fit@hcmus
VNUHCM - UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY

University of Science - VNU-HCM
Faculty of Information Science

Multi-modal Machine Comprehension Group

Presenter:

Dr. Le Thanh Tung

MMC Group

fit@hcmus

1. Multi-modal Comprehension
   - Visual Question Answering
   - Vision-language Model
   - Contrastive Learning
   - Youtube Engagement
2. Speech Processing (for Vietnamese) - ntienhuy
   - Text to Speech, Speech to text
   - Voice identification/verification, Noise Reduction
3. NLP: Text Classification, Sentiment Analysis
4. Explainable AI

- **Visual Question Answering** is the task to predict/generate an answer of textual questions from the content of image.

Fig 1: The examples of Visual Question Answering dataset (VQA v2.0)

- **Data Collection**: by blind people
  - Image is taken by personal phone
  - Question is recorded in the conversation
- **Annotation**: Crowdsource

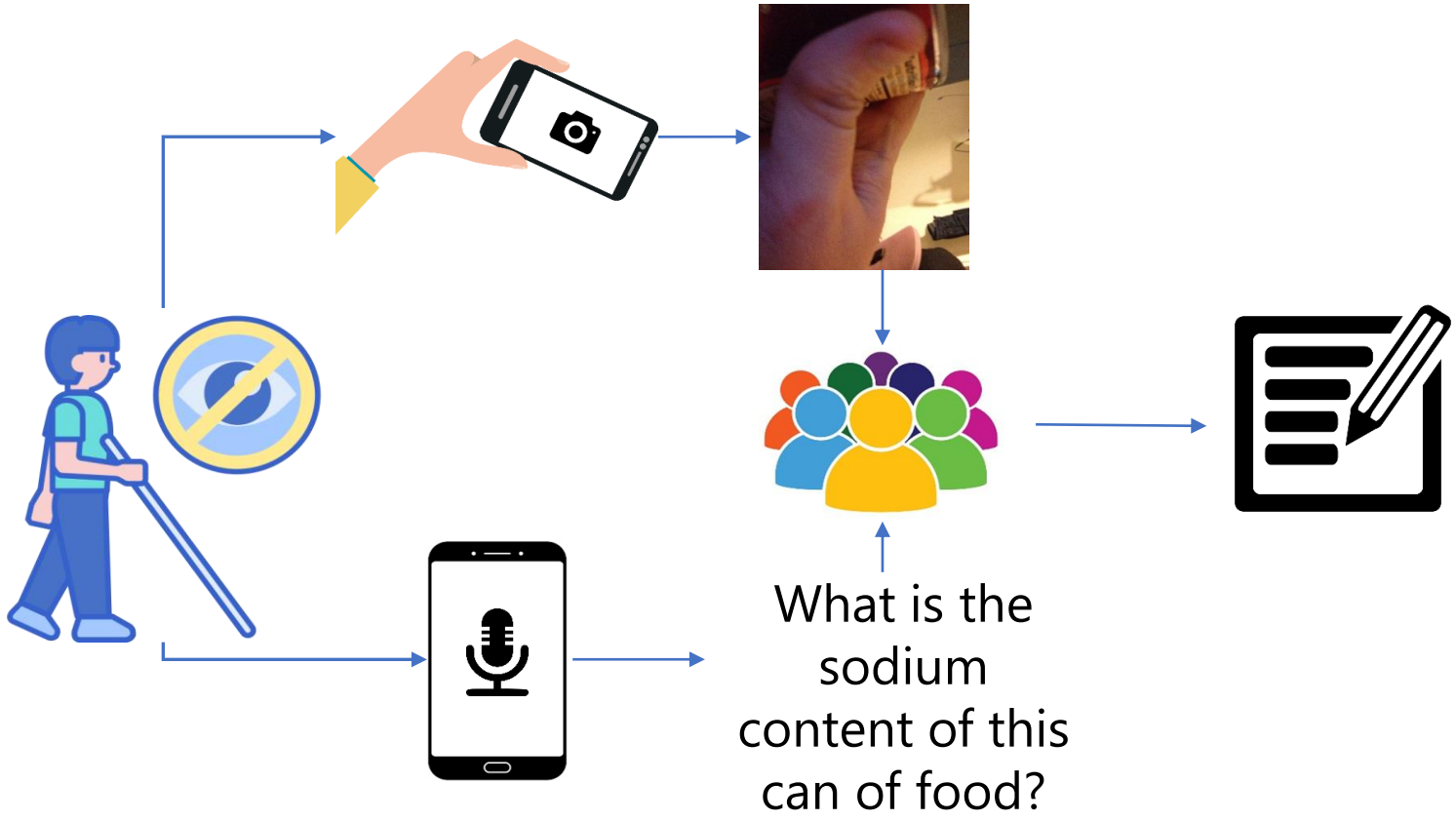

Fig 2: The detailed procedure to collect the VQA dataset for blind people (VizWiz-VQA 2.0)

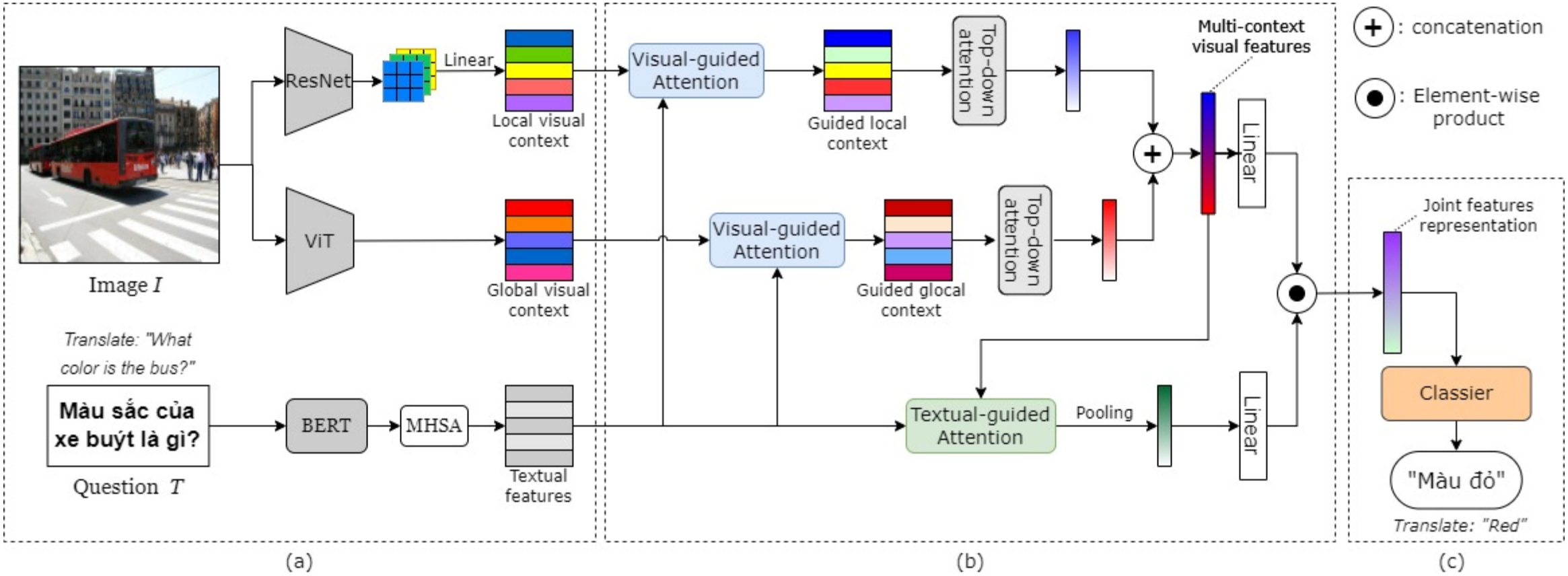

Fig 3: The detailed architecture of Multi-vision Embedding in Contextual Attention Model (Duc Anh Nguyen et al. 2022)

