Disentangled Self-Attention Models

Han Hu (胡瀚)

Visual Computing Group

Microsoft Research Asia (MSRA)

September 8th, 2020

On behalf of authors: Minghao Yin*, Zhuliang Yao*, Yue Cao, Xiu Li, Zheng Zhang, Steve Lin and Han Hu

Outline

- A Brief Introduction of Self-Attention Models
- The Degeneration Problem and Diagnosis
- Approach and Results

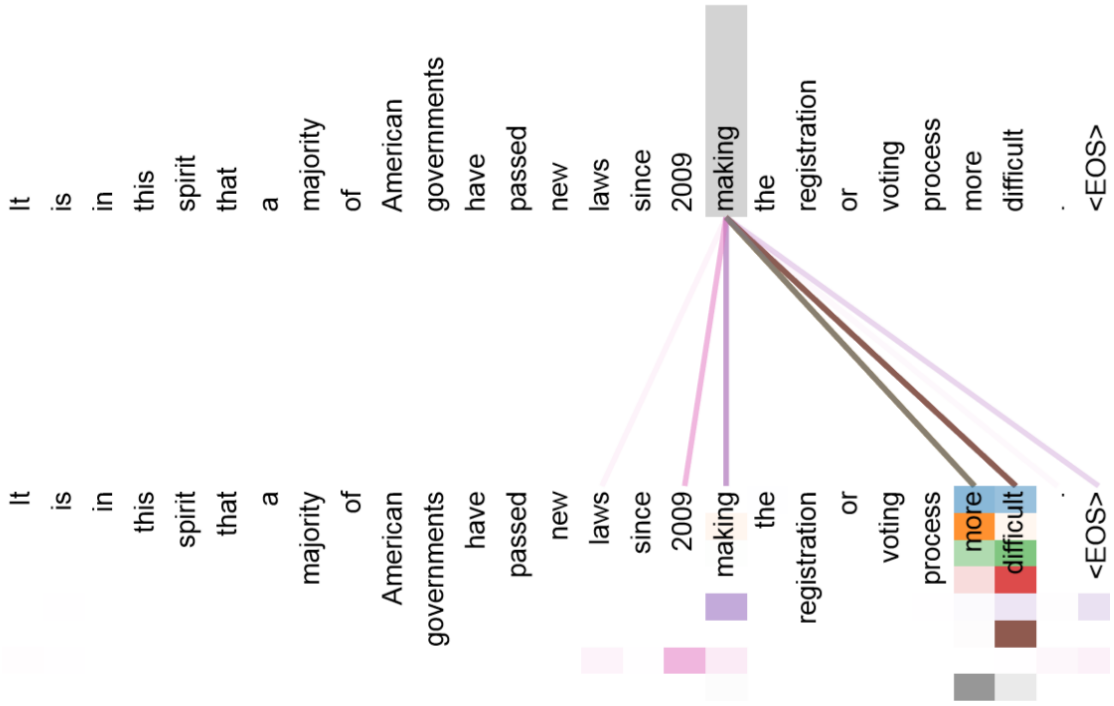
Outline

- **A Brief Introduction of Self-Attention Models**
- The Degeneration Problem and Diagnosis
- Approach and Results

Self-Attention Models Dominate the NLP Field

- Transformer (Google)

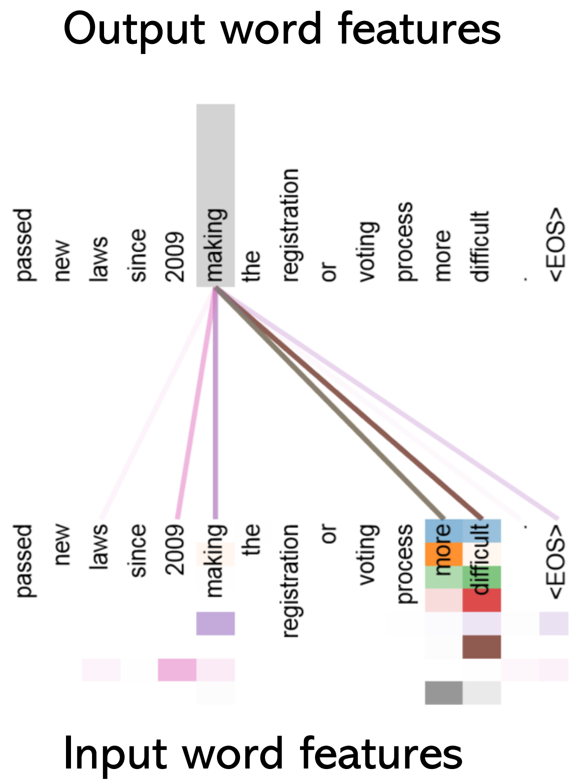
- GPT (Open AI)
- BERT (Google)
- MASS, UniLM, VL-BERT (MSRA)

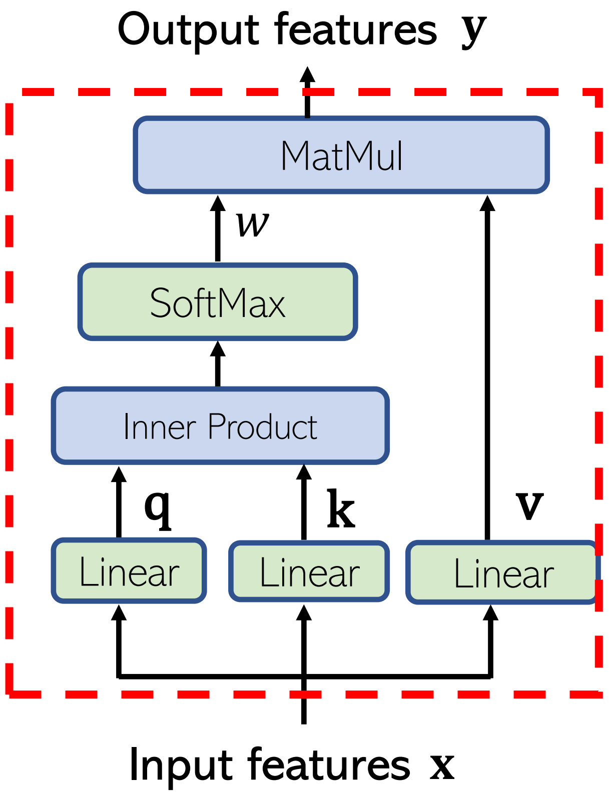


Ashish Vaswani et al, Attention is all you need, NeurIPS'2017

What is a Self-Attention Module?

