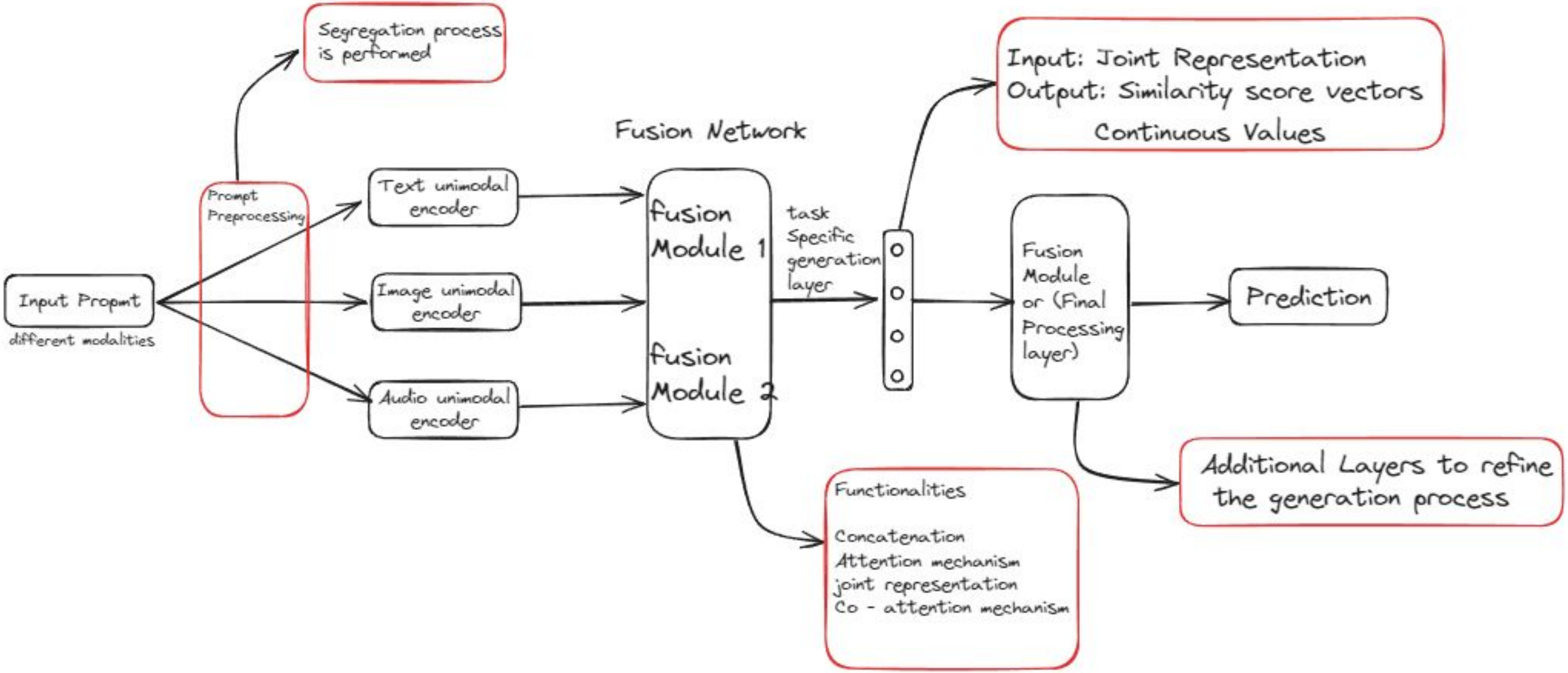
Multimodal Architecture

By : Kaif Shaheem J

Overview of Multimodal Architecture

- Input Segmentation
- Unimodal Encoders
- Fusion Network
- Task Specific Generation Layer
- Final Processing layer
- Prediction

Multimodal Architecture



Unimodal Encoders

⇒ Text Encoders - BERT, GPT, Bag of words TF-IDF

⇒ Image Encoders - CNNs, VGG, ResNet, Vision Transformer

⇒ Audio Encoders - Wav2Vec, CNNs, LSTM

Fusion Network

- Fusion Network Consists of Fusion Modules, which perform the integration of information from different Modalities.
- Fusion Modules can vary in **complexity & design**.
- But **primary goal** is to **combine into unified representation**.
- Integrating information from multiple modalities and create a **joint representation** that captures the multimodal interactions.

