Face Processing in Video

Tác giả thứ nhất[1,3] **Tác giả thứ 2**[2] **Tác giả thứ 3** [3,1] *and* **Tác giả thứ 4** [3,1]

[1] Trường ĐH……

[2] University of Science HCMC, Vietnam

[3] National Institute of Informatics

What ?

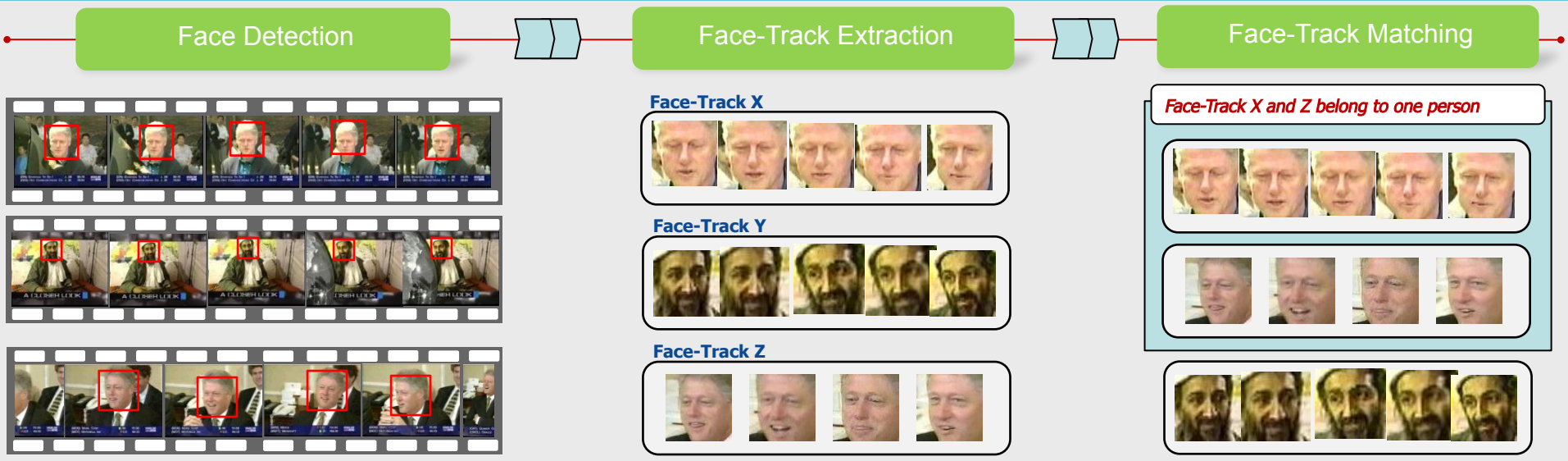
We introduce a framework to process and retrieve faces in video, in which we have:

- Proposed a robust method to extract face-tracks in news video.
- Built the largest face database compare to current popular worldwide face databases.
- Evaluated several face-track matching methods.

Why ?

- The human face is one of the most important objects in video since it provides rich information for spotting people of interest and is the basis for interpreting facts. Therefore, detecting and recognizing faces appearing in video are essential tasks of video indexing and retrieval applications.
- Most studies have focused on static images rather than ***large-scale*** and ***real video dataset*** .

Overview



Description

1. Face Detection

- The face detector implemented in OpenCV based on Viola method was used for detecting frontal views of faces in every frame of our video sequences.
- A high threshold was used to reduce the number of false positives.

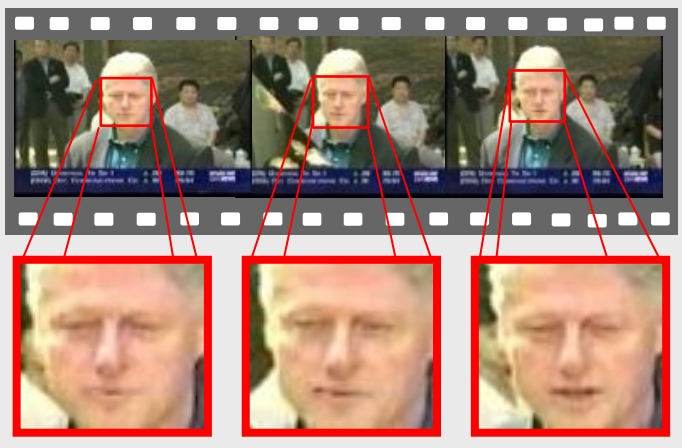


Figure 1. Detection results with a high threshold.

2. Face-Track Extraction

- We used Kanade–Lucas–Tomasi (KLT) method to create and track key/interest points between frames.
- The number of key points that pass through pairs of faces in consecutive frames was computed to make decision on grouping faces into face-tracks.
- Several treatments are proposed to handle tracking traps in news video.
  - Flash-frame detector.
  - Adaptively generating key points.

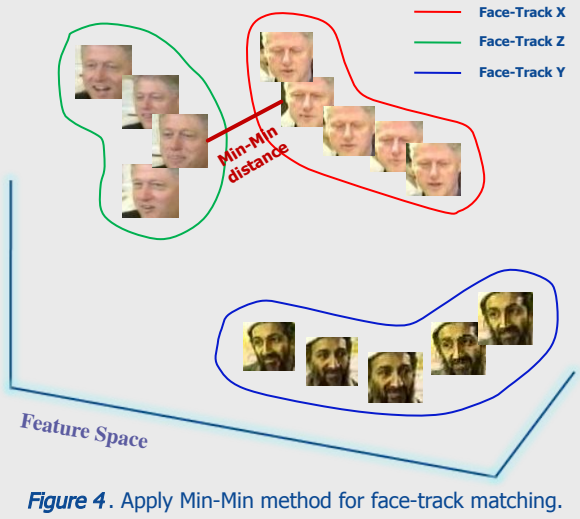


Figure 4. Apply Min-Min method for face-track matching.

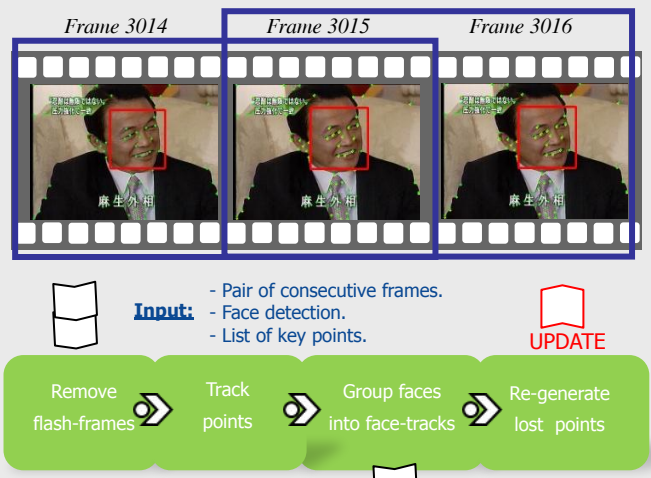


Figure 2. Process-flow of face tracker.