- Transforms the word/token input feature by encoding its relationship with other words/tokens
- A weighted average of **Value**, where the weight is the normalized inner product of **Query** and **Key**



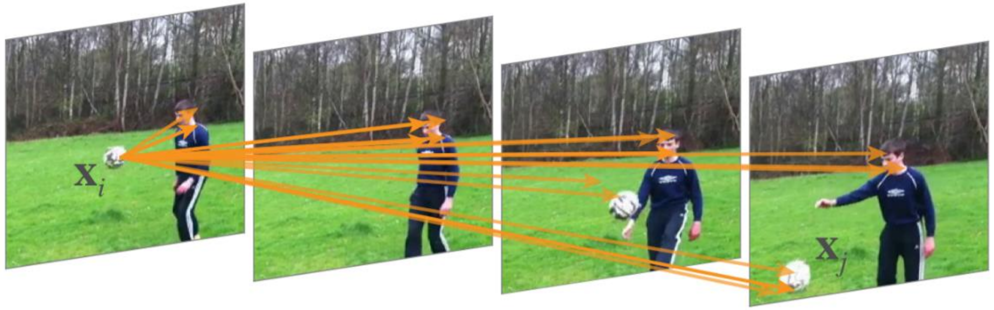


$$\mathbf{y}_i = \sum_{j\in\Omega} w(\boldsymbol{q}_i, \mathbf{k}_j)\,\mathbf{v}_j$$

$$w(\mathbf{q}_i, \mathbf{k}_j) \sim exp(\mathbf{q}_i^T \mathbf{k}_j)$$

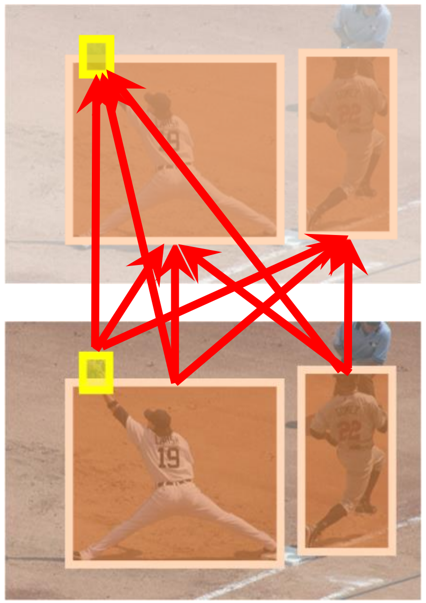
Two Pioneer Works in Vision

Non-Local Neural Networks [CVPR'2018]



- ✓ Inserted in backbone networks to complement convolution
- ✓ Improves various applications: object detection, semantic segmentation, action recognition and etc

Relation Networks [CVPR'2018]



- ✓ Models Object-to-Object Relationship
- ✓ The first fully end-to-end object detector

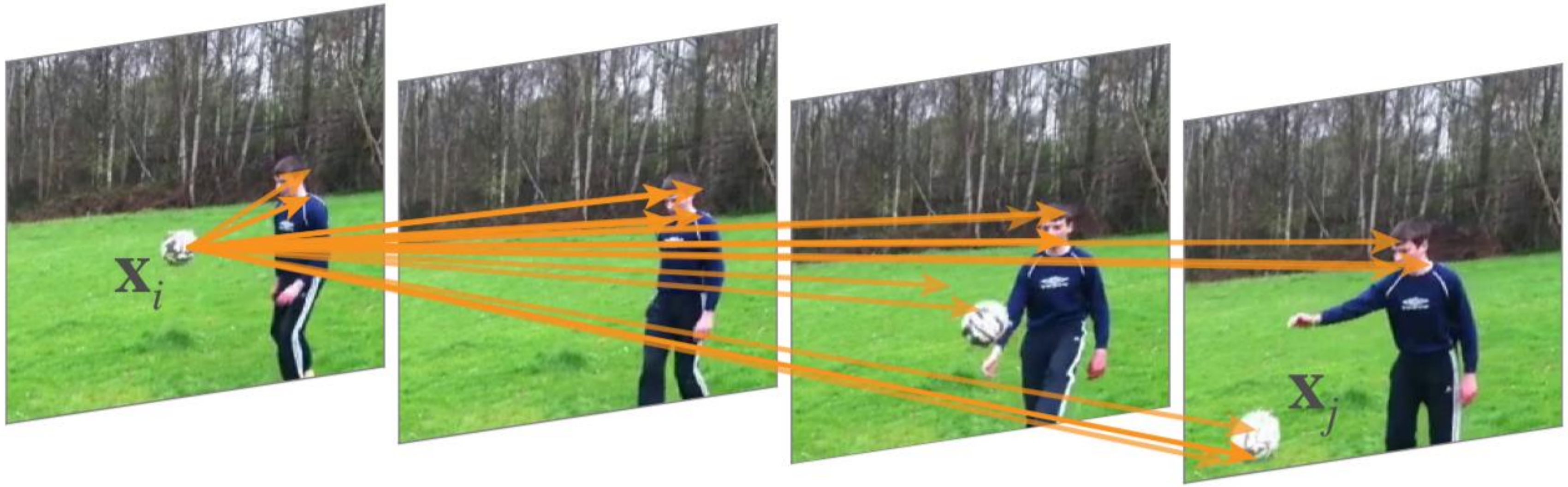
Summary of Representative Works

- Pixel-to-Pixel Relationship
  - Non-Local Neural Networks [CVPR'2018]
  - Local Relation Networks [ICCV'2019]
  - Standard-Alone Self-Attention Models [NeurIPS'2019]
- Object-to-Pixel Relationship
  - Learning Region Features [ECCV'2018]
  - End-to-End Object Detector (DETR) [ECCV'2020]
- Object-to-Object Relationship
  - Relation Networks [CVPR'2019]
  - Various Video Applications
    - Video Action Recognition, Multi-Object Tracking, Video Object Detection

Outline

- A Brief Introduction of Self-Attention Models
- **The Degeneration Problem and Diagnosis**
- Approach and Results

Self-Attention Encodes **Pairwise** Relationship



Does it learn pairwise relationship well?