Task Specific Generation Layer

Task-specific generation layer may integrate large language models (LLMs) to enhance performance on tasks that involve **natural language understanding, generation, or interaction**.

The **Specific roles of these LLMs** depend on the complexity and requirements of the task.

Task Specific Generation Layer

Types of LLMs

- **BERT** (Bidirectional Encoder Representations from Transformers)
- **GPT** (Generative Pre-trained Transformer)
- **BART** (Bidirectional and Auto-Regressive Transformers)

Task Specific Generation Layer

Roles of LLMs

- Text Understanding.
- Generations.
- Sequence to Sequence tasks.

The number of LLMs present in the task-specific generation layer can vary based on the architecture and task requirements:

- Single LLM (for Text Generations)
- Multiple LLMs (VQA)

Final Processing Layer

Purpose and Functionality

The final processing layer's primary goals are to:

- **Refine**: Apply additional processing to enhance the quality and relevance of the task-specific output.
- **Integrate**: Combine outputs from various parts of the model, if necessary.
- **Normalize**: Ensure that the output is in the correct format and range for the prediction layer.

⇒ Normalization Layers

⇒ Additional Dense Layers

⇒ Activation Functions

Prediction Layer

The prediction layer is the final step in the model, responsible for producing the specific output required for the task (e.g., class labels, generated text).

It maps the refined feature vector to the final output format.

Components,

- Dense (Fully Connected) Layers
- Softmax Layer
- Sigmoid Layer

Final Processing Layer:

- **Function**: Further refine and normalize the output from the task-specific generation layer.
- **Focus**: Ensuring the feature vector is in its optimal form for prediction.
- **Components**: Normalization, dense layers, activation functions, attention mechanisms.

Prediction Layer:

- **Function**: Produce the final output of the model.
- **Focus**: Mapping the refined feature vector to the task-specific output format.
- **Components**: Dense layers, softmax/sigmoid/regression layers, decoder layers

Prediction layer

Example:

Visual Question Answering (VQA)

- **Input**: Joint feature representation from visual and textual data.
- **Components**: Dense layers followed by a softmax layer to predict the answer from a predefined set of possible answers.
- **Output**: Predicted answer as a class label.

