香港中文大學
The Chinese University of Hong Kong

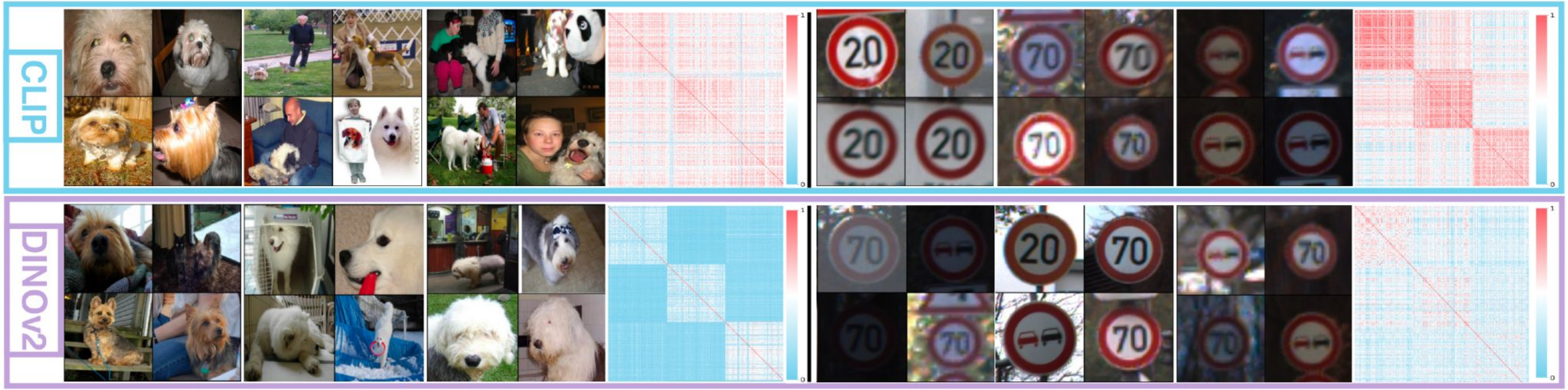
NEURAL INFORMATION PROCESSING SYSTEMS

When Kernels Multiply, Clusters Unify: Fusing Embeddings with the Kronecker Product

Youqi Wu, Jingwei Zhang, Farzan Farnia

Complementary Strengths of Embeddings

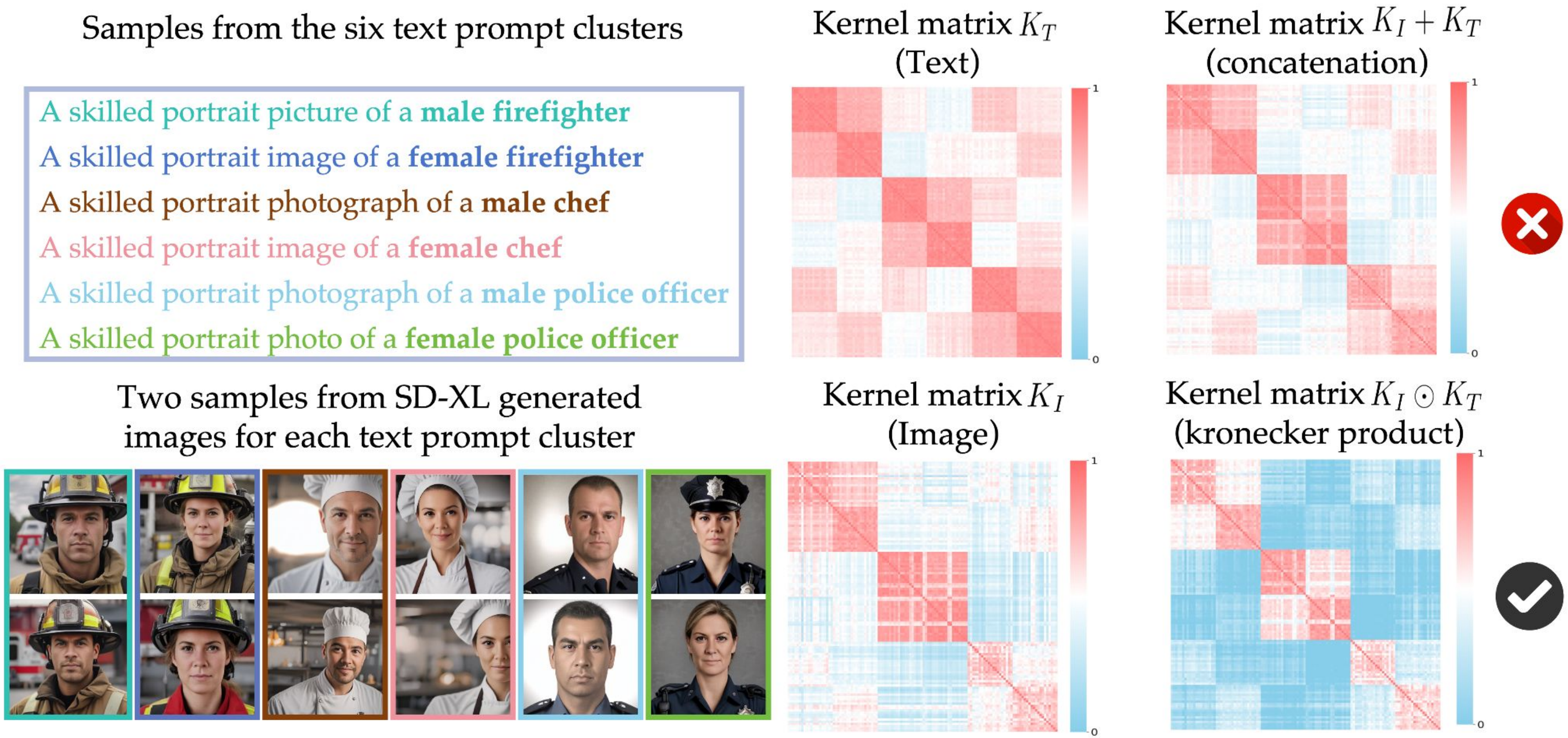
- Different embeddings capture distinct and complementary features.
- CLIP: good at traffic signs
- DINOv2: good at dog breeds