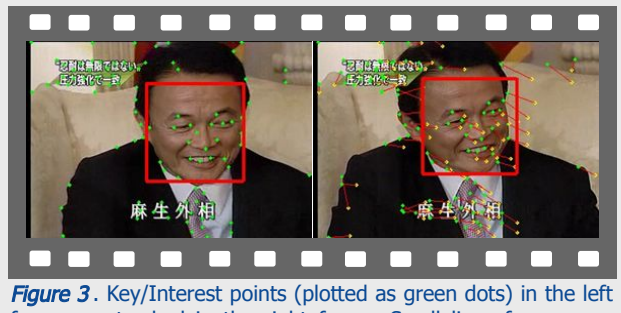


Figure 3. Key/Interest points (plotted as green dots) in the left frame are tracked in the right frame. Small lines from green dots are motion of these points. Two faces in these frames share 22/23 points.

3. Face-Track Matching

- Face-track matching is done by applying similarity estimation methods (e.g. Min-Min distance) on the feature space.

NII

Tên tác giả chính– Trường Đại học

TEL : ******* Email : *********

RESEARCH AND DEVELOP A DEEP LEARNING MODEL FOR DETECTING AND RECOGNIZING HUMAN ACTION

Nguyễn Đặng Đức Mạnh[1]

[1] University of Information Technology

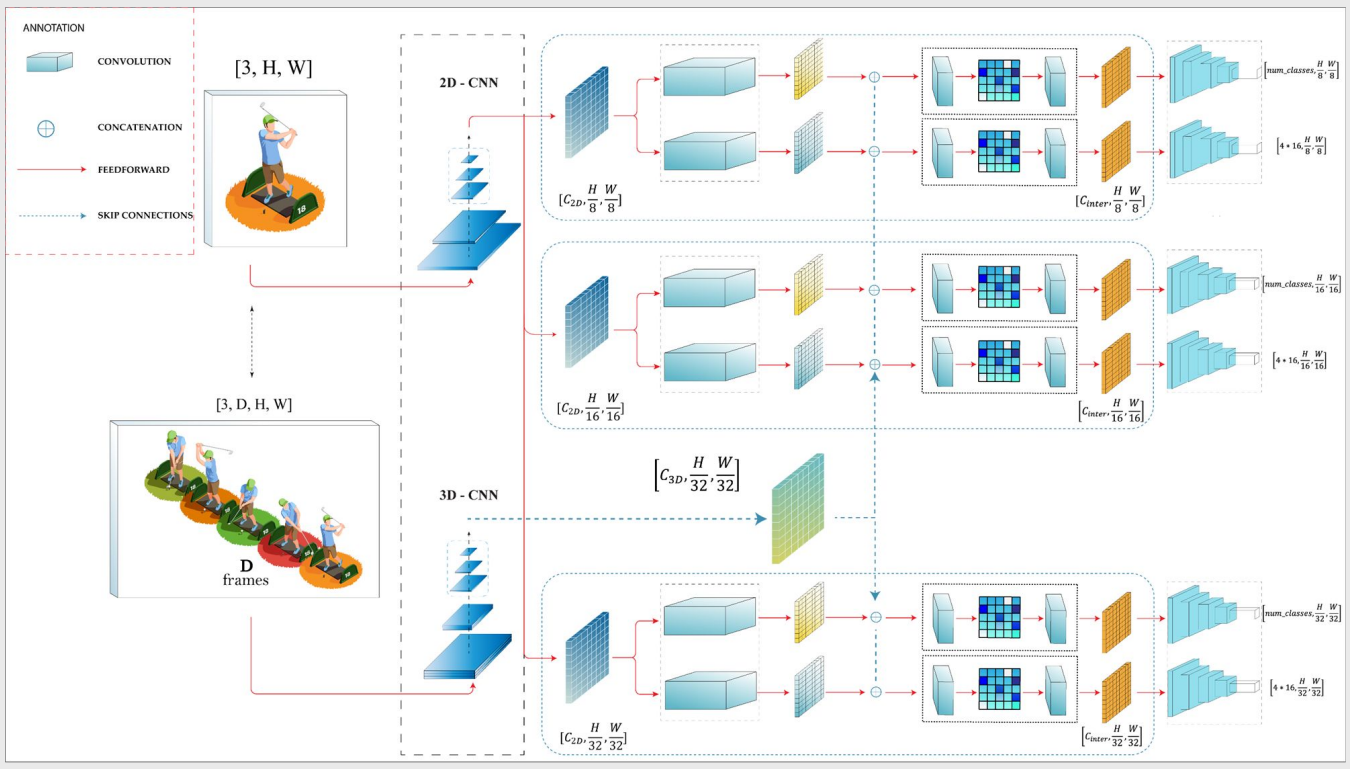
What is HADR?

In the field of computer vision, HADR (Human Action Detection and Recognition) is a task that requires models to detect (bounding box) and recognize (classify) human actions from a given short clip.

Why do we need bounding box?

- **Label**: Useful in video classification tasks, focusing on the general content of the video.
- **Starting + ending point:** Useful in tasks that require querying where a specific action occurs in the video.
- **Bounding box**: Meaningful in more specific query tasks (e.g., person A performing action B), effectively utilizing spatial information in each frame.