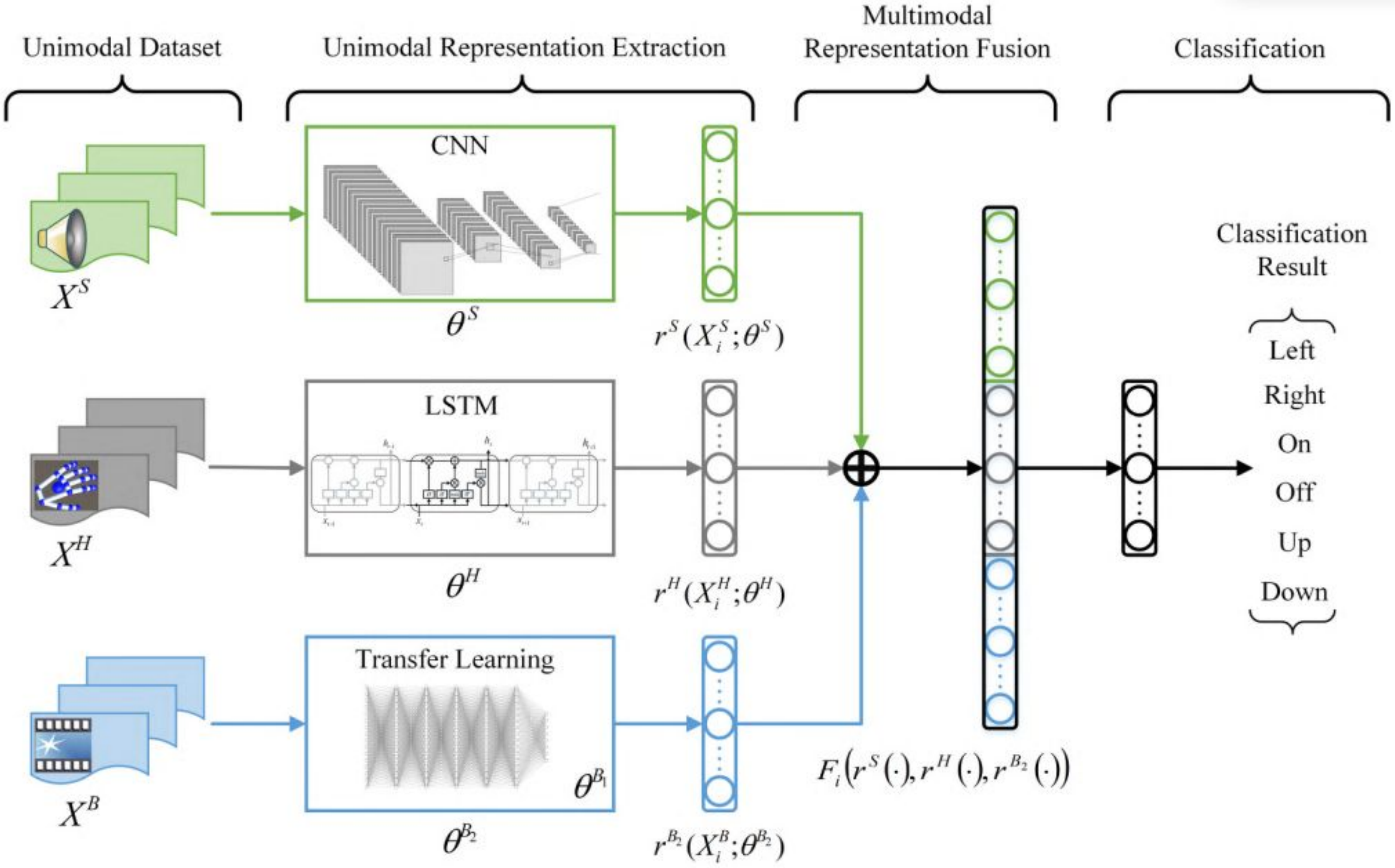