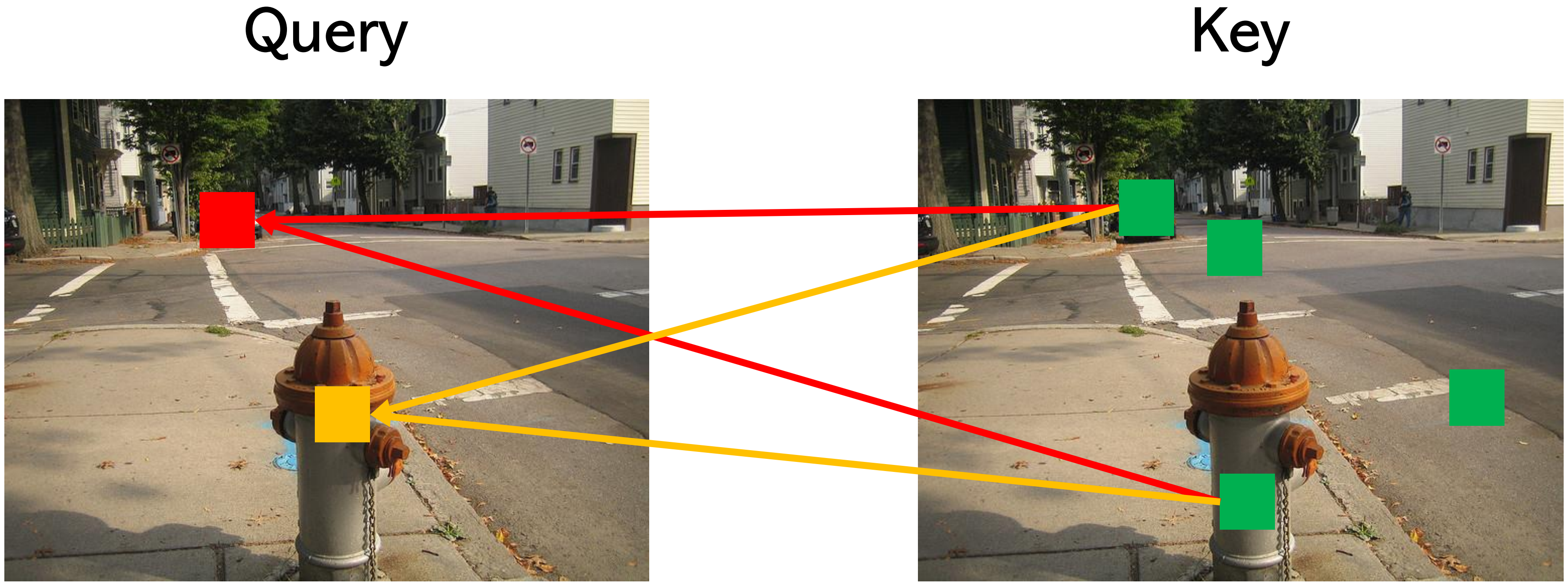
Expectation of Learnt Relation

- Different queries affected by **different** key

Query **Key**

What does the Self-Attention Learn?

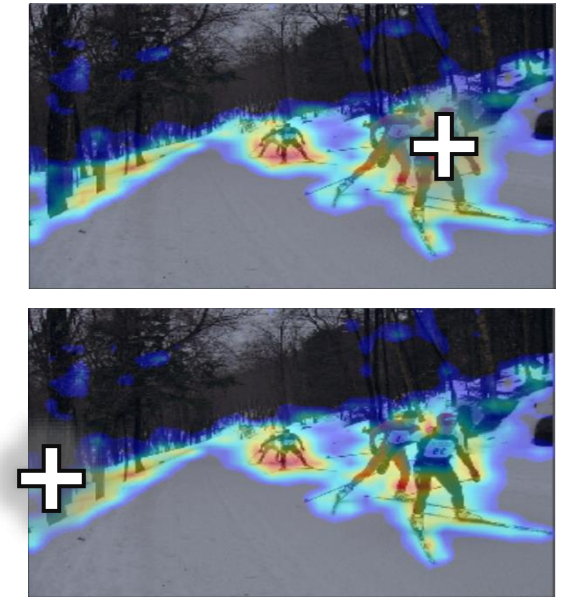
- Different queries affected by the **same** keys
- **Pairwise** in expectation → **Unary** in actual



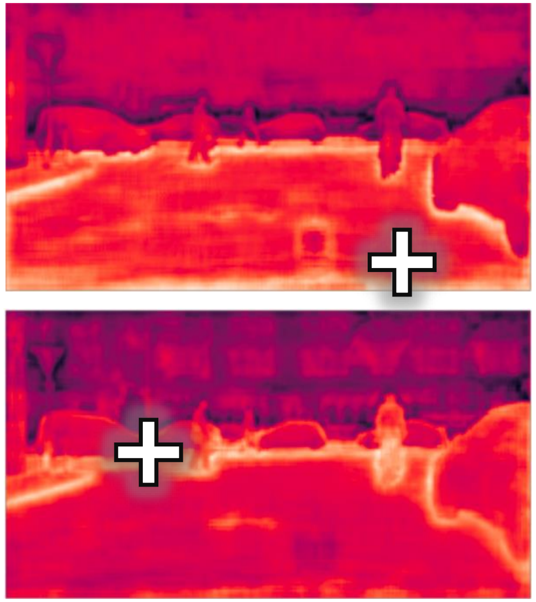
Yue Cao*, Jiarui Xu* , Stephen Lin, Fangyun Wei and **Han Hu**. ***GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond***. ICCVW'2019

Visualizations on Real Tasks

- indicates the query point
- The activation map for different queries are similar
- The self-attention model degenerates to a unary model



Object Detection



Semantic Segmentation

[GCNet, ICCVW'2019]

https://arxiv.org/pdf/1904.11492.pdf

WHY?

Revisit Self-Attention Formulation

- The self-attention formulation has a '*hidden*' unary term:

$$w(\mathbf{q}_i, \mathbf{k}_j) \sim exp(\mathbf{q}_i^T \mathbf{k}_j) = exp\Big(\underbrace{(\mathbf{q}_i - \boldsymbol{\mu}_q)^T (\mathbf{k}_j - \boldsymbol{\mu}_k)}_{\text{(whitened) pairwise}} + \underbrace{\boldsymbol{\mu}_q^T \mathbf{k}_j}_{\text{(hidden) unary}}\Big)$$

* $\boldsymbol{\mu}_q$ and $\boldsymbol{\mu}_k$ are global average of $\mathbf{q}$ **and** $\mathbf{k}$

Behavior of the Pairwise and Unary Terms

method	fomulation	mIoU
Baseline	none	75.8%
Joint (Self-Attention)	$\sim exp\left(\mathbf{q}_i^T \mathbf{k}_j\right)$	$\boxed{78.5\%}$
Pairwise Alone	$\sim exp\left((\mathbf{q}_i - \boldsymbol{\mu}_q)^T (\mathbf{k}_j - \boldsymbol{\mu}_k)\right)$	77.5%
Unary Alone	$\sim exp\left(\boldsymbol{\mu}_q^T \mathbf{k}_j\right)$	$\boxed{\mathbf{79.3\%}}$