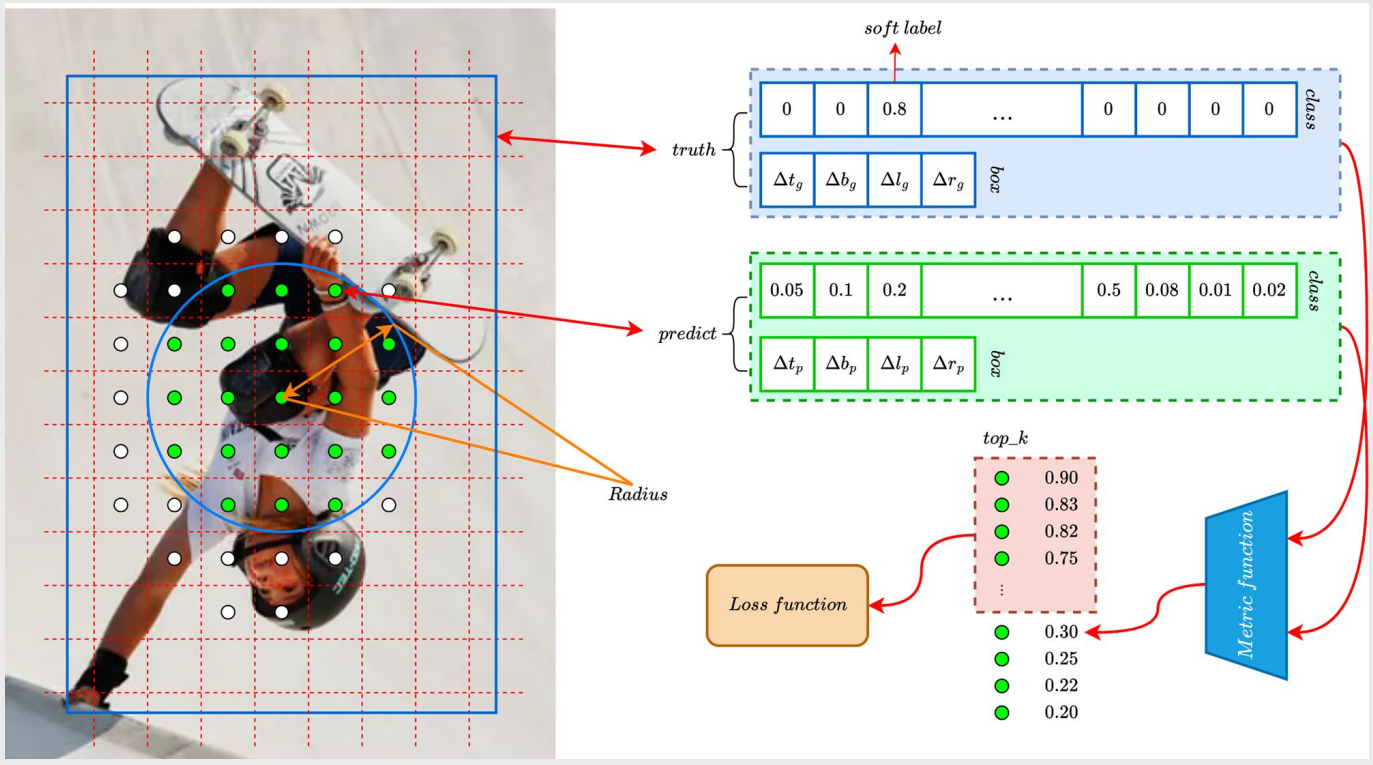
Architecture Overview



- **Backbone 2D**: Used to extract spatial features, we utilize YOLOv8
- **Backbone 3D:** Used to extract motion features over time, we utilize several available 3D backbones including I3D, ResNeXT.
- **Neck:** Enhancing the semantic features of the feature maps by leveraging the Feature Pyramid Network architecture.
- **Head:** Used to predict labels and bounding boxes, the output branch of the model.

Label Assignment

Illustrates in detail the pipeline of the label assignment stage. The candidates deemed eligible (green circles) must lie within the truth bounding box and be within a distance no greater than the radius R from the object's center. Subsequently, the candidates meeting the criteria are evaluated through a metric function, and only the top_k candidates with the highest metric are considered as positive.



NII

Tên tác giả chính– Trường Đại học

TEL : ********* Email : *********

APPLYING SEGMENT ANYTHING MODEL FOR THREE-STAGE SEGMENT-BASED GRAY-SCALE IMAGE COLORIZATION

Lý Văn Nhật Tiến[1]

[1] Trường ĐH Công nghệ thông tin - ĐHQG TPHCM

What ?

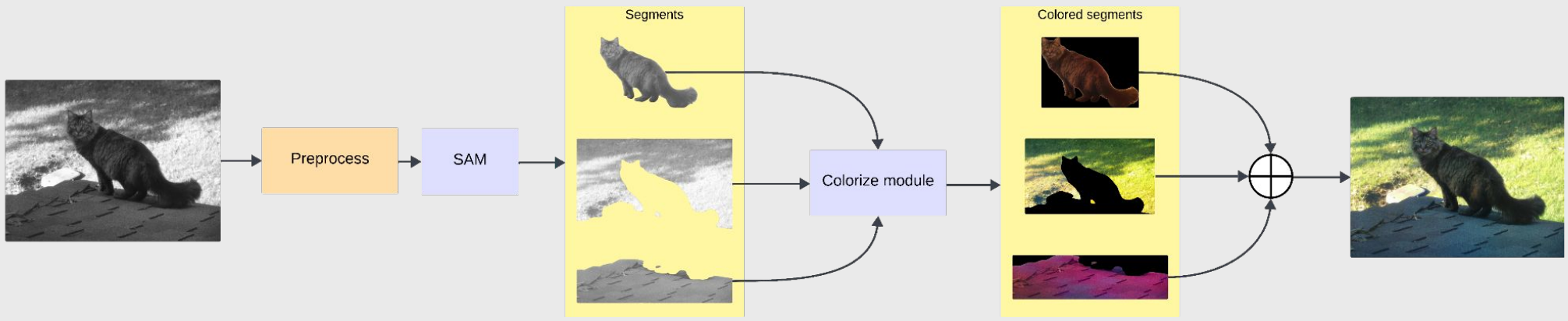
We introduce new approach to colorize grayscale images:

- Proposed a pipeline to extract segments in grayscale images, color these segments and combine into output image with realistic color.
- Evaluated the pipeline performance on several benchmark and compare with existing methods.

Why ?

- Segments are the cornerstone of an image and provide useful information if harnessed correctly. Existing methods mainly focused on pixels and the correlation between them and the whole image or local context but within a ***fixed*** area rather than a dynamic one.
- With the recent advance in segmentation methods (typically SAM), extracting segments can now be performed easily and therefore made this approach more viable.

Overview



Description

1. Extract segments

- We propose the ***Segment Anything Model*** for this stage for its robustness and computational efficiency.
- Define a maximum number of segments in order not to over-segmentation, leaving the pipeline with details too fine-grained to contain any useful information.

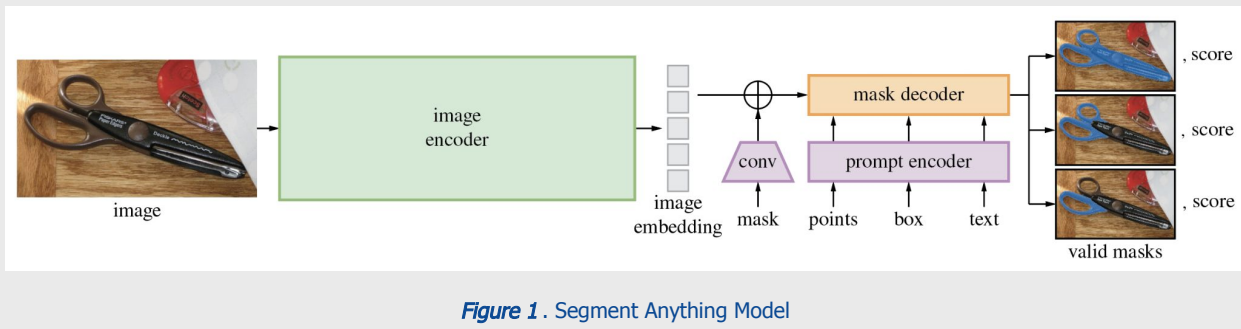


Figure 1. Segment Anything Model

2. Colorize module

- We compare vision models with performance, speed and lightweight-ness as criteria.
- Select best models to experiment as implementation for this module.
- As of the aforementioned criteria, ***Stable Diffusion*** is a viable candidate.