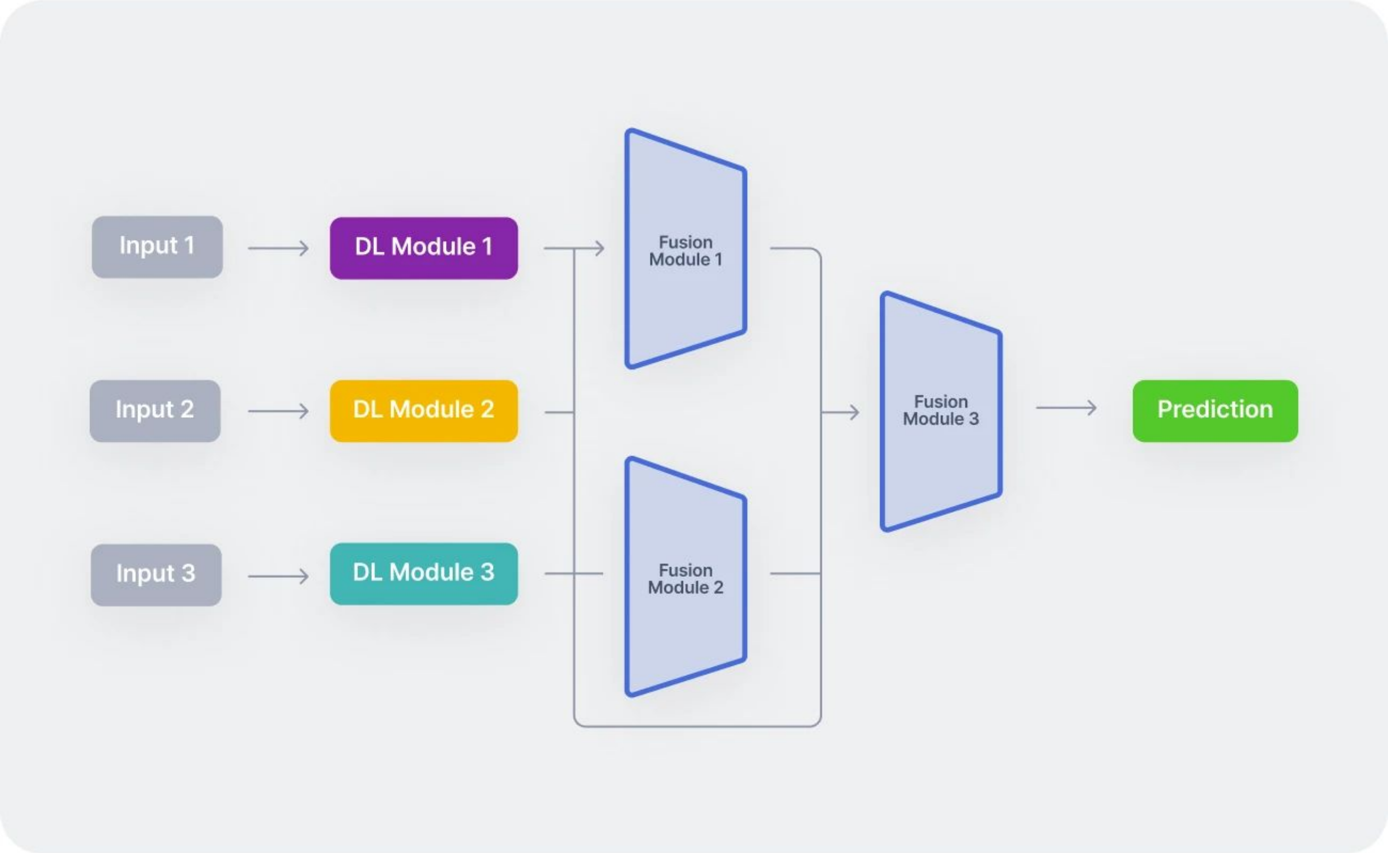