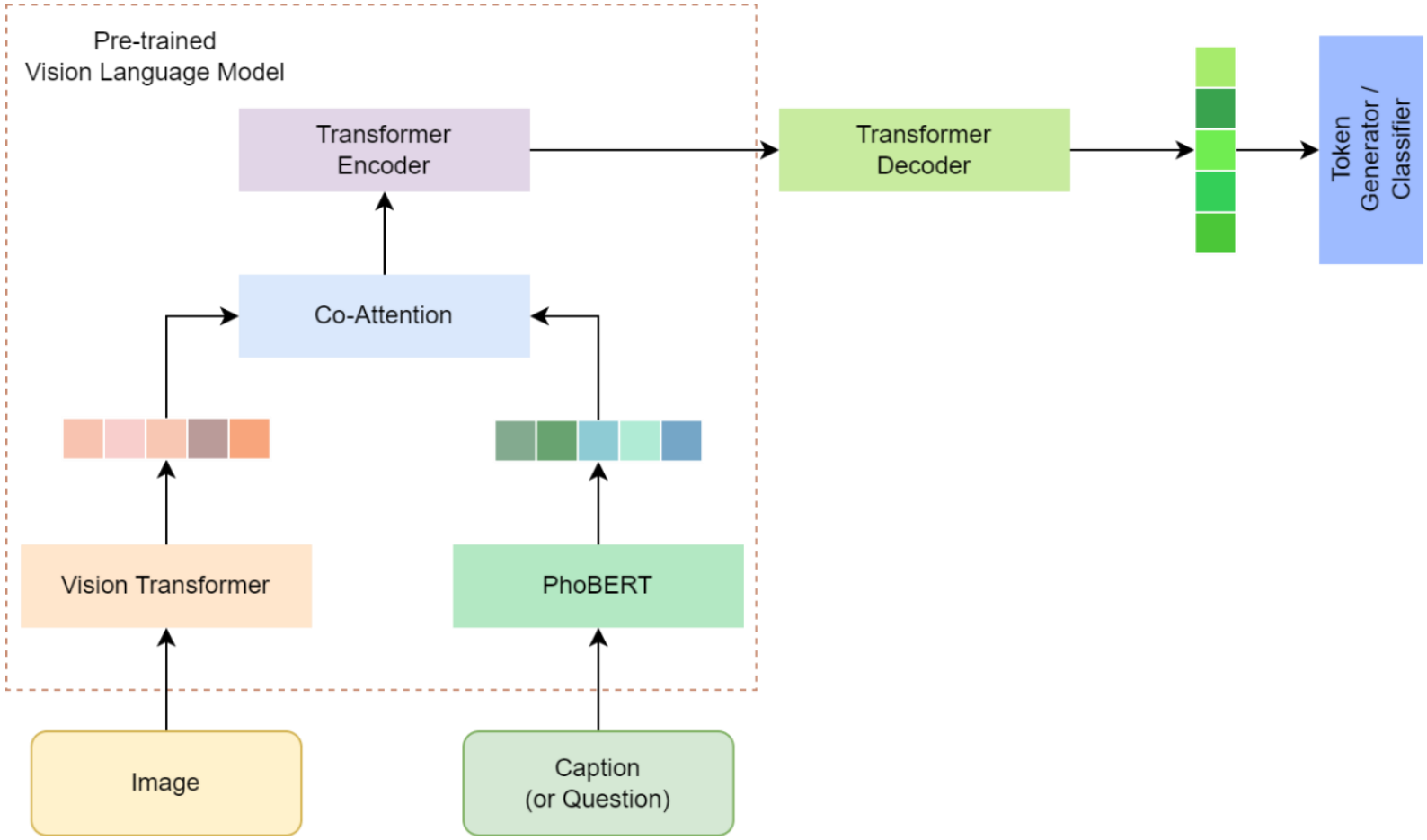

Fig 4: The detailed architecture of pre-trained vision-language model (Huy Nguyen et al.)