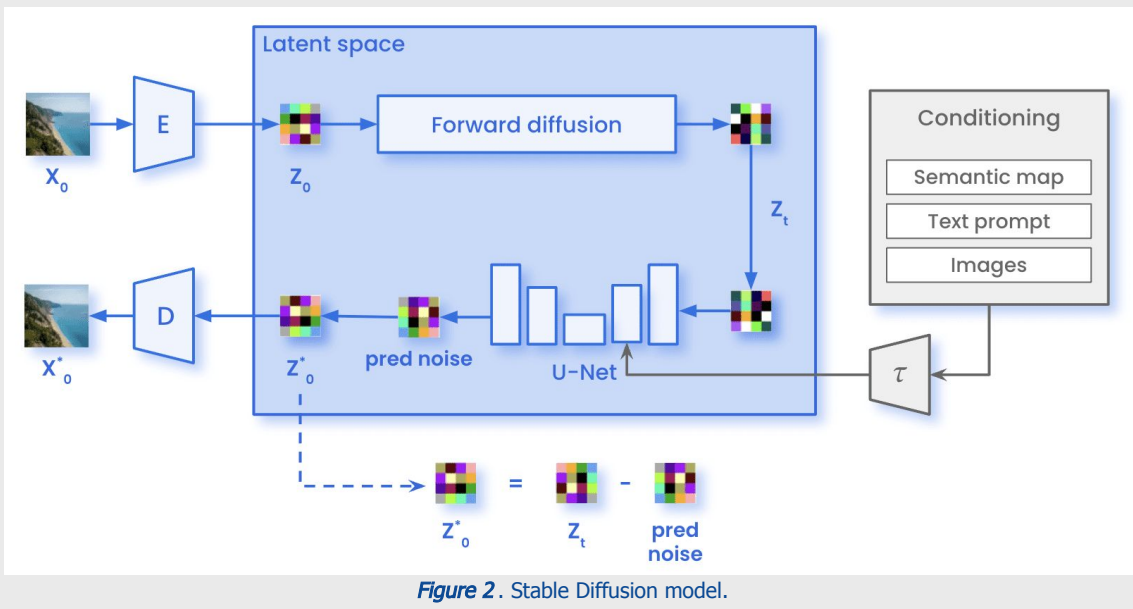


Figure 2. Stable Diffusion model.

3. Combine module

- This module is responsible for combining colored segments into the final result.
- We propose using a simple method (image matrix adding as one). However, this is an important component and will be experimented with more advanced methods

NII

Lý Văn Nhật Tiến– Trường Đại học Công nghệ thông tin - ĐHQG TPHCM
TEL : 0942319042 Email : 21521525@gm.uit.edu.vn

Contextual anomaly detection in multivariate time series using graph neural network and multi-context attention mechanisms

Nguyễn Anh Hải Ngọc — **University of Information Technology**

What ?

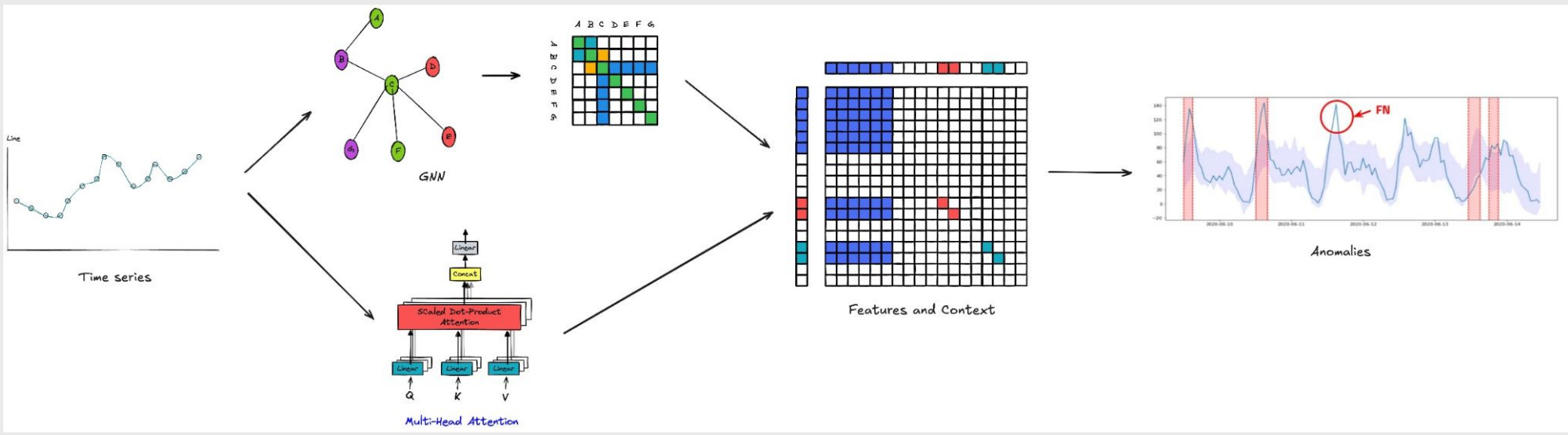
We introduce a new method to detect anomalies in multivariate time series

- Combine GNN with multi-context attention mechanism
- Improve the performance of anomaly detection using attention mechanism to contextualize anomalies in different dimensions and time frames

Why ?

- The anomaly detection is the most important problem in analyzing time series data
- Most approaches have focused on contextual anomaly detection on fixed dimensions and/or time frames

Overview



Description

1. Graph Neural Network (GNN)

The Graph Neural Network is used to decompose time series data into nodes

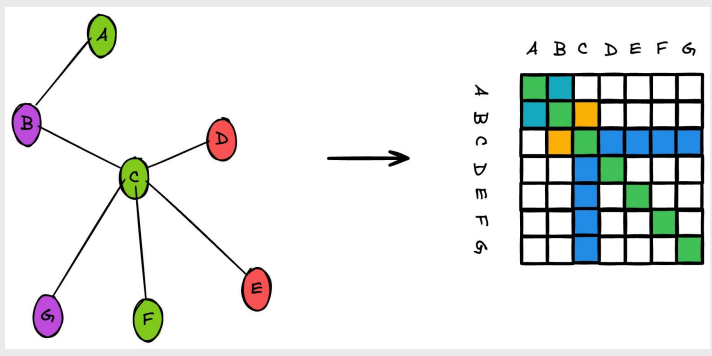


Figure 3. The GNN decompose data into nodes

2. Multi-context attention mechanism

The multi-context attention mechanism is used to learn the contextual data from the nodes in the network.