FIGURE 2. Illustration of the proposed multimodal fusion architecture.

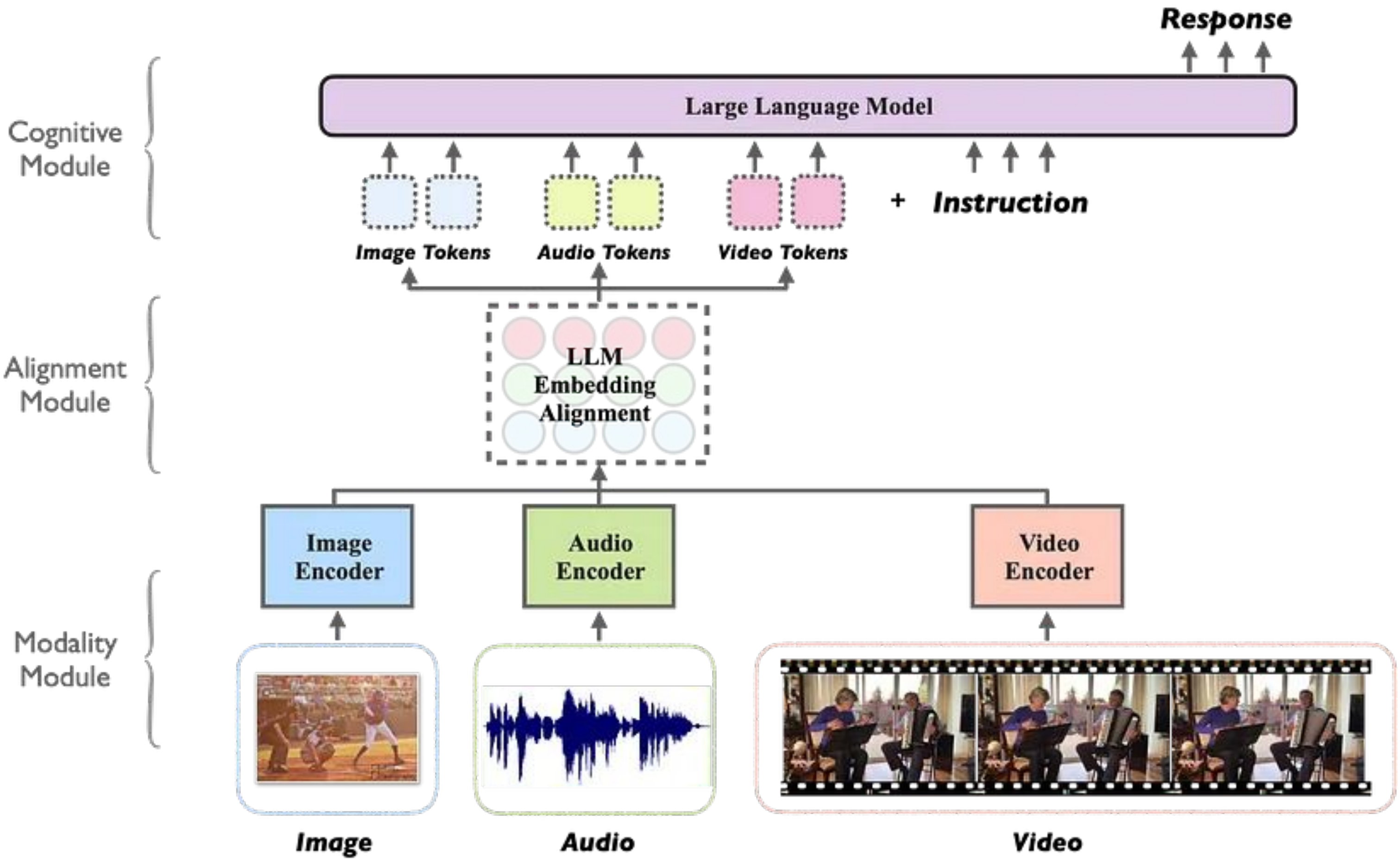

Input 1
DL Module 1
Input 2
DL Module 2
Input 3
DL Module 3
Fusion Module 1
Fusion Module 2
Fusion Module 3
Prediction

V7

Macaw LLM

Macaw-LLM is a multi-modal language model that can generate natural language texts based on the input content from image, audio, video, or text modalities.

Macaw LLM


Response
Large Language Model
Cognitive Module
Image Tokens
Audio Tokens
Video Tokens
+ Instruction
Alignment Module
LLM Embedding Alignment
Image Encoder
Audio Encoder
Video Encoder
Modality Module
Image
Audio
Video

Macaw LLM

Modality Module:

Where existing LLMs primarily focus on processing textual information.

To incorporate additional modalities such as visual and audio data, we integrate extra modality encoders into MACAW-LLM.

This enhancement enables our MACAW-LLM to handle multiple modalities effectively.

Macaw LLM

Alignment Module:

Since each modality encoder is trained independently, the learned representations of different modalities may not be directly compatible.

To address this, we propose the alignment module, which unifies the representations from different modalities, enabling effective integration of multi-modal information.

