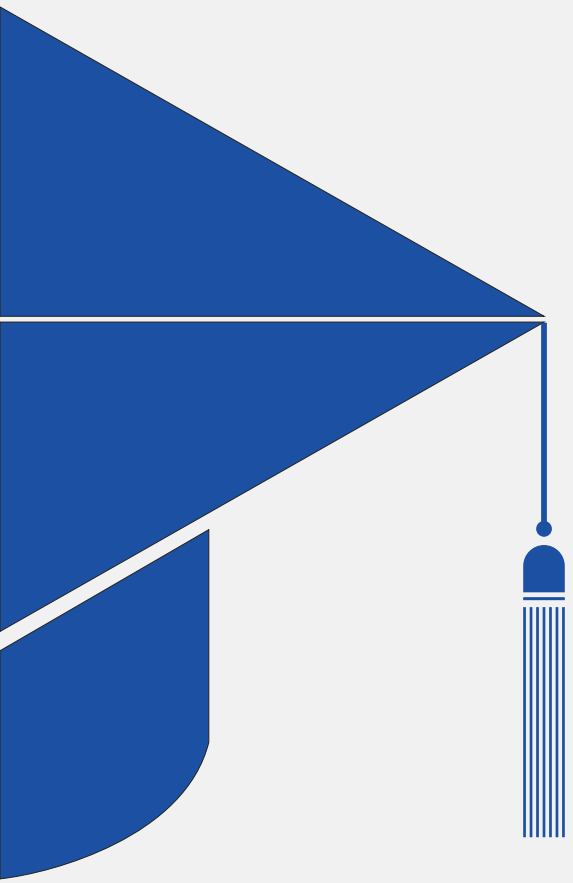


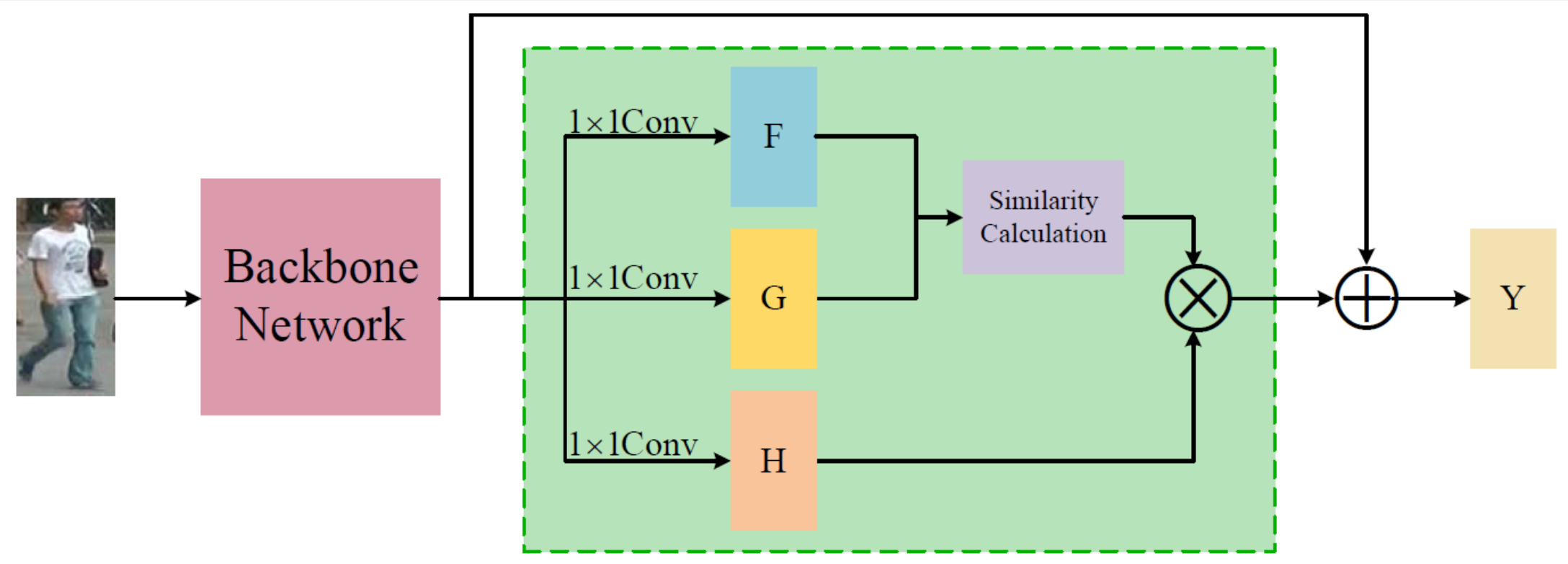
Decoupled Self-attention Module for Person Re-identification