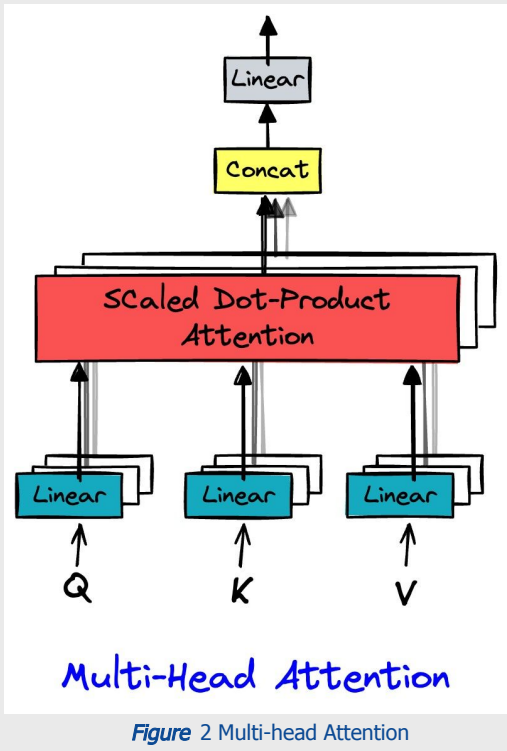
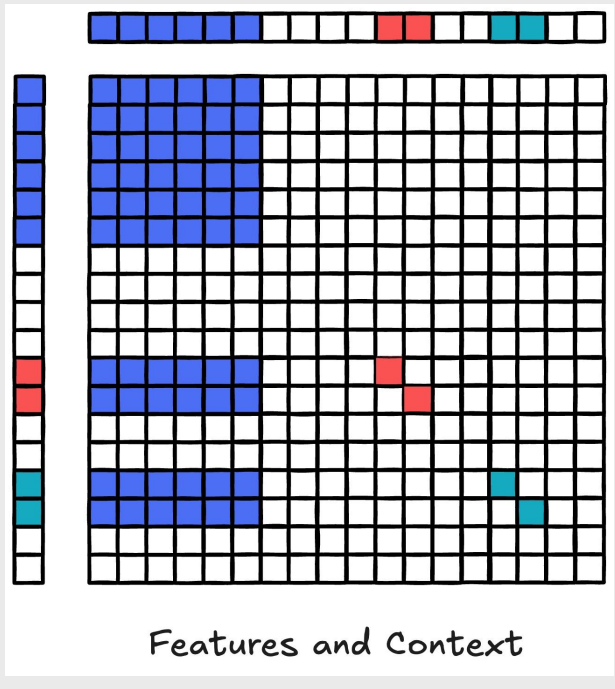


Figure 2 Multi-head Attention

3. Combine the context and the features

The output of attention mechanism then be combined into the final features result. This can be inputted into dense layer to reduce dimension



NII
Nguyễn Anh Hải Ngọc – Trường Đại học Công nghệ Thông tin
TEL : 0376374908 Email : 22520955@gm.uit.edu.vn

ENHANCING KNOWLEDGE DISTILLATION PERFORMANCE THROUGH ATTENTION TRANSFER FOR CLASSIFICATION TASKS.

Nguyễn Viết Đức[1,1] **Đoàn Văn Hoàng**[1,2]

[2] University of Information and Technology
UIT, Vietnam

What ?

We introduce now approach to enhance performance for classification task in IoT device

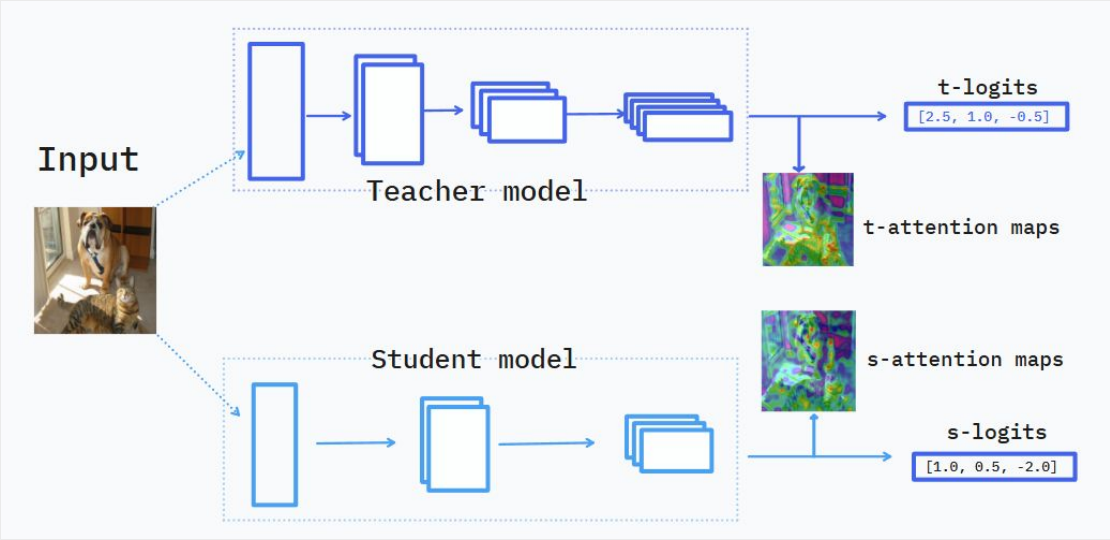
- Proposed a robust architecture combine Knowledge Distillation with Attention Transfer
- Evaluated the pipeline performance on several benchmark and compare with existing methods.

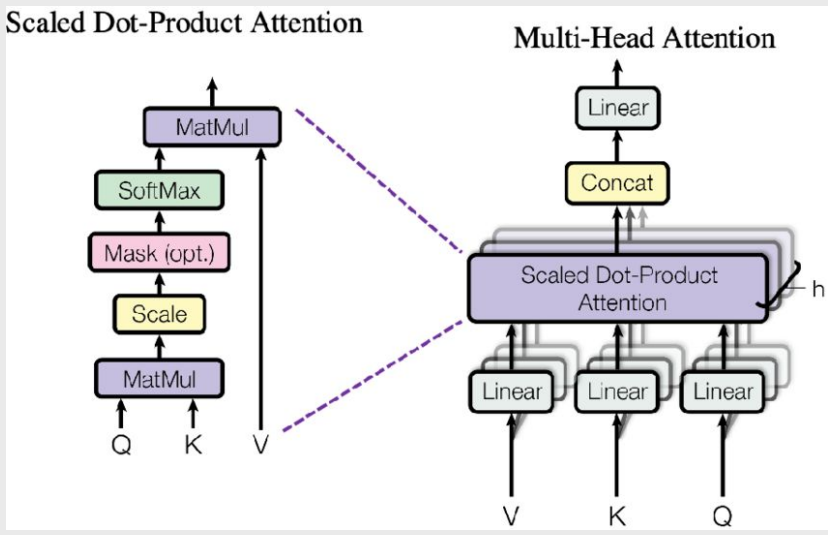
Why AT in Classification Tasks?

- Traditional KD limitations: In some cases, logits alone don't capture all the nuances of the data (e.g., spatial relationships in an image).
- AT advantage: By transferring attention maps, the student learns not just the output but also where and what the teacher is focusing on, leading to better generalization and accuracy.