Clustering results and kernel matrix heatmaps for CLIP and DINOv2 on ImageNet dog breeds and GTSRB dataset.

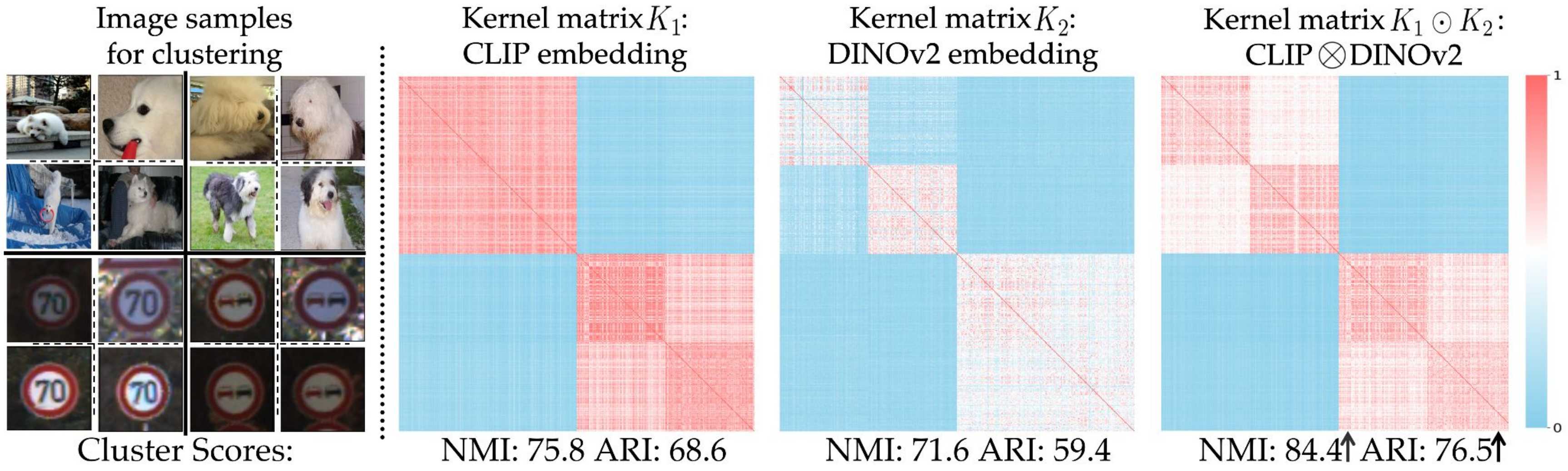
How can we fuse embeddings to combine their strengths?

- **Concatenation** or **kronecker product ?**



Kernel Multiplication = Agreement Rule

- Each embedding defines a kernel $k_\psi(x, y)$ describing similarity.
- Kernel multiplication has below properties:
  - Samples are similar only if all parent embeddings agree.
  - Captures the intersection of similarity structures.



