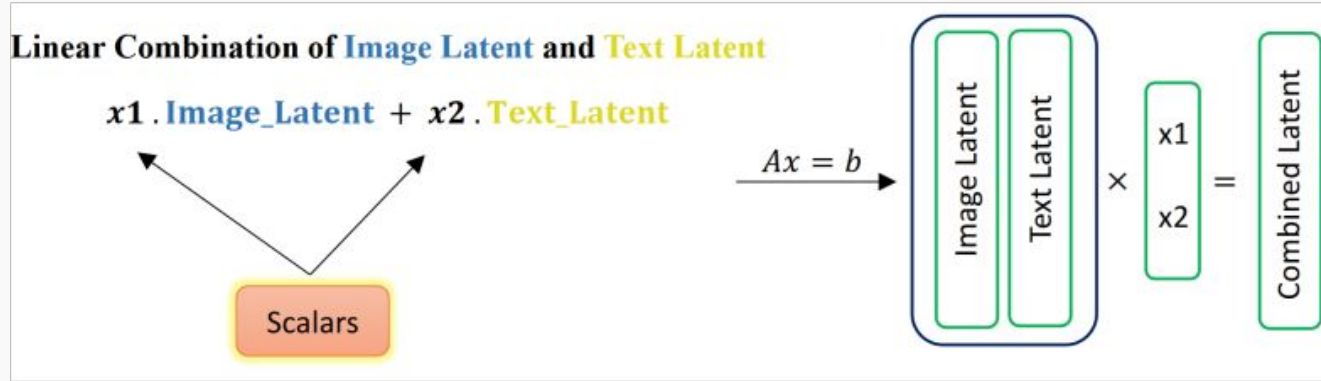
Knowledge Distillation



Why DistilBERT?



Feature Linear Combination





Results

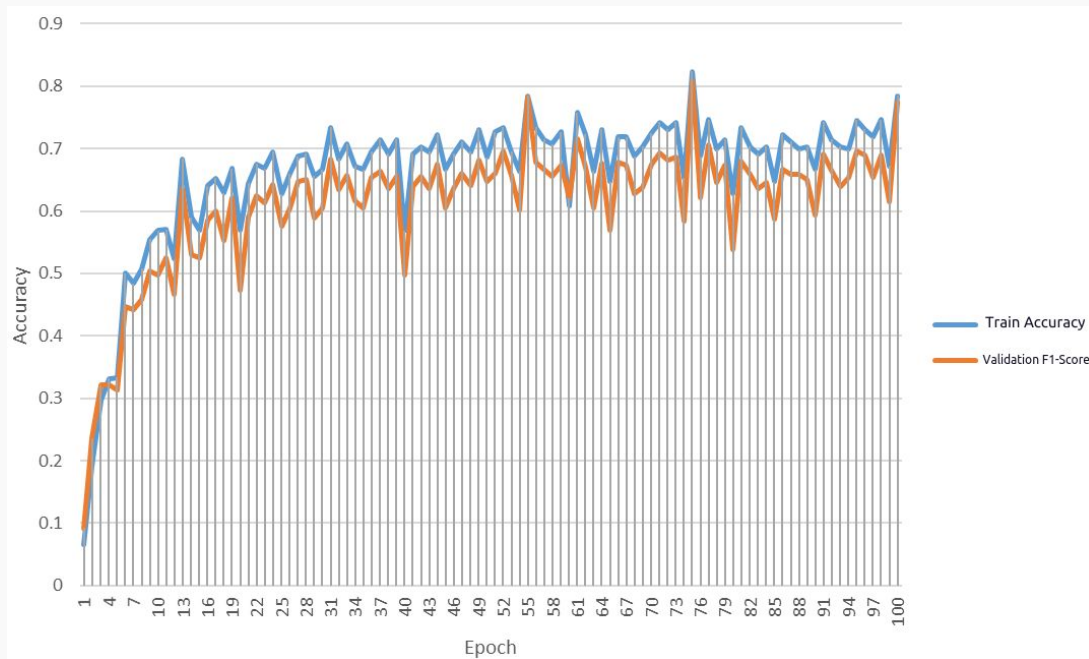
Evaluating Performance and
Discussing Outcomes

Our Results

THE EVALUATION OF THE IMAGE MODELS ON THE ARTELINGO DATASET

Model/Accuracy	Train ACC	Validation Acc
ResNet50	0.711386	0.698935688
VGG	0.717367	0.706483998
VIT	0.702534	0.687801932

Our Results



Conclusion

Key Takeaways and Future
Research Paths

Conclusions

- Successful integration of cultural awareness in multimodal emotion recognition
 - Contributions to human-computer interaction and service development
 - Suggestions for future research: Inclusion of speech, video, physiological signals
-

Thanks!

A decorative graphic consisting of a thick, light gray horizontal bar spanning the width of the slide. Two thin, dark gray vertical lines are positioned on either side of the bar, extending from the top to the bottom of the slide. A thin, dark gray horizontal line is also present at the bottom of the slide, starting from the left edge and extending to the right.