——Chao Zhao, Zhenyu Zhang, Jian Yang, Yan Yan

- **The necessity of introducing self-attention mechanism**

  - Convolution neural network can only capture the correlation of local location.
  - Self-attention can reduce the impact of background noise, let the network focus on the human body, and focus on learning body feature.
  - Self-attention can reduce the influence of illumination change, because the location with high similarity to activation value should also have high activation value.

- **The traditional form of self-attetion (SA)**



$$SA\left(F,G,H\right)=softmax\left(FG^{\top}\right)H$$

where $F,G,H\in\mathbb{R}^{hw\times c}$ is feature maps after using 1x1 convolution onthe original feature maps.

- **The traditional form of self-attetion (SA)**

$$SA(F, G, H) = softmax\left(FG^{\top}\right)H$$

$$F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_{hw} \end{bmatrix} \in \mathbb{R}^{hw \times c}$$

The similarity between F_i and G_j is

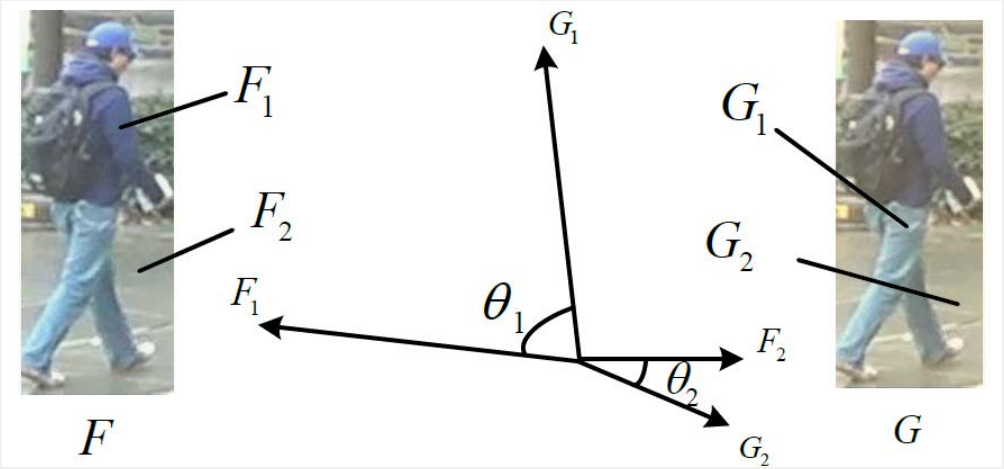
$$Sim(F_i, G_j) = F_i G_j^{\top} = \langle F_i, G_j \rangle$$

The above formula can be further written as

$$Sim(F_i, G_j) = \|F_i\| \|G_j\| \cos\langle F_i, G_j \rangle$$

$$Sim(F_i, G_j) = \|F_i\| \|G_j\| \cos\langle F_i, G_j \rangle$$

In this form, the norm reflects the activation degree of a location on the feature map, and cosine similarity reflects whether the two locations have similar semantic information.



Semantic response and semantic difference have totally different meanings, but they are mingled together in similarity calculation in self-attention.

- **The reasons of decoupling self-attention module**
  - In order to better analyze the reason of the instability of self attention mechanism in person Re-ID.
  - Semantic response and semantic relevance are mingled together in the general form of similarity calculation in self-attention, making the loss of independence.