Quantitative results on semantic segmentation (Cityscapes)

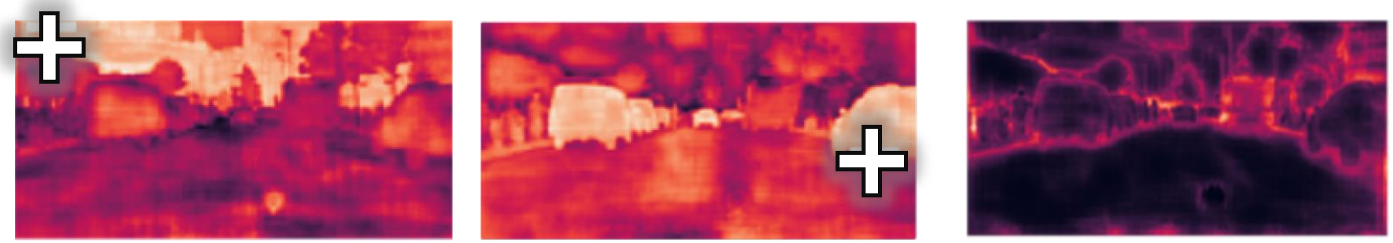
- The **unary** term alone outperforms **the standard joint model**
- The pairwise and unary terms are **not well learnt** when combined in the self-attention formulation