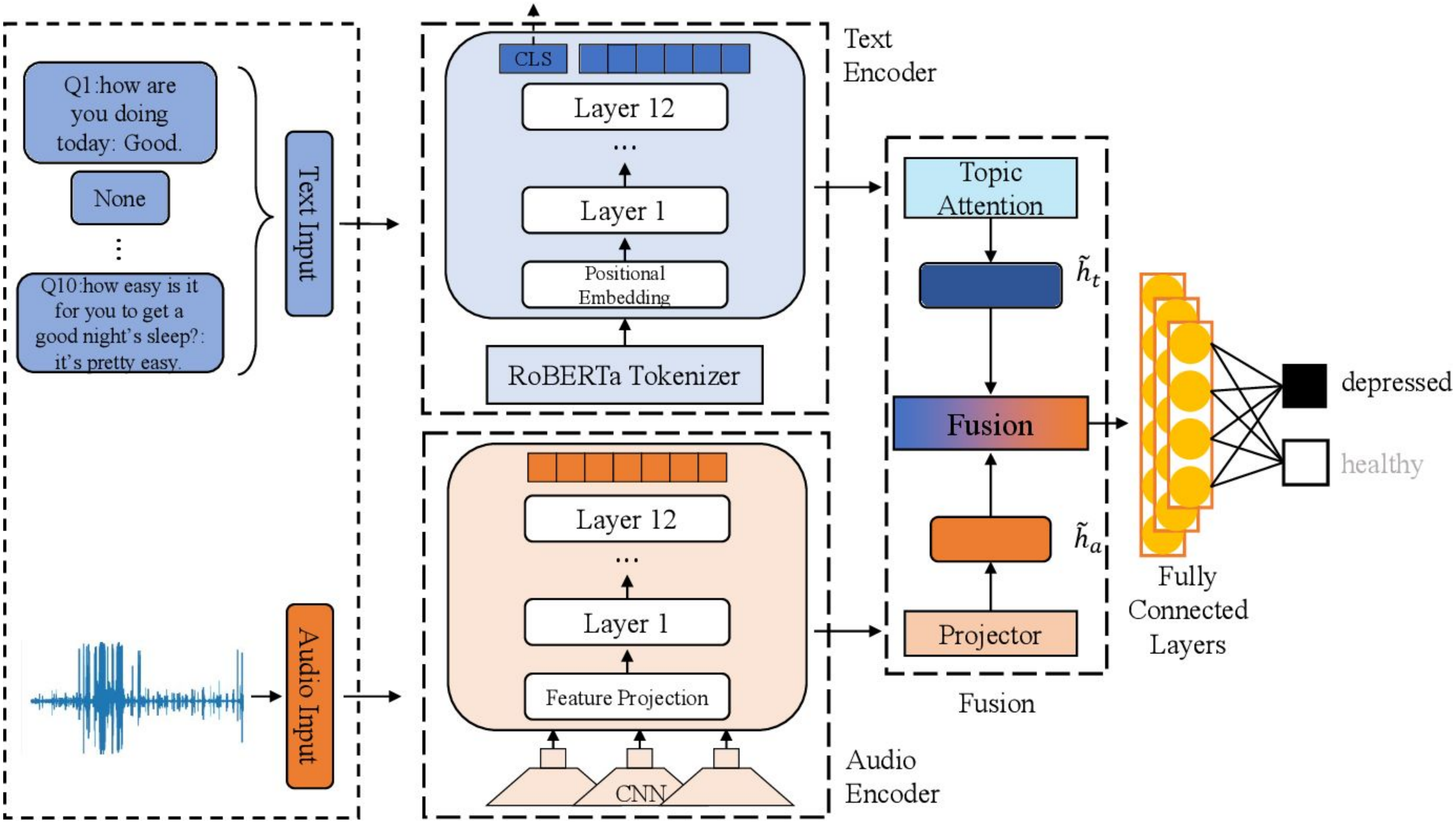
Macaw LLM

Cognitive Module:

In the MACAW-LLM framework, this cognitive module not only serves as the basis for understanding human instructions but also functions as the textual modality encoder.

This means that it encodes textual information into a format that the system can understand and process further.

By using a pre-trained LLM as the cognitive module, MACAW-LLM benefits from the rich representations and **language understanding** abilities of these models, enabling it to **effectively interpret** and **respond to human instructions** or input in various applications or tasks.


Q1:how are you doing today: Good.
None
Q10:how easy is it for you to get a good night's sleep?: it's pretty easy.
Text Input
Audio Input
CLS
Layer 12
...
Layer 1
Positional Embedding
RoBERTa Tokenizer
Text Encoder
Layer 12
...
Layer 1
Feature Projection
CNN
Audio Encoder
Topic Attention
$\tilde{h}_t$
Fusion
$\tilde{h}_a$
Projector
Fusion
Fully Connected Layers
depressed
healthy

Resources

Research paper - Multimodal Deep Learning

Thank You