The similarity matrix in the self-attention model is

$$Sim = FG^{\top} = \begin{bmatrix} \langle F_1, G_1 \rangle & \cdots & \langle F_1, G_{hw} \rangle \\ & \ddots & \vdots \\ \langle F_{hw}, G_1 \rangle & & \langle F_{hw}, G_{hw} \rangle \end{bmatrix}$$

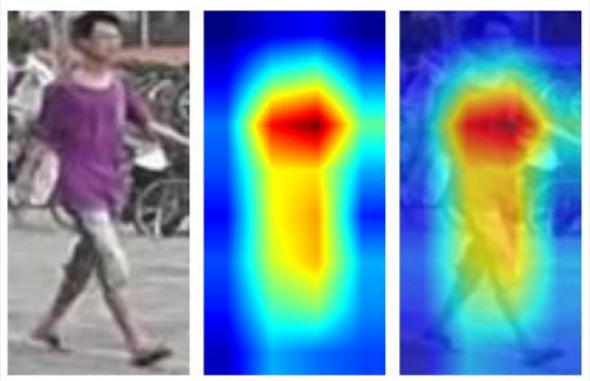
To simplify the form of the self-attention model and make it easier to understand, the softmax function for normalization is omitted

$$Y = \begin{bmatrix} \langle F_1, G_1 \rangle & \cdots & \langle F_1, G_{hw} \rangle \\ & \ddots & \vdots \\ \langle F_{hw}, G_1 \rangle & & \langle F_{hw}, G_{hw} \rangle \end{bmatrix} \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_{hw} \end{bmatrix}$$

For location i ($1 \leq i \leq hw$), the feature after SA will be

$$Y_i = \sum_{j=1}^{hw} (F_i G_j^\top) H_j = \sum_{j=1}^{hw} \langle F_i, G_j \rangle H_j$$
$$= \|F_i\| \sum_{j=1}^{hw} \|G_j\| \cos \langle F_i, G_j \rangle H_j$$

It shows that the feature in location i obtained from the self-attention mechanism is essentially a weighted summation based on the inner-product similarity between the features at all locations and the current location i.



- **Reasons for instability of self-attention module in person Re-ID**
  - There is a big difference between the norm of semantic response degree and the scale of cosine which represents semantic relevance, so the contribution to similarity calculation is quite different

$$Sim(F_i, G_j) = \|F_i\| \|G_j\| \cos\langle F_i, G_j \rangle$$

  - Once there is a huge norm $\|G_j\|$, the similarity between location j and all locations will be larger than others, especially after softmax function, then the feature will only depend on G_j , which affects back propagation, leads to instability of self-attention module.

$$Y_i = \sum_{j=1}^{hw} (F_i G_j^\top) H_j = \sum_{j=1}^{hw} \langle F_i, G_j \rangle H_j$$

$$= \|F_i\| \sum_{j=1}^{hw} \|G_j\| \cos\langle F_i, G_j \rangle H_j$$

In order to fully utilize correlation and balance the contribution of norms and angle in similarity calculation, we propose a generic form as follows:

$$Sim(F_i, G_j) = f(\|F_i\|, \|G_j\|) \cdot g(cos\ \langle F_i, G_j \rangle)$$

then the form of the decoupled self-attention module will be

$$Y_i = \sum_{j=1}^{hw} softmax(f(\|F_i\|, \|G_j\|) \cdot g(cos\ \langle F_i, G_j \rangle))\ H_j$$