KrossFuse: Kronecker Fusion of Embeddings

- **Fusing Uni-modal Embeddings using their Kronecker Product:**
  - Considering kernel functions $k_1 : \mathcal{Z}_1 \times \mathcal{Z}_1 \to \mathbb{R}$ and $k_2 : \mathcal{Z}_2 \times \mathcal{Z}_2 \to \mathbb{R}$ operating in the embedding spaces, each of the embeddings γ_1 and γ_2 provide a kernel function for inputs $x, y \in \mathcal{X}$:

$$k_{\gamma_1}(x,y) = k_1\big(\gamma_1(x), \gamma_1(y)\big), \quad k_{\gamma_2}(x,y) = k_2\big(\gamma_2(x), \gamma_2(y)\big).$$

  - The product of kernel functions ⇔ The kronecker product of kernel feature maps

$$k_{\gamma_1}(x,y) \cdot k_{\gamma_2}(x,y)$$

$$\phi_{\gamma_1,\gamma_2}(x) = \phi_1\big(\gamma_1(x)\big) \otimes \phi_2\big(\gamma_2(x)\big)$$

KrossFuse: Kronecker Fusion of Embeddings

- **Extending KrossFuse for Kronecker Fusion of Uni-modal and Cross-Modal Embeddings:**
    - Define the following symmetrized cross-modal embedding $\widetilde{\gamma} = (\widetilde{\phi}_{\gamma,X}, \widetilde{\phi}_{\gamma,T})$ to play the role of the uni-modal embedding $\gamma = (\gamma_X)$ that missing text part γ_T in the Kronecker fusion process:

$$\widetilde{\phi}_{\gamma,X}(x) := \frac{1}{\sqrt{2}}\Big[\sqrt{\frac{C}{d}} + \phi(\gamma_X(x))\,,\ \sqrt{\frac{C}{d}} - \phi(\gamma_X(x))\Big]^\top \qquad \widetilde{\phi}_{\gamma,T}(t) := \sqrt{\frac{C}{2d}} \cdot \big[\underbrace{1,\dots,1}_{2d \text{ times}}\big]^\top$$

    - KrossFuse combines the cross-modal embedding $\psi = (\psi_X, \psi_T)$ (e.g. CLIP) and the uni-modal embedding γ (e.g. DINOv2) by taking the Kronecker product of ψ and $\widetilde{\gamma}$ in each modality as:

$$E_X(x) := \phi\big(\Psi_X(x)\big) \otimes \widetilde{\phi}_{\gamma,X}(x),$$
$$E_T(t) := \phi\big(\Psi_T(t)\big) \otimes \widetilde{\phi}_{\gamma,T}(t)$$

RP-KrossFuse: Scalable Embedding Fusion via Random Projection

- Kronecker space is huge (512 × 768 ≈ 393k dims).

- Applies random projection to each cross-modality embedding
  - We generate random matrix a $U_i \in \mathbb{R}^{l \times d_1}$ whose entries are independent random variables with uniform distribution over $[-\sqrt{3}, \sqrt{3}]$ that has unit variance.
  - The RP-KrossFuse embedding of each of inputs x ∈ X and t ∈ T will be

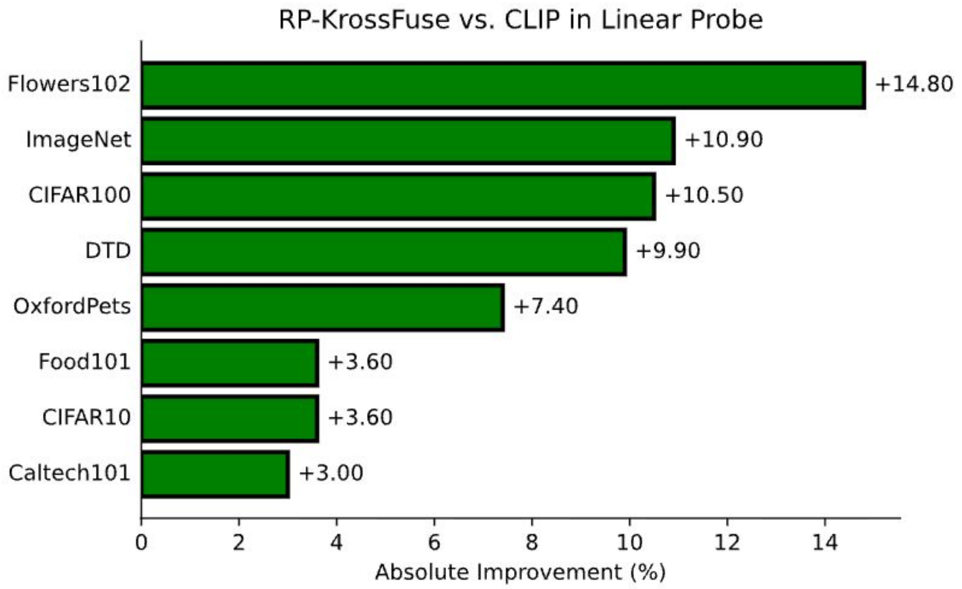
$$\widetilde{\psi}_X(x) = \frac{1}{\sqrt{l}}\big(U_1\psi_{1,X}(x)\big) \odot \big(U_2\psi_{2,X}(x)\big)$$