Visual Meaning of Each Term

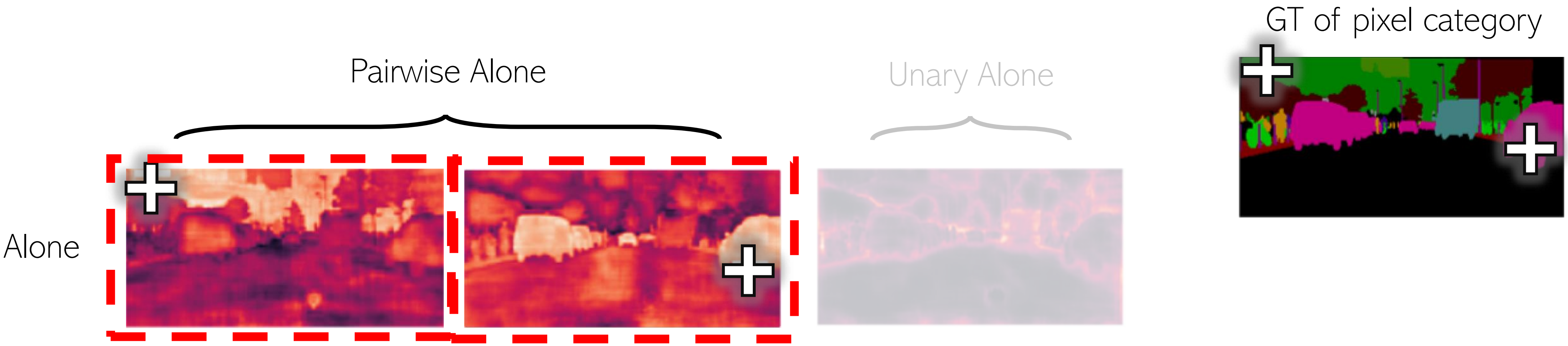
Pairwise Alone

Unary Alone

Alone

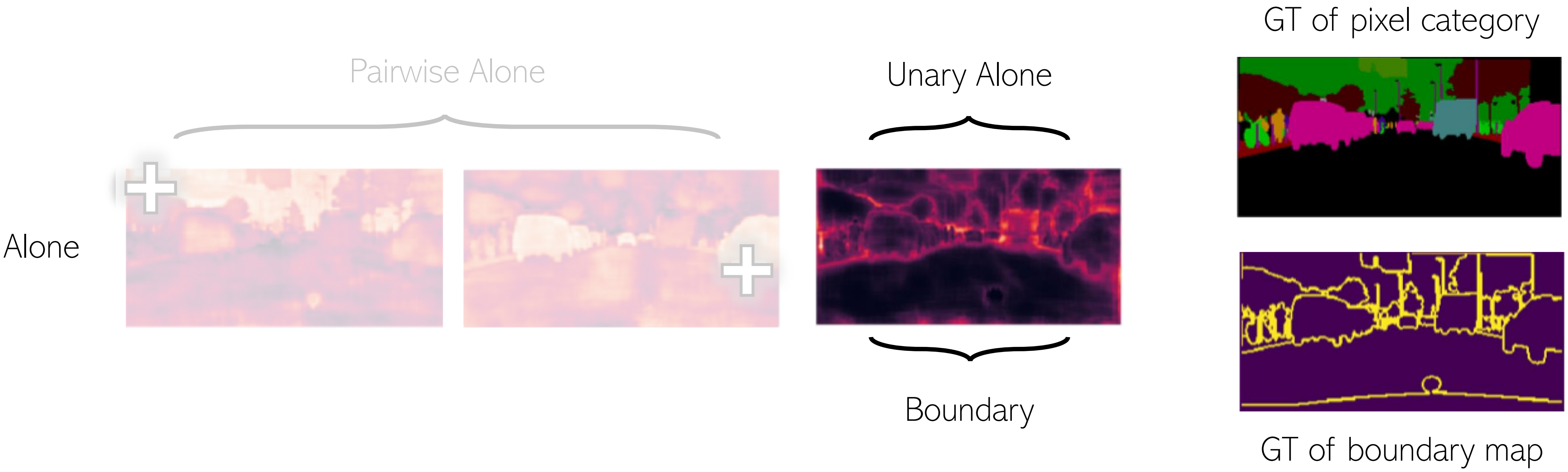


Visual Meaning of Each Term



- The pairwise term tends to learn relations within the **same category region**

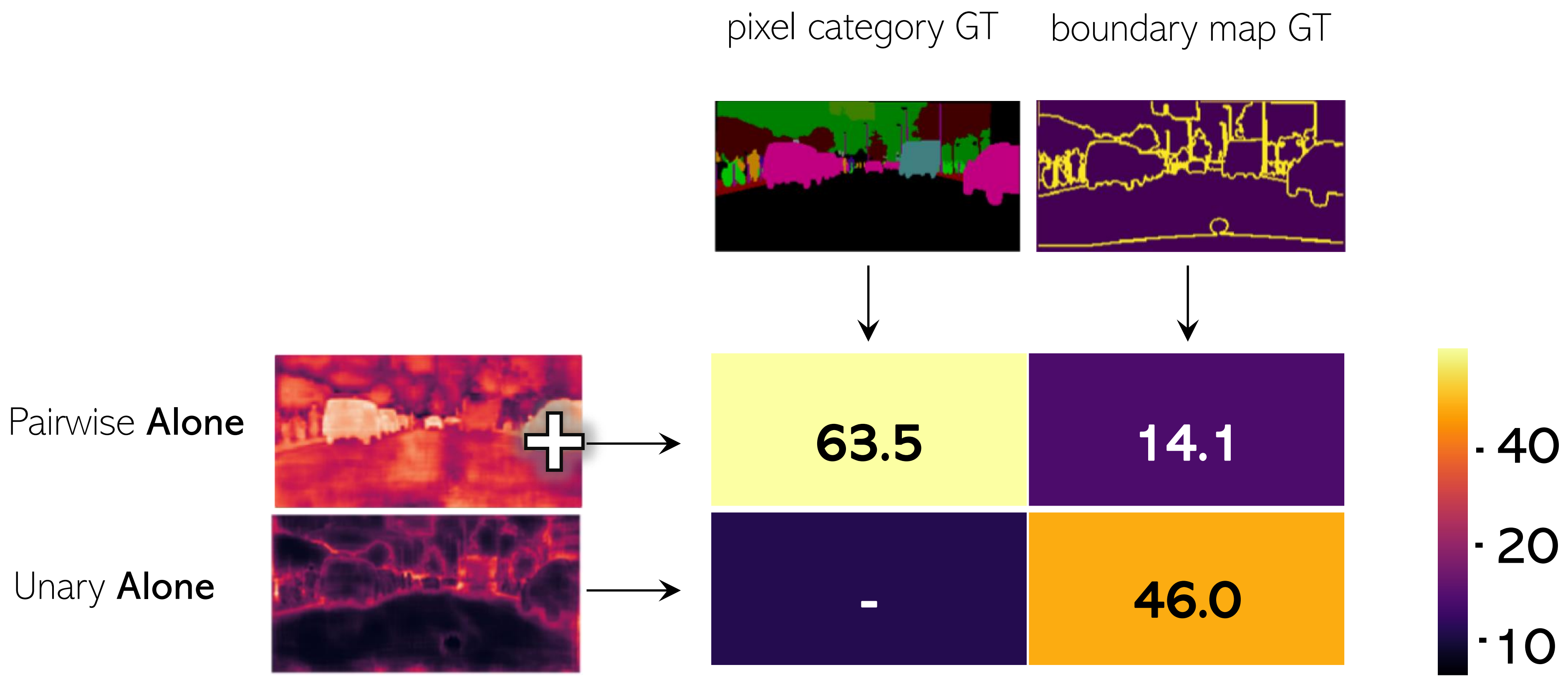
Visual Meaning of Each Term



- The pairwise term tends to learn relations within the **same category region**
- The unary term tends to focus on **boundary pixels**

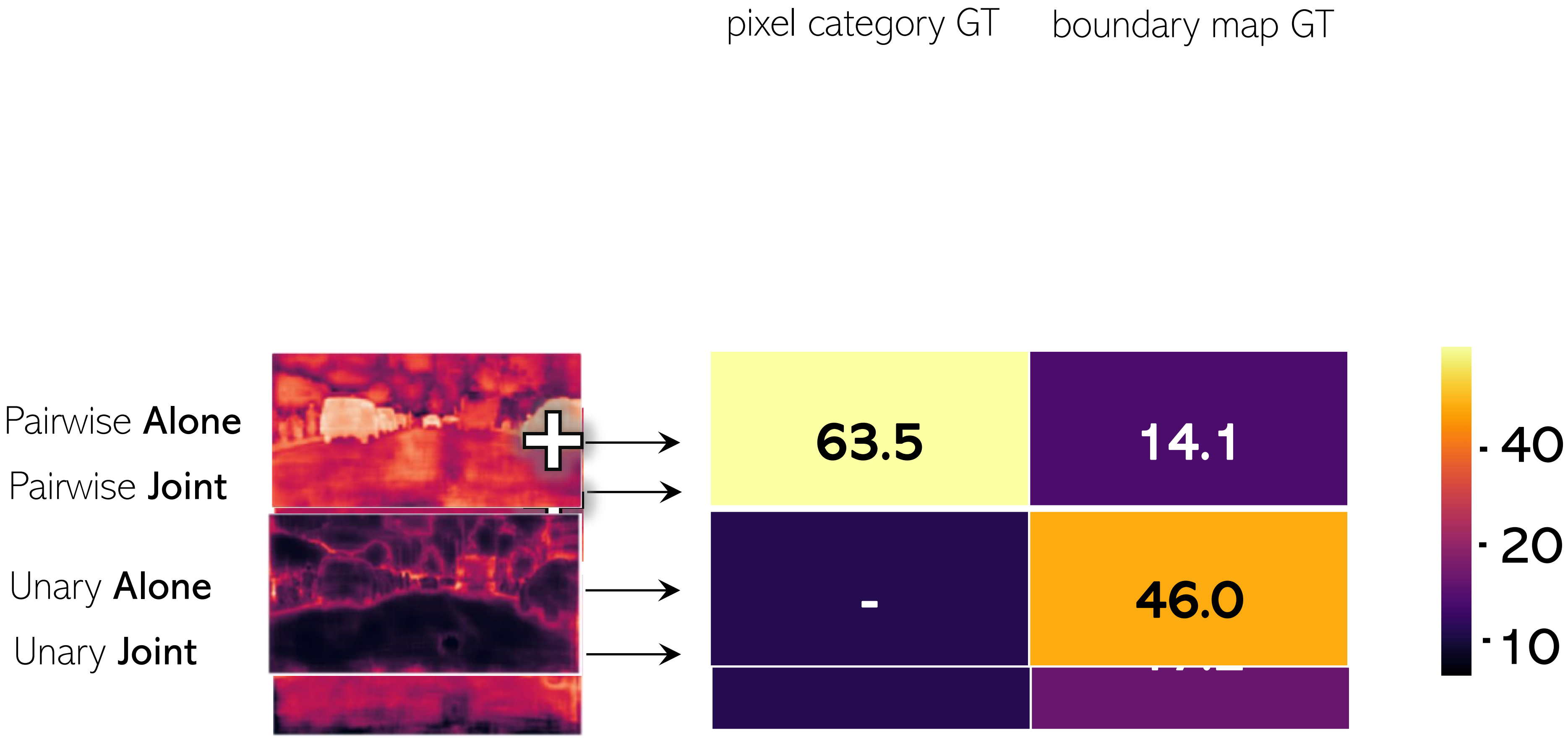
Visual Meaning of Each Term

- Statistical correlation



Comparison with Standard 'Joint' Model

- Statistical correlation



Why is 'Joint' Worse than 'Alone'?