- **Activation Function of Norms**
  - **Logarithm Norm**

$$Sim(F_i, G_j) = log(1 + \|F_i\|\|G_j\|)g(cos\langle F_i, G_j\rangle)$$

  - **Scaled Norm**

$$\begin{aligned} f(\|F_i\|, \|G_j\|) &= \sqrt{\frac{1}{c}\sum_{k=1}^{c} F_{ik}^2}\sqrt{\frac{1}{c}\sum_{k=1}^{c} G_{jk}^2} \\ &= \frac{1}{\sqrt{c}}\|F_i\|\frac{1}{\sqrt{c}}\|G_j\| = \frac{1}{c}\|F_i\|\|G_j\| \end{aligned}$$

$$Sim(F_i, G_j) = \frac{1}{c}\|F_i\|\|G_j\|g(cos\langle F_i, G_j\rangle)$$

  - **Non Norm**

$$Sim(F_i, G_j) = g(cos\langle F_i, G_j\rangle)$$

- **The angular Activation Function**
  - **The cosine angular activation function**

$$g(cos\left\langle F_i, G_j\right\rangle) = cos\left\langle F_i, G_j\right\rangle$$

  - **The square cosine (SqCosine) angular activation function**

$$g(cos\left\langle F_i, G_j\right\rangle) = sign\left(cos\left\langle F_i, G_j\right\rangle\right) \cdot cos^2\left\langle F_i, G_j\right\rangle$$

Advantages of our decoupled self-attention module

- Self-attention is decoupled into norms describing semantic response degree and cosine describing semantic difference, which can help us further understand how each part affects the performance of self-attention.
- Decoupled self-attention module allow us to employ various functions to better model semantic response degree and semantic difference, which introduces more nonlinearity to enhance the generalization ability and robustness of the model. Without extra parameters, the decoupled self-attention module can achieve better performance and stability than original self-attention by adopting rational functions.
- Our decoupled self-attention module is very flexible and architecture-agnostic, which can be easily added to any stages in any frameworks.

- **Results**

TABLE I

COMPARISON OF DIFFERENT FORMS OF THE DECOUPLED SELF-ATTENTION MODULE WITH THE BACKBONE NETWORK OF RESNET50 ON MARKET-1501 AND DUKEMTMC-REID. WE COMBINE DIFFERENT ACTIVATION FUNCTIONS OF NORMS AND ANGLE.

Models	Market-1501				DukeMTMC-ReID			
	rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
ResNet Baseline	87.6	95.4	98.0	71.9	79.4	88.9	94.3	60.6
LogNorm+Cosine	**92.3**	97.0	98.7	75.2	**82.8**	90.8	94.8	62.6
ScaledNorm+Cosine	92.0	96.4	98.8	75.4	82.4	**91.3**	**95.0**	**64.1**
NonNorm+Cosine	91.8	**97.3**	**98.9**	**75.8**	82.7	91.1	94.8	63.5
LogNorm+Sqcosine	91.8	96.6	98.5	75.4	81.9	90.4	94.0	62.8
ScaledNorm+SqCosine	91.3	96.9	98.7	75.4	82.4	**91.4**	**94.9**	**63.9**
NonNorm+SqCosine	**92.3**	**97.1**	**98.8**	**75.8**	**82.7**	90.8	**94.9**	63.6

TABLE III

COMPARISON OF DIFFERENT FORMS OF THE DECOUPLED SELF-ATTENTION MODULE WITH THE BACKBONE NETWORK OF DENSENET121 ON MARKET-1501 AND DUKEMTMC-REID. WE COMBINE DIFFERENT ACTIVATION FUNCTIONS OF NORMS AND ANGLE.

Models	Market-1501				DukeMTMC-ReID			
	rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
DenseNet Baseline	90.5	96.1	98.4	73.8	81.2	90.1	94.3	61.4
LogNorm+Cosine	90.9	96.5	98.5	74.7	82.6	**90.8**	94.4	62.2
ScaledNorm+Cosine	**91.1**	**96.8**	**98.7**	**74.9**	82.2	**90.8**	94.3	**63.6**
NonNorm+Cosine	**91.1**	96.6	98.6	**74.9**	**82.8**	90.3	**94.8**	63.1
LogNorm+Sqcosine	90.7	96.5	98.5	74.5	81.8	90.9	94.5	63.0
ScaledNorm+SqCosine	**91.7**	**96.7**	**98.6**	**75.9**	**82.1**	**91.1**	94.5	63.2
NonNorm+SqCosine	91.0	96.3	98.4	75.1	**82.1**	90.6	**94.6**	**63.4**

《Decoupled Self-attention Module for Person Re-identification》

- **Results**

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART MODELS ON MARKET-1501 AND DUKEMTMC-REID, WHERE ECN IS A DOMAIN ADAPTATION MODEL.