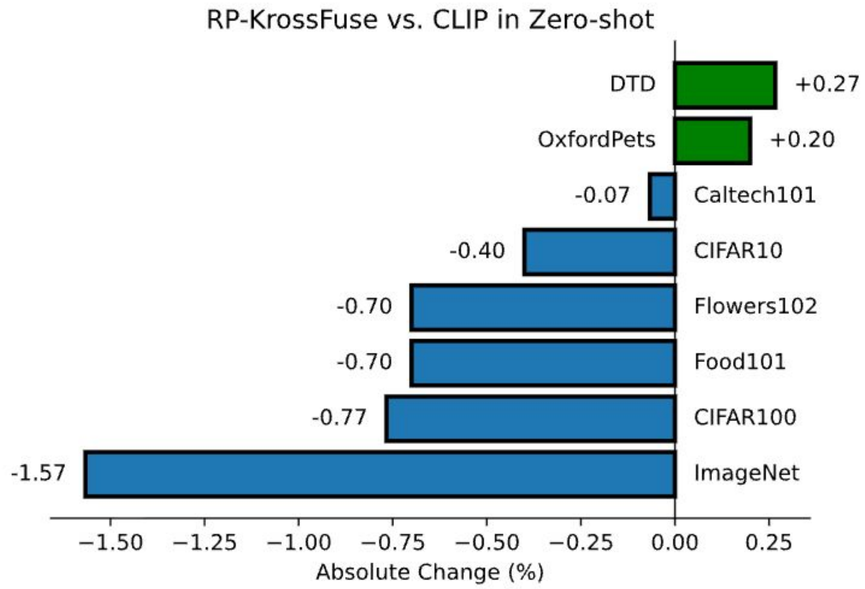
$$\widetilde{\psi}_T(t) = \frac{1}{\sqrt{l}}\big(U_1\psi_{1,T}(t)\big) \odot \big(U_2\psi_{2,T}(t)\big)$$

$\odot$ denotes the element-wise Hadamard product

Comparison of RP-KrossFuse and CLIP

- RP-KrossFuse is able to gain consistent improvements over CLIP in linear probe on all datasets. (Left)
- RP-KrossFuse’s declines in zero shot accuracy are mostly under 1%, which are far outweighed by the gains in linear probe. (Right)





Cross-modal Few-shot Learning

Table 3: Cross-modal few-shot classification results across datasets. "Ours" = RP-KrossFuse (proj. dim. 3000); "I"/"T" denote image/text domains.

Shots	Method	Caltech [16]	Food [9]	DTD [11]	Aircraft [46]	ImageNet [13]	MSCOCO [42]	Average
1	CLIP [60] (I)	70.9	37.8	35.4	14.6	24.3	8.7	32.0
	CLIP [60] (I+T)	78.9	58.7	44.9	17.8	33.8	31.6	44.3
	DINOv2 [53] (I)	84.3	57.9	47.2	15.4	54.0	16.4	45.8
	Ours (I)	84.6	55.7	48.3	19.4	51.8	21.5	46.9
	Ours (I+T)	**86.0**	**64.6**	**51.7**	**20.3**	**54.9**	**43.5**	**53.5**
2	CLIP [60] (I)	78.9	47.8	44.2	18.2	30.2	11.2	38.4
	CLIP [60] (I+T)	82.7	60.7	47.3	19.8	36.0	47.2	49.0
	DINOv2 [53] (I)	88.3	63.4	57.3	17.3	61.9	23.1	51.9
	Ours (I)	89.2	63.6	57.3	23.6	60.9	36.8	55.2
	Ours (I+T)	**90.1**	**68.0**	**59.5**	**24.8**	**62.1**	**51.5**	**59.3**
4	CLIP [60] (I)	83.3	57.7	51.9	20.6	36.8	23.9	45.7
	CLIP [60] (I+T)	84.6	64.8	52.0	21.1	42.4	57.5	53.7
	DINOv2 [53] (I)	90.4	69.8	64.0	20.9	67.0	38.5	58.4
	Ours (I)	90.8	71.8	64.4	28.1	66.6	52.8	62.4
	Ours (I+T)	**91.1**	**73.8**	**65.0**	**28.2**	**67.2**	**58.1**	**63.9**
8	CLIP [60] (I)	84.5	65.5	53.7	24.2	42.1	44.9	52.5
	CLIP [60] (