- Self-Attention is the **multiplicative** combination of pairwise term ($\boldsymbol{w_p}$) and unary term ($\boldsymbol{w_u}$) :

$$w(\mathbf{q}_i, \mathbf{k}_j) \sim exp\left((\mathbf{q}_i - \boldsymbol{\mu}_q)^T(\mathbf{k}_j - \boldsymbol{\mu}_k) + \boldsymbol{\mu}_q^T \mathbf{k}_j\right)$$

$$= \underbrace{exp\left((\mathbf{q}_i - \boldsymbol{\mu}_q)^T(\mathbf{k}_j - \boldsymbol{\mu}_k)\right)}_{\text{Pairwise } \boldsymbol{w_p}} \times \underbrace{exp(\boldsymbol{\mu}_q^T \mathbf{k}_j)}_{\text{Unary } \boldsymbol{w_u}}$$

Combination by Multiplication is Bad

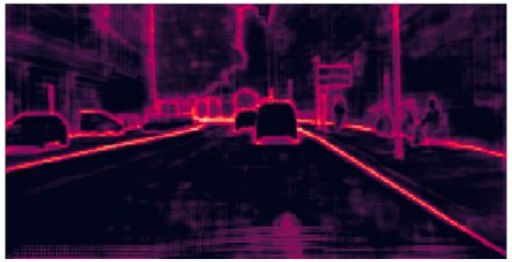
- Multiplication couples two terms in gradient computation

$$\boxed{\frac{\partial L}{\partial \boldsymbol{w_p}}} = \frac{\partial L}{\partial \boldsymbol{w}} \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{w_p}} \sim \frac{\partial L}{\partial \boldsymbol{w}} \boxed{\boldsymbol{w_u}} \qquad \boxed{\frac{\partial L}{\partial \boldsymbol{w_u}}} = \frac{\partial L}{\partial \boldsymbol{w}} \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{w_u}} \sim \frac{\partial L}{\partial \boldsymbol{w}} \boxed{\boldsymbol{w_p}}$$

- Multiplication acts like **intersection**, resulting in empty if two terms encode different visual clues

Pairwise
(Same category region)

$\cap$



Unary
(Boundary)

$=$

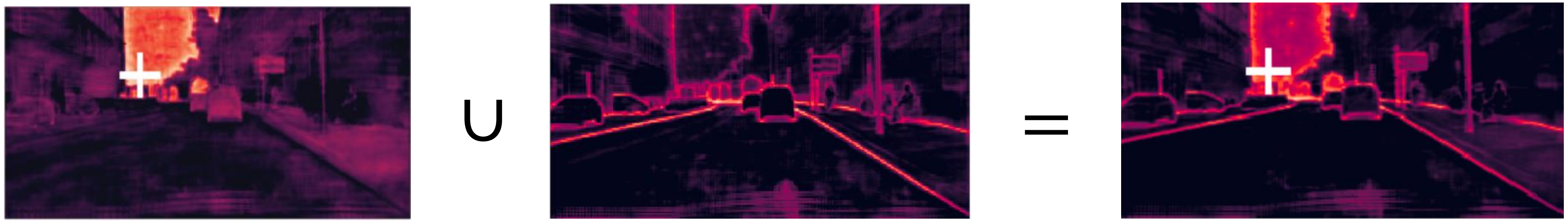
Empty

Outline

- A Brief Introduction of Self-Attention Models
- The Degeneration Problem and Diagnosis
- **Approach and Results**

From Intersection (Mul) to Union (Add)

- **Union** instead of intersection:



- Implement by **addition**

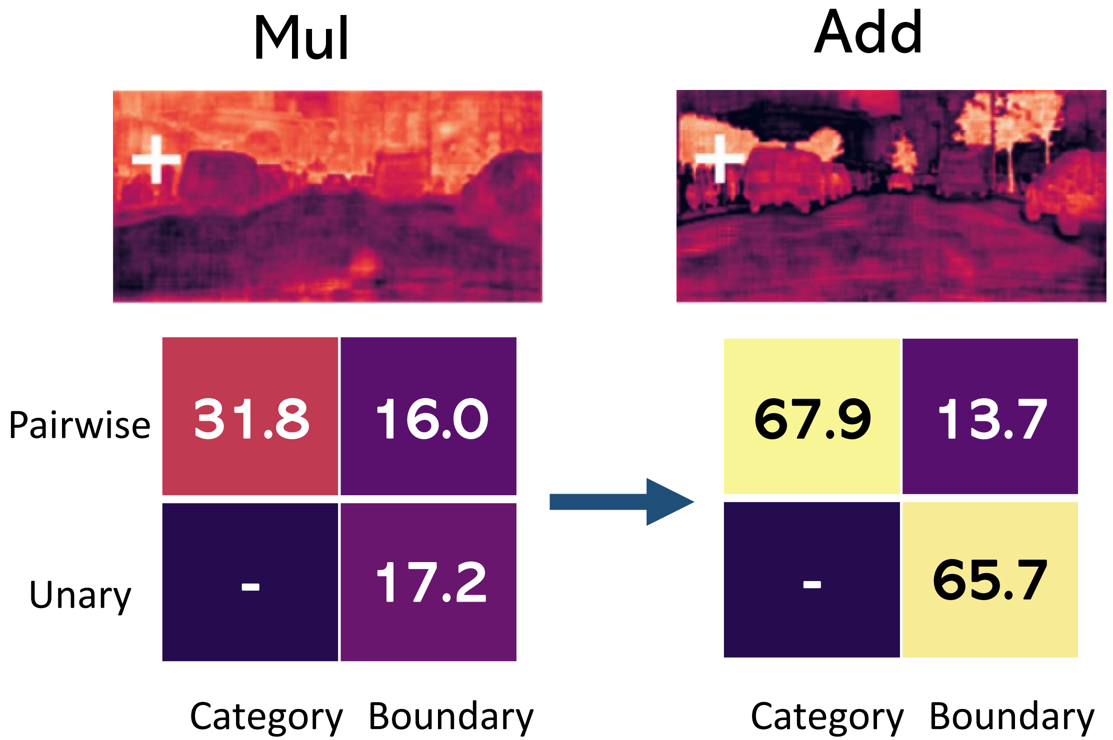
$$w(\mathbf{q}_i, \mathbf{k}_j) \sim exp\big((\mathbf{q}_i - \boldsymbol{\mu}_q)^T (\mathbf{k}_j - \boldsymbol{\mu}_k)\big) * exp(\boldsymbol{\mu}_q^T \mathbf{k}_j)$$

- Gradients are **disentangled** by **addition**

From Intersection (Mul) to Union (Add)

- 0.7 mIoU improvements on Cityscapes
- Significantly clearer visual meaning

method	mIoU
Baseline	75.8%
Mul (Self-Attention)	78.5%
Add (Ours)	**79.2%**



Are There Other Coupling Factors?

- The key is **shared** in the pairwise term and unary term
- The shared key can be further **disentangled**:

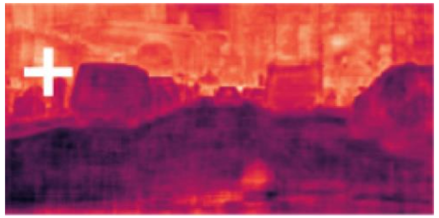
$$w(\mathbf{q}_i, \mathbf{k}_j) \sim \overbrace{exp\left((\mathbf{q}_i - \boldsymbol{\mu}_q)^T (\boxed{\mathbf{k}_j} - \boldsymbol{\mu}_k)\right)}^{\text{pairwise}} + \overbrace{exp(\boxed{\mathbf{k}_j})}^{\text{unary}}$$

$$exp\left((\mathbf{q}_i - \boldsymbol{\mu}_q)^T (\boxed{\mathbf{W}^{p}\mathbf{k}_j} - \boldsymbol{\mu}_k)\right) + exp(\boxed{\mathbf{W}^{u}\mathbf{k}_j})$$

Disentangle the Key Transformations

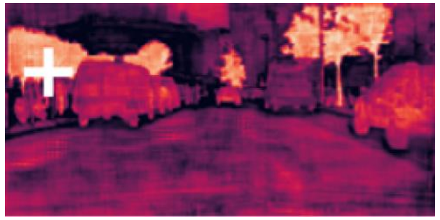
- The pairwise and unary terms learn clearer visual meaning

Mul



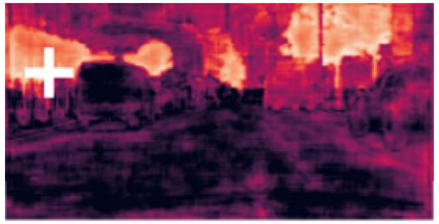
	Category	Boundary
Pairwise	31.8	16.0
Unary	-	17.2

Add (Key Shared)



	Category	Boundary
Pairwise	67.9	13.7
Unary	-	65.7

Add (Separate Keys)



	Category	Boundary
Pairwise	75.9	13.0
Unary	-	69.6

Results by Two Disentangle Techniques

- **2.0** mIoU improvements than self-attention
- **4.7** mIoU improvements than baseline

method	mIoU
Baseline	75.8%
Mul (Self-Attention)	78.5%
Add(Shared key)	79.2%
Add(Disentangled key)	**80.5%**

On Three Semantic Segmentation Benchmarks

- Disentangled Non-Local Neural Networks
  - Multiplication to Addition
  - Shared keys to Disentangled keys

method	backbone	mIoU(%)
Deeplab v3	ResNet101	81.3
OCNet	ResNet101	81.7
Self-Attention	ResNet101	80.8
Ours	ResNet101	**82.0**
HRNet	HRNetV2-W48	81.9
Self-Attention	HRNetV2-W48	82.5
Ours	HRNetV2-W48	**83.0**

Cityscapes

method	backbone	mIoU(%)
ANN	ResNet101	52.8
EMANet	ResNet101	53.1
Self-Attention	ResNet101	50.3
Ours	ResNet101	**54.8**
HRNet v2	HRNetV2-W48	54.0
Self-Attention	HRNetV2-W48	54.2
Ours	HRNetV2-W48	**55.3**

ADE20K

method	backbone	mIoU(%)
ANN	ResNet101	45.24
OCNet	ResNet101	45.45
Self-Attention	ResNet101	44.67
Ours	ResNet101	**45.90**
HRNet v2	HRNetV2-W48	42.99
Self-Attention	HRNetV2-W48	44.82
Ours	HRNetV2-W48	**45.82**

PASCAL-Context

Disentangled Non-Local Network is General

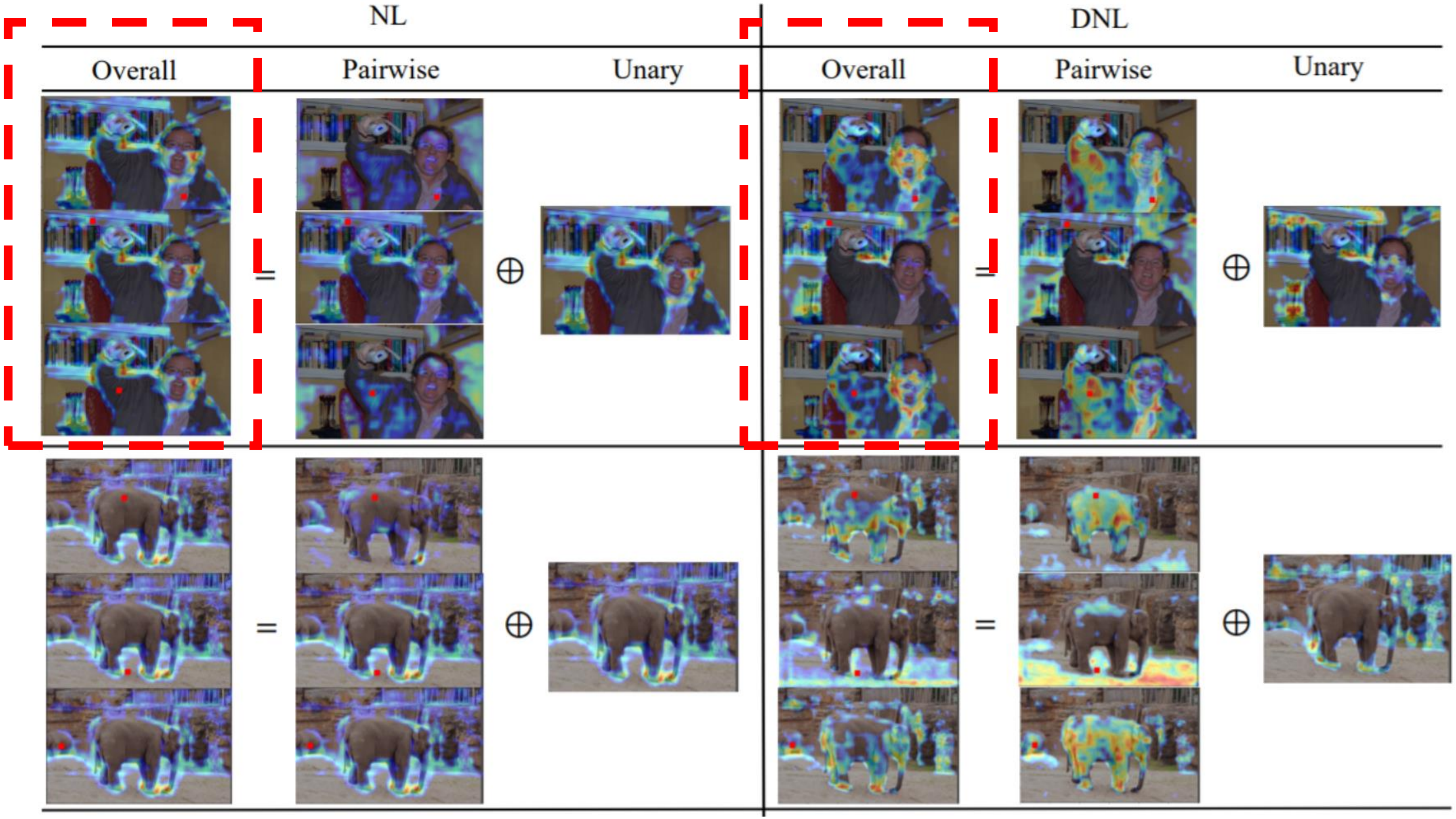
- Object detection & instance segmentation, COCO2017 dataset