Overview

a) Overall Architecture





b)Seft-attention mechanism

Description

1.Teacher Model: Using **ResNet-50**, a large, pre-trained model as a source of knowledge transfer with attention maps extracted from intermediate layers using self-attention mechanism to highlight characteristic spatial regions and logits feature

2.Student Model: **MobileNetV2**, Smaller, computationally efficient model is trained to:

- Match the logits of the teacher using **KL Divergence Loss**.

$$L_{KD} = KL(Z_t || Z_s)$$

- Mimic the teacher's attention maps using **Mean Squared Error (MSE) Loss.**

$$L_{AT} = MSE(A_t, A_s)$$

Aggregate loss function:

$$L_{total} = \alpha L_{KD} + \beta L_{AT}$$

3.Outcome: The student achieves high performance with reduced size, making it suitable for deployment on devices with limited computational power.

NII

Tên tác giả chính– Trường Đại học
TEL : ******* Email : *********

DETECTING AND CLASSIFY CARS FROM IMAGES BY COMBINATION OF YOLO AND RESNET MODELS

Vũ Huy Hoàng[1,3] **Đoàn Tiến Đạt**[1,2]

{[1] 21522104, [2] 21521933}

[3] Trường ĐH Công nghệ Thông tin, ĐHQG TPHCM

What ?

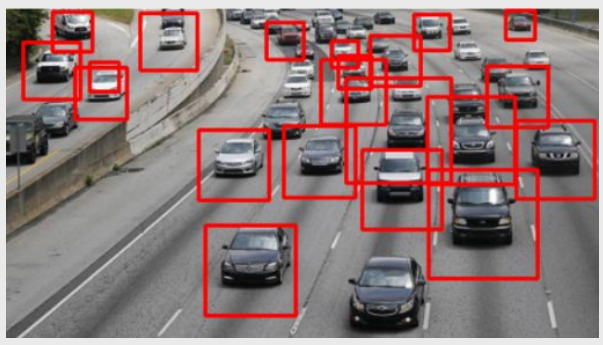
- Building a car recognition method from images with data with inconsistent labels and annotated data from many different sources.
- Introducing a framework for handling inconsistent data using deep learning networks and hybrid models.
- Evaluate the effectiveness of the method with different data sets, comparing with current methods.

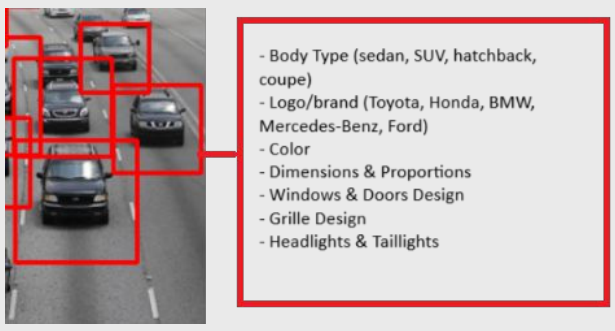
Why ?

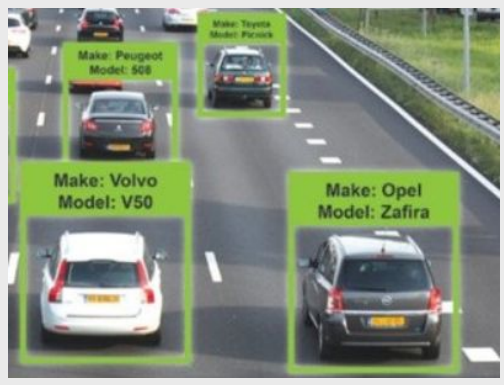
- Enhance the ability to automatically identify cars in traffic monitoring systems, parking lot management, or related applications.
- Currently, car image data often has inconsistent labels, causing difficulties in model training.
- The proposed method has the potential to overcome the disadvantages of traditional recognition systems and improve performance.

Overview

Detect cars from image → Cars data extraction → Return data of cars







Description

1. Detect vehicles from images

The published solution approach uses deep learning models to automatically detect cars from images, even in data with minimal or missing labels.
By combining many advanced techniques, the system is capable of recognizing a variety of vehicles in different environmental conditions, including streets, parking lots and heavy traffic.

Use the YOLO (You Only Look Once) model:

- The YOLO model enables real-time object detection with high speed and good accuracy. This is an ideal choice for vehicle identification in traffic monitoring systems thanks to its ability to process frames continuously and detect multiple vehicles at the same time on crowded traffic routes.

2. Extract data from vehicle

After detecting the vehicle, the system extracts relevant detailed information, including model (sedan, SUV, coupe), brand such as Toyota, Honda, BMW..., size, proportion, window design, grille and headlight/taillight system.

Using the ResNet model:

- The ResNet (Residual Network) model is used to classify images, ensuring accurate recognition of vehicle brands and models. With the ability to effectively process complex data sets, ResNet helps the system achieve high accuracy in classifying technical details.
- The data is processed through a hybrid model, allowing predictions to accurately recognize the vehicle's location and specific characteristics despite incomplete or inconsistent conditions.

3. Returns vehicle data

The end result of the system is to provide detailed information about the vehicle, including important attributes such as brand, model and specifications.

Hybrid Model:

- The system uses YOLO to detect the vehicle's location and then uses ResNet to classify the brand and model, ensuring both speed and accuracy.
- This system not only improves the accuracy of vehicle recognition but also improves the performance of practical applications such as traffic monitoring and parking management.

NII

Đoàn Tiến Đạt – Trường ĐH Công nghệ Thông tin, ĐHQG TPHCM
TEL : 0933520344 Email : 21521933@gm.uit.edu.vn

PROTECTING IMAGES FROM MACHINE LEARNING MODELS USING NATURALLY CAMOUFLAGED ADVERSARIAL PATCHES.

Huynh Chi Nhan

University of Information Technology - National University of Vietnam

Nguyen Ho Nam

Department of Computer Science

What ?

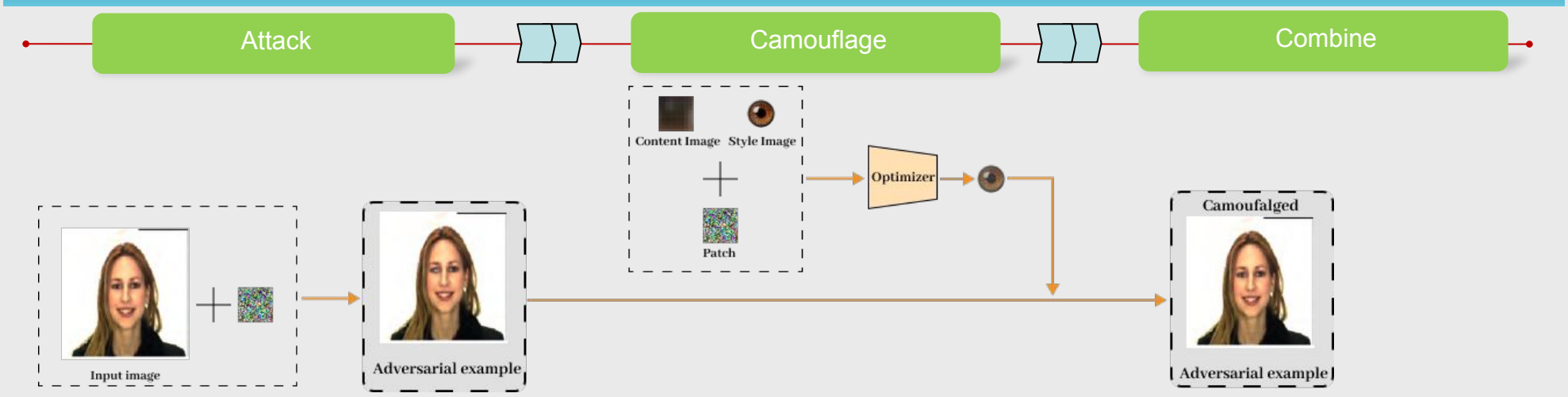
The study on generating Camouflaged Adversarial Patches aims to achieve the following objectives:

- Experiment with the implementation of Adversarial Patches to evaluate their performance on the YOLOv8 detection model.
- Enhance the content and style of the patches using Style Transfer techniques.
- Provide foundational knowledge and mathematical principles related to Adversarial Patches and Style Transfer.

Why ?

- Personal images on the internet can easily be exploited to create Deepfake videos. So, how can users protect their personal images?
- To address this issue, we propose using Camouflaged Adversarial Patches, a method that disrupts the facial extraction process of users. This solution not only safeguards personal images but also preserves the aesthetic and natural appearance of the original photos.

Overview



Description

1. Dataset and detection model

For classification, we experiment on the YOLOv8 model on the CelebA dataset.

Figure 1. CelebA dataset.

2. Adversarial Patches

Since we do not have access to the parameters of the YOLO model, we construct the **Adversarial Patch** based on the model's input and output. The eyes are a critical region in face detection; therefore, we create a **Mask** based on the size and position of the eyes. To cause the model to misclassify the ground truth of the bounding box and class, we optimize a patch (**Patch**) to maximize the loss of the object detector with respect to the ground truth label and bounding box, when the patch is applied through a predefined function.

$$\hat{P}_u = \arg\max_P \mathbb{E}_{x,s}\left[L(A(x,s,P);\hat{y},\hat{B})\right]$$

During back-propagation, we update the pixels of **Patch**.

3. Style Transfer

We provide both a style image and a content image. The style image is sourced externally, while the content image represents the region of the original image covered by the patch.

In a neural network, style optimization is performed using features extracted from the lower-dimensional layers, whereas content optimization leverages high-dimensional features from deeper layers.

The optimization process uses cross-entropy as the primary loss function and focuses on two key components:

1. **Content Loss**: Ensures that the structural information of the original image (or the patch area) remains recognizable.

$$L_{\text{content}}(C,\delta) = \frac{1}{2}\sum_{i,j}\left(C_{ij}^l - \delta_{ij}^l\right)^2$$

2. **Style Loss**: Transfers the texture and visual features of the selected style image (e.g., rust texture) onto the patch. Style features are extracted from the style image using the Gram matrix of the convolutional layers in a pre-trained neural network.

$$L_{\text{style}}(S,\delta) = \sum_l \frac{1}{N_l M_l}\sum_{i,j}\left(G_{ij}^l(\delta) - G_{ij}^l(S)\right)^2$$

4. Camouflaged Adversarial Patch

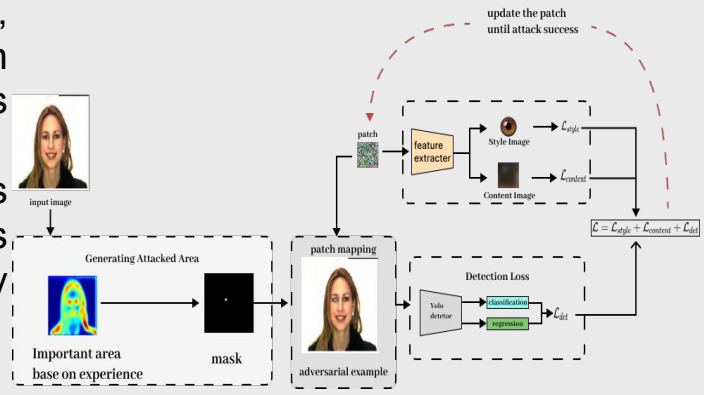


Figure 2. Diagram of Camouflaged Adversarial Patch.

NII

Huynh Chi Nhan – University of Infomation Technology

TEL : 0376856852 Email : 22520996@gm.uit.edu.vn

IMPROVING AND OPTIMIZING MULTIMODAL MODELS FOR AUTOMATED X-RAY REPORT GENERATION

Minh Quan Tran[1]

[1] University of Information Technology. Vietnam National University, Ho Chi Minh City.

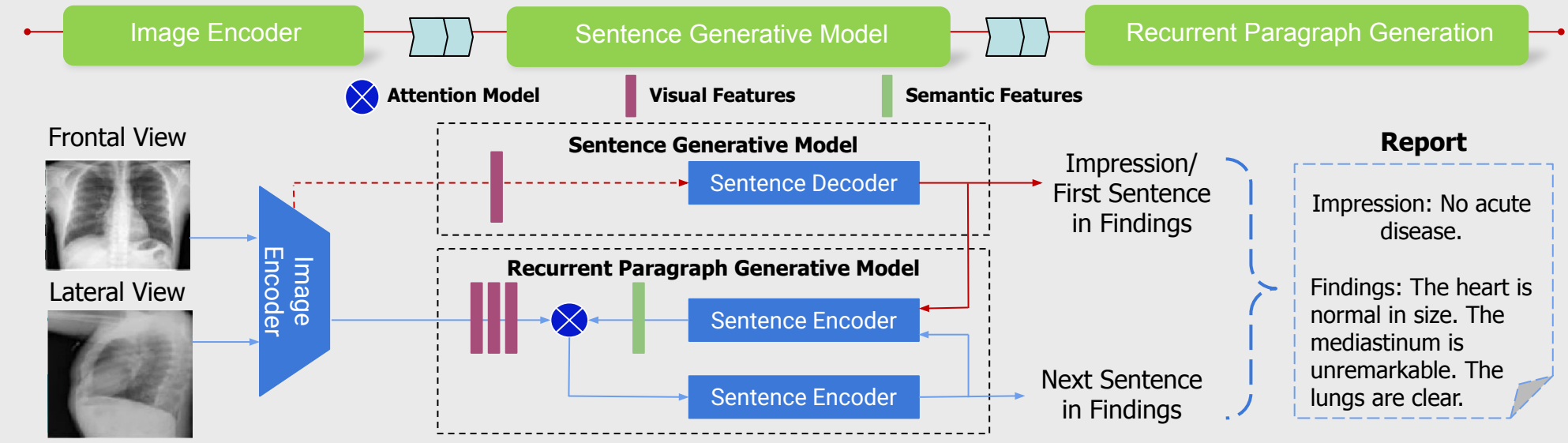
Motivations

Previous research mainly used deep learning models such as CNN and RNN to generate reports from X-ray images, but still had difficulty in combining multi-modal information and low accuracy. The goal of this project is to **improve these models** by applying *Vision Transformers* and *BioBERT*, **improving semantic understanding and image processing**. The project will s**olve the problem of overfitting** and **increase accuracy**, thereby supporting doctors in diagnosis and decision making.