Models	Market		DukeMTMC	
	rank-1	mAP	rank-1	mAP
ECN [35]	75.1	43.0	63.3	40.4
ECN+Decoupled module	**75.5**	**43.8**	**63.9**	**40.5**
OSNet [36]	94.8	84.9	88.6	73.5
OSNet+Decoupled module	**95.1**	**85.2**	**88.8**	**75.3**
MHN [12]	94.8	85.2	89.5	77.5
MHN+Decoupled module	**95.1**	**85.9**	**89.8**	**78.1**

TABLE IV
ABLATION STUDY FOR EACH PART OF SIMILARITY CALCULATION IN SELF-ATTENTION ON MARKET-1501 AND DUKEMTMC-REID. THE BACKBONE NETWORK IS RESNET50. NO$\|F_i\|$, NO$\|G_j\|$ AND NOCOSINE DENOTE THAT THERE IS NO $\|F_i\|$, NO $\|G_j\|$ AND NO COSINE IN SIMILARITY CALCULATION, RESPECTIVELY.

Models	Market		DukeMTMC	
	rank-1	mAP	rank-1	mAP
Original SA	82.3	61.7	54.4	32.8
No$\|F_i\|$	90.5	73.3	82.2	62.0
No$\|G_j\|$	90.5	73.7	81.4	61.6
NoCosine	41.9	23.0	44.3	26.6
NonNorm	**91.1**	**74.9**	**82.8**	**63.1**

TABLE V
COMPARISON WITH OTHER FORMS OF SELF-ATTENTION WHICH HAS BEEN PROPOSED ON MARKET-1501 AND DUKEMTMC-REID. THE BACKBONE NETWORK IS RESNET50 HERE. DP AND SDP DENOTE DOT-PRODUCT AND SCALED DOT-PRODUCT, RESPECTIVELY.

Models	Market		DukeMTMC	
	rank-1	mAP	rank-1	mAP
Original SA	82.3	61.7	54.4	32.8
DP [15]	91.2	74.6	81.9	63.3
SDP [14]	89.7	71.8	81.4	61.6
ScaledNorm	**92.0**	**75.4**	**82.4**	**64.1**