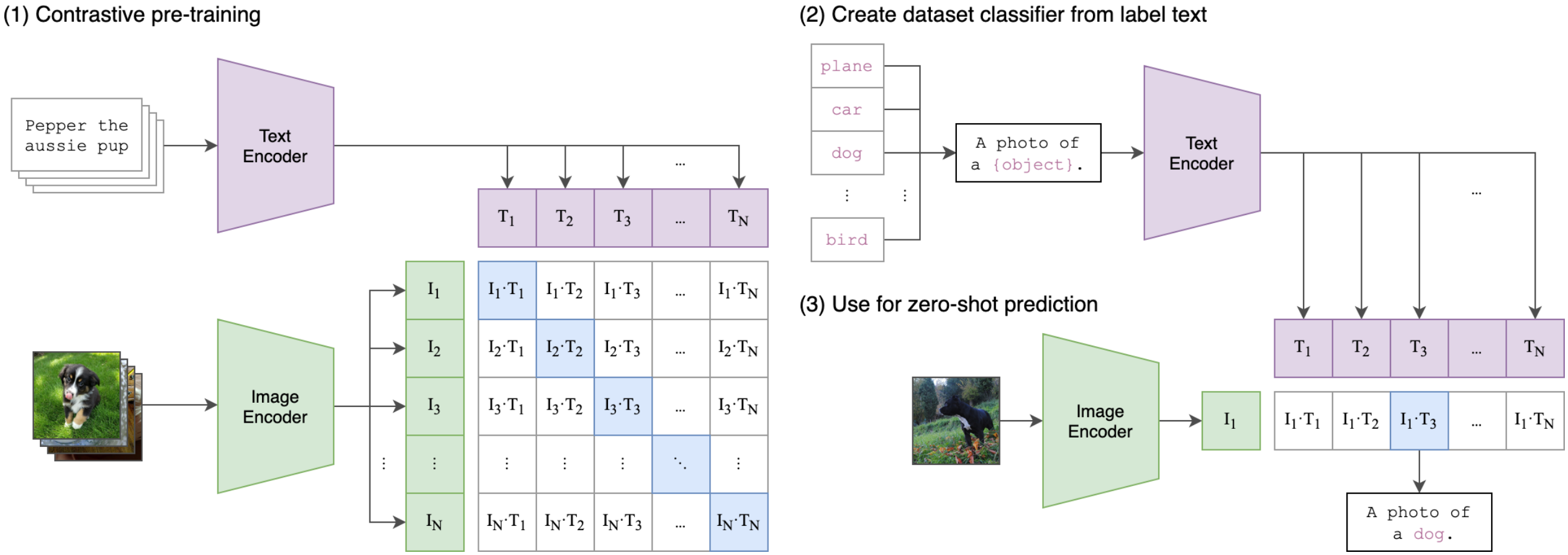

Fig5: Learning Transferable Visual Models From Natural Language Supervision (Alec Radford et al. 2021)