Targets

- **Improve accuracy:** Optimize the ability to generate reports from X-ray images using multimodal models, improving diagnostic accuracy.
- **Minimize overfitting:** Use large and diverse data sets to improve the generalization ability of the model, ensuring reliability.
- **Support clinicians:** Develop automated systems to help doctors analyze images and make decisions more quickly and accurately.

Overview



Description

1. Image Encoder

- An image encoder is first applied to extract both global and regional visual features from the input images.
- The image encoder is a Convolutional Neural Network (CNN) that automatically extracts hierarchical visual features from images

2. Sentence Generative Model

- In general, both the one-sentence impression and the first sentence in the findings paragraph contain some high level descriptions of the image.
- A sentence generative model that takes the global visual features learned by the image encoder as input.
- Such a model can be trained to generate the impression. It can also be jointly trained with the recurrent generative model to generate the first sentence in the findings as an initialization of the recurrent model

3. Recurrent Paragraph Generation

- Recurrent paragraph generative model takes the sentence and regional image features as input and generates findings paragraph sentence by sentence.
- It has two main components: sentence encoder and attentional sentence decoder. Sentence Encoder is used to extract semantic vectors from text descriptions. Attentional Sentence Decoder takes regional visual features and the previously generated sentence as a multimodal input, and generates the next sentence.

4. Research Plan

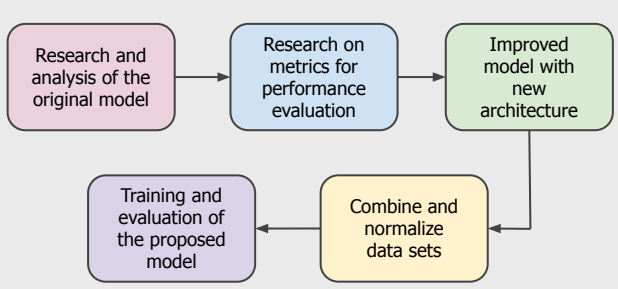


Figure 2. Research plan diagram

Figure 1. Example of original report compared to report generated by recurrent attention model

Minh Quan Tran – University of Information Technology. Vietnam National University. Ho Chi Minh City.
TEL : 0337459509 Email : 22521191@gm.uit.edu.vn

RESEARCH AND DEVELOPMENT OF PLANT DISEASE DETECTION SYSTEM

Nguyễn Minh Lộc **Huỳnh Chấn Kiệt**

Trường Đại Học Công nghệ thông tin

What ?

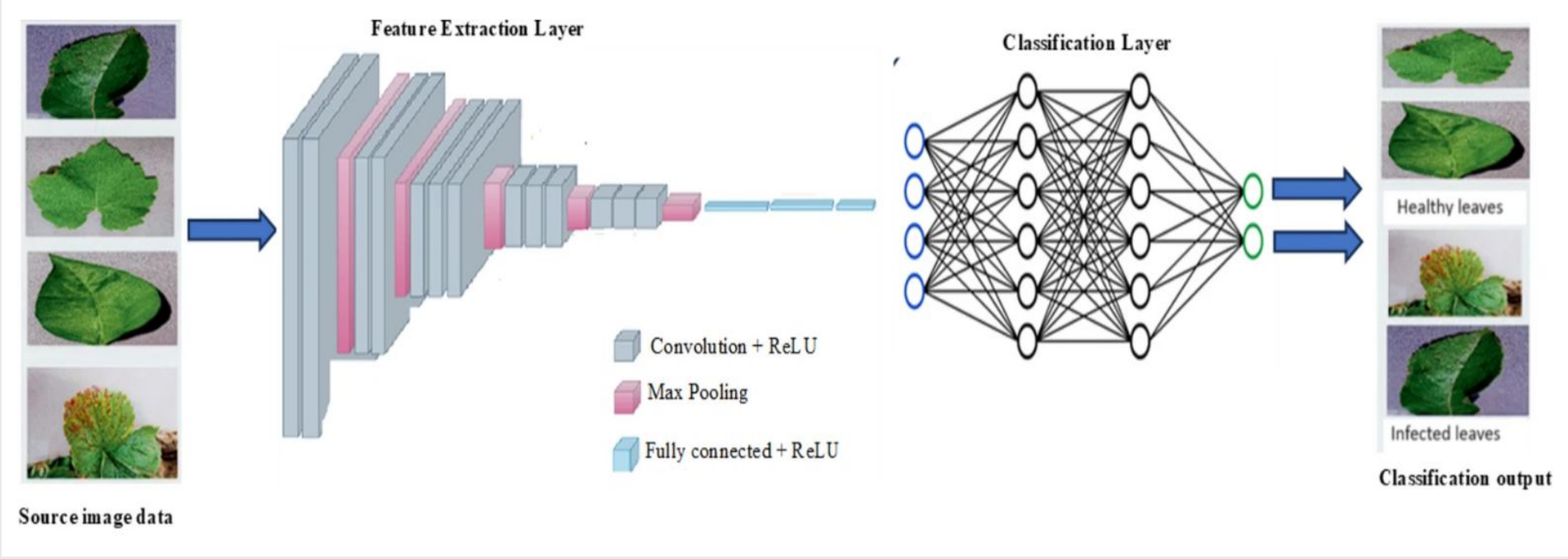
We introduce a system to detect and classify diseases on plant leaves, in which we have:

- Identify the diseased area of the leaf
- Built the evaluation model based on the pre-prepared data.
- Evaluated results by several methods.

Why ?

The condition of plant leaves is one of the most critical indicators in agriculture, as it provides valuable information for identifying diseases and assessing plant health. Therefore, detecting and classifying diseases on plant leaves are essential tasks in precision agriculture and crop management applications.

Overview



Description

1. Preprocessing of Plant Leaf Images

- Data Collection: Images are gathered from farms and nurseries, with the PlantVillage dataset used as a base.
- Preprocessing Steps: Standardizing image sizes, adjusting brightness, and removing noise or blurred images to enhance data quality.

2. Building a Multi-Branch Deep Learning Architecture

- EfficientNet Backbone: Used for feature extraction due to its balanced performance in accuracy and processing speed.
- Attention Mechanism: Designed to focus on diseased regions of the leaf, combined with Grad-CAM to generate heatmaps for interpretability.

3. Model Optimization and Evaluation

- Loss Function: A combination of cross-entropy and focal loss to address imbalanced data:

$$L = \alpha CE(y, \hat{y}) + \beta FL(y, \hat{y})$$

- Evaluation Metrics: Accuracy, precision, recall, F1-score, and inference time
- Comparative Analysis: Model performance compared with existing methods to validate effectiveness.

NII

Tên tác giả chính– Trường Đại học
TEL : ******* Email : *********