# 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

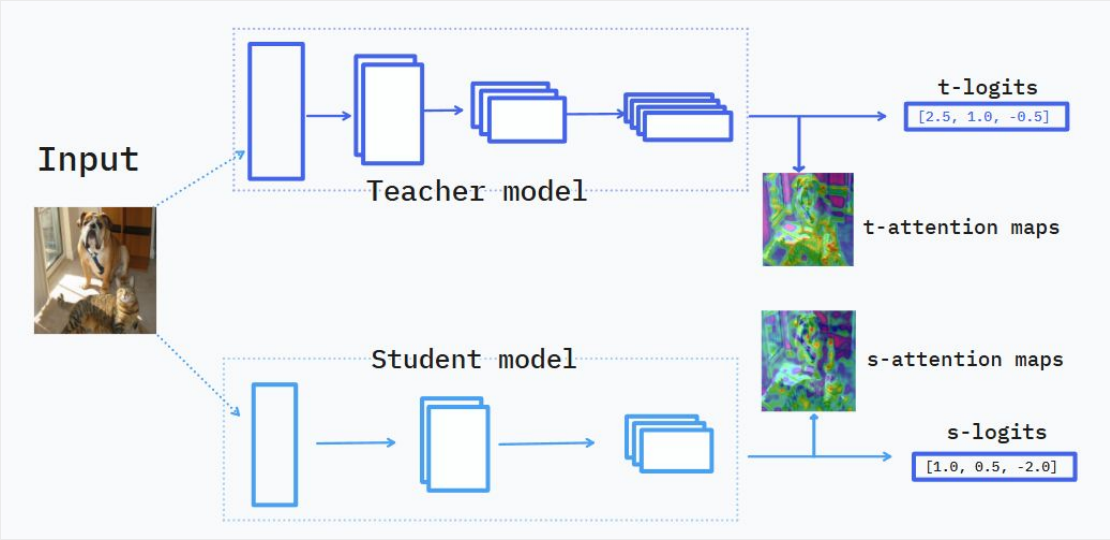
- Proposed a robust architecture combine Knowledge Distillation with Attention Transfer
- Evaluated the pipeline performance on several benchmark and compare with existing methods.

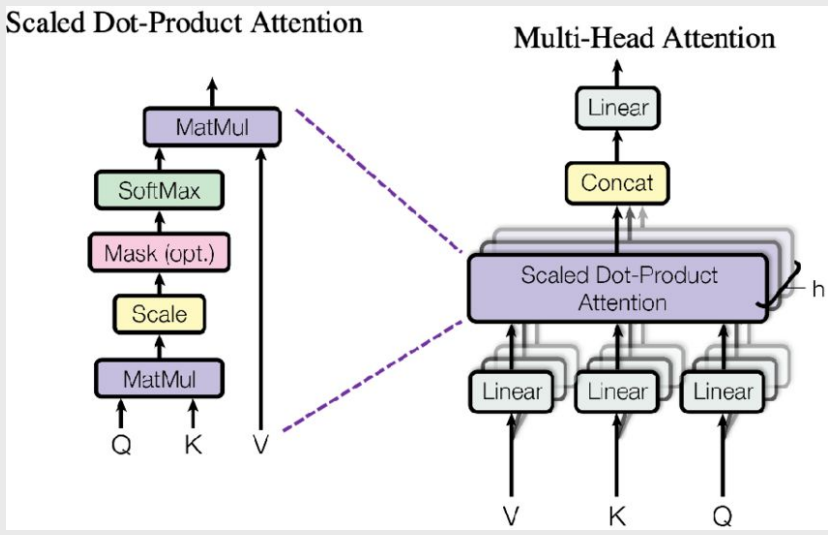
### Why AT in Classification Tasks?

- 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 : *********