method	$\mathrm{mAP}^{\mathrm{bbox}}$	$\mathrm{mAP}^{\mathrm{mask}}$
Baseline	38.8	35.1
Self-Attention	40.1	36.0
Disentangled Self-Attention (ours)	**41.4**	**37.3**

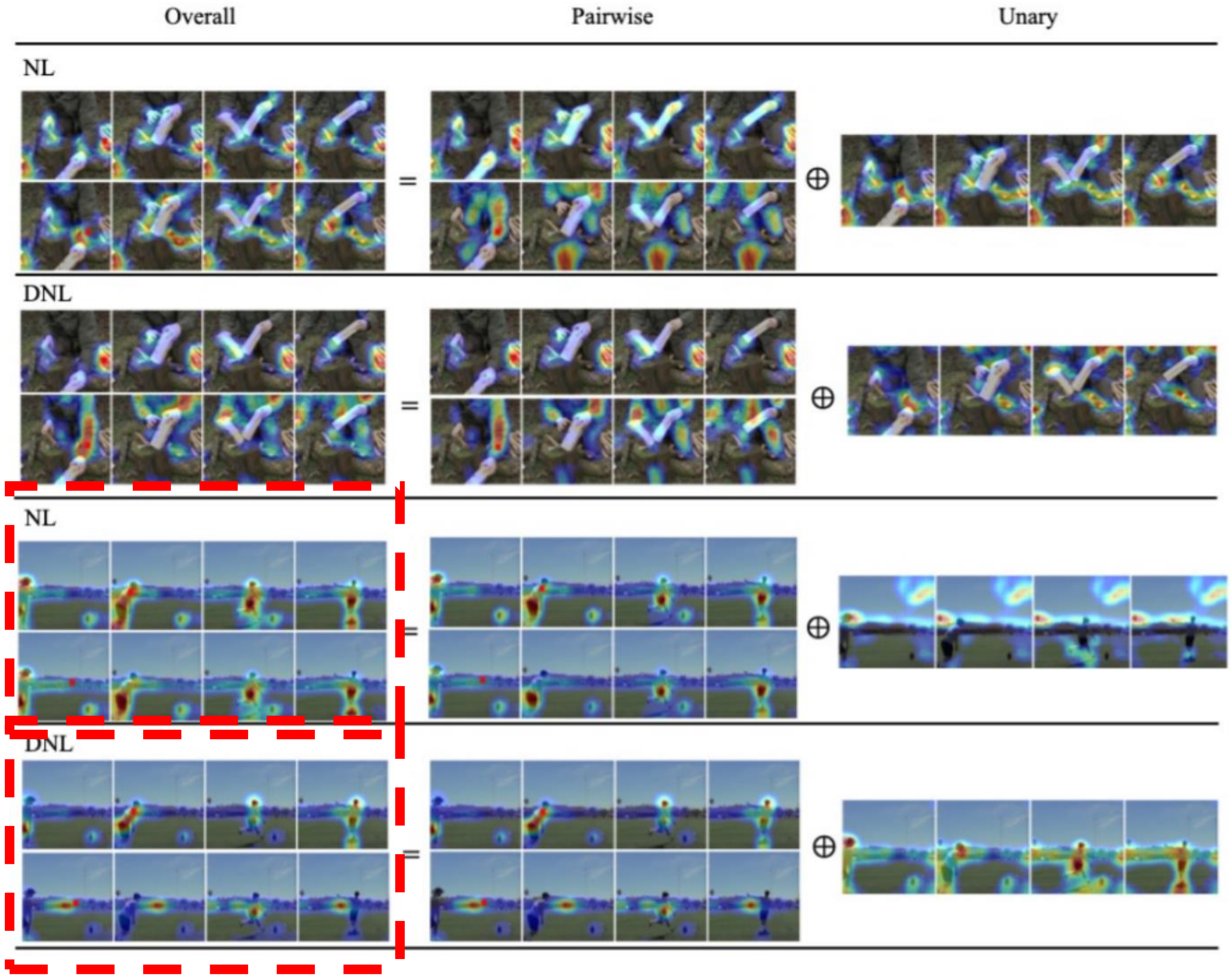
- Action recognition, Kinetics dataset

method	**Top-1 Acc**	**Top-5 Acc**
Baseline	74.9	91.9
Self-Attention	75.9	92.2
Disentangled Self-Attention (ours)	**76.3**	**92.7**

Visualization (Object Detection)



Visualization (Action Recognition)



Summary

- Are self-attention models learnt well on visual tasks?
  - **No [GCNet, ICCVW'2019],**
- How can it be more effective?
  - **Disentangled design [DNL, ECCV'2020]**

DNL code

Semantic Segmentation

Object Detection

in mmsegmentation

https://github.com/yinmh17/DNL-Semantic-Segmentation
https://github.com/Howal/DNL-Object-Detection
https://github.com/open-mmlab/mmsegmentation/tree/master/configs/dnlnet

Q&A