Youtube Engagement – Explainable AI

fit@hcmus

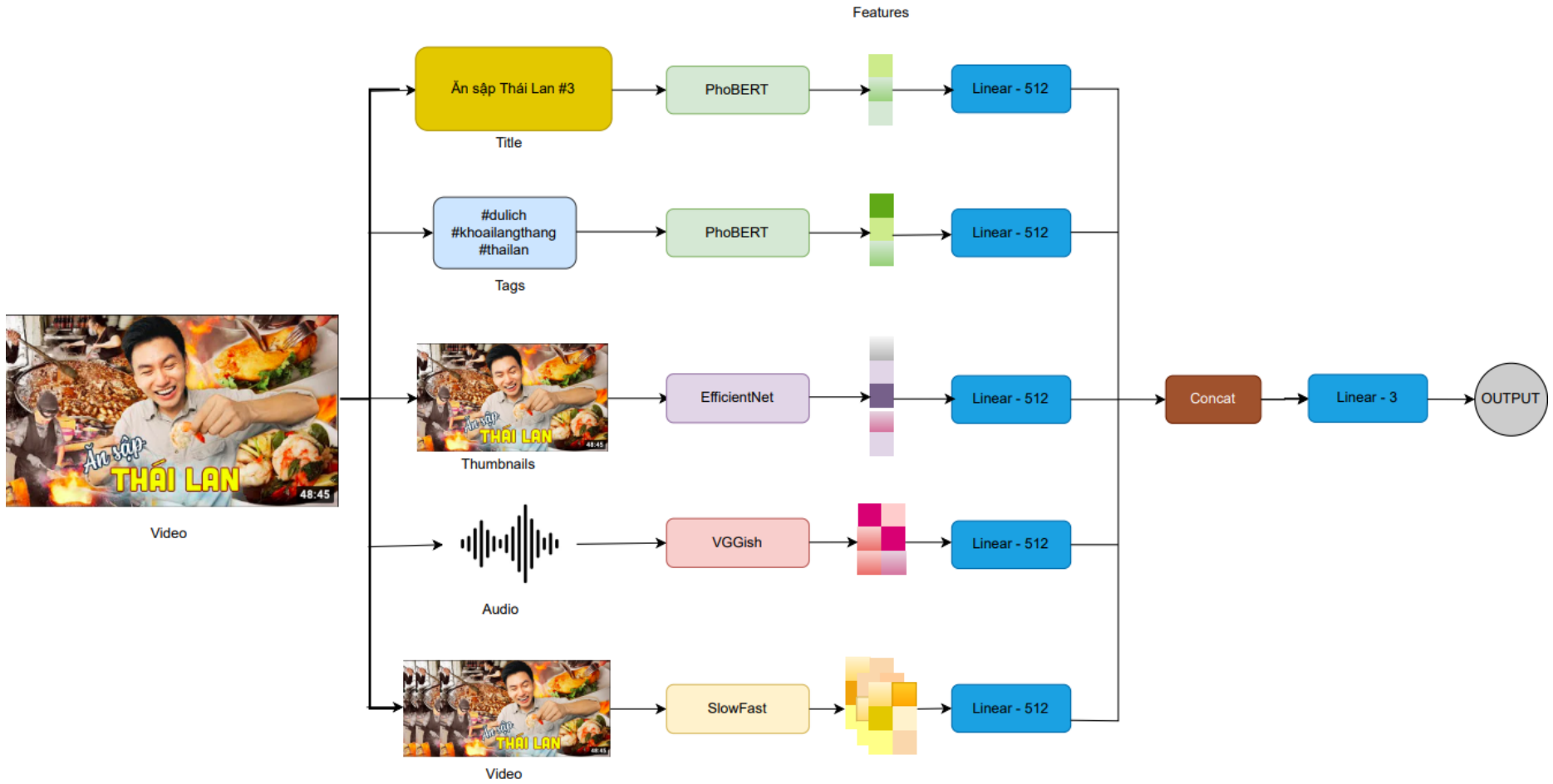

Fig6: Youtube Engagement Analytics via Deep Multimodal Fusion Model (Minh-Vuong Nguyen-Thi et al. 2022)

Contact

- Email:
  - Lê Thanh Tùng: lttung@fit.hcmus.edu.vn
  - Nguyễn Tiến Huy: ntienhuy@fit.hcmus.edu.vn
  - Subject: [CH2023][K3x]
  - Content: CV + Transcript +