- **Results**

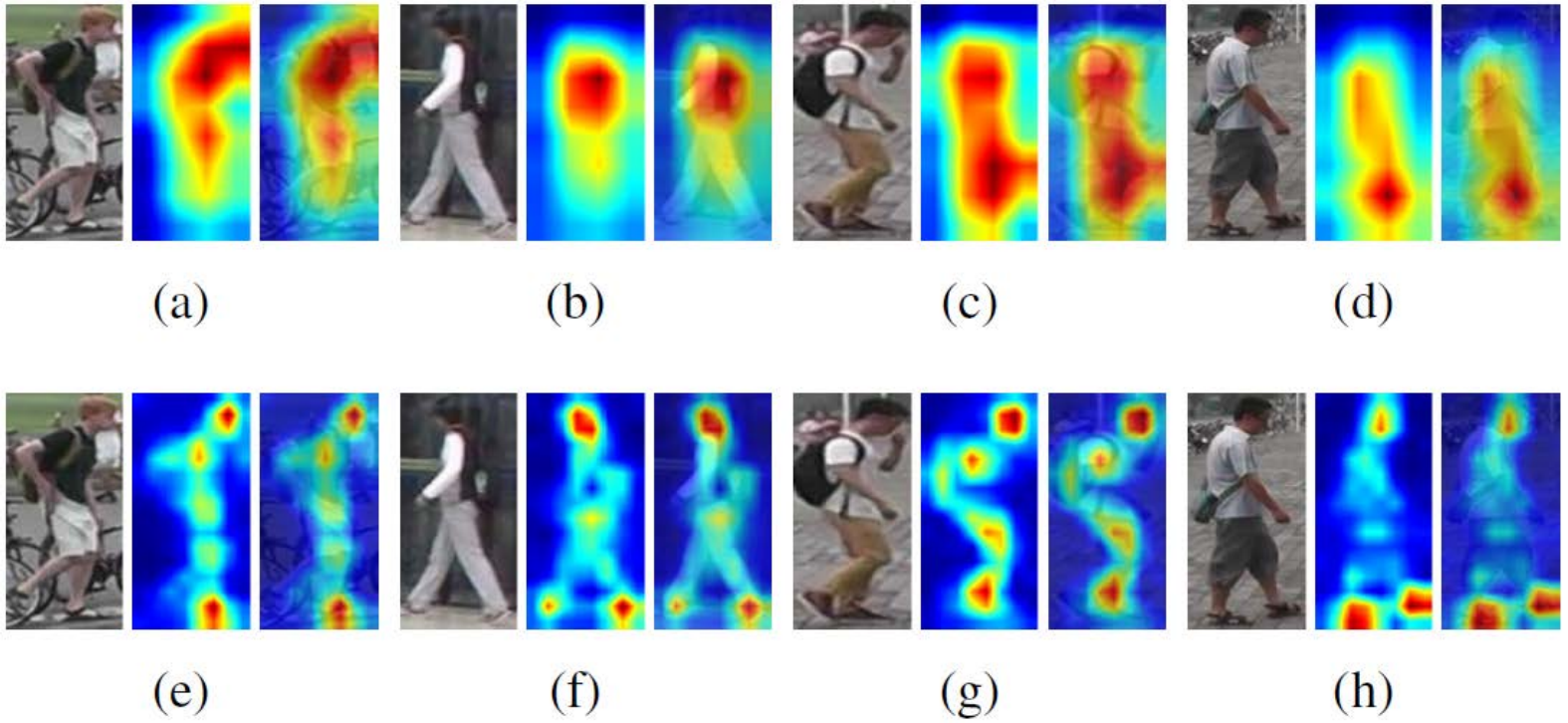
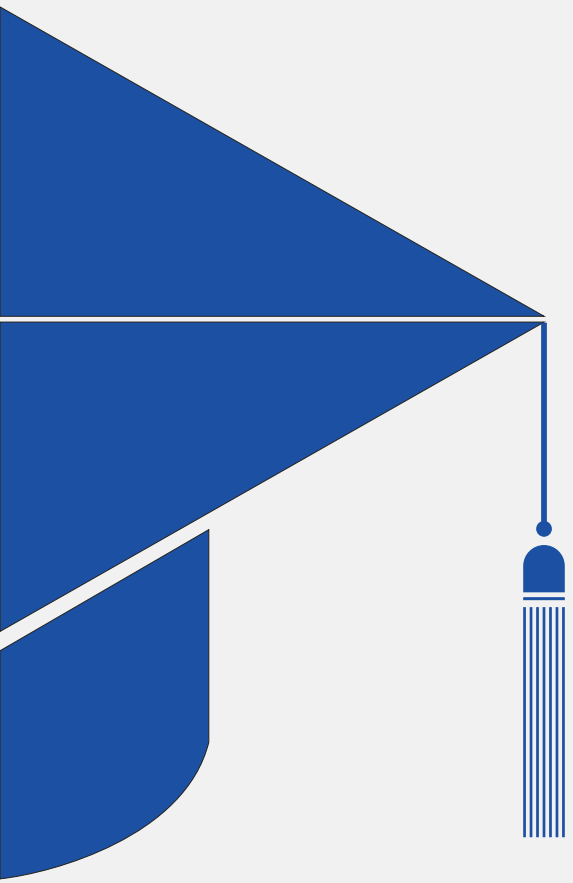


Fig. 5. Activation maps of the model w or w/o the decoupled self-attention module. The first row is activation maps without the decoupled self-attention module, and the second row is activation maps with the decoupled self-attention module.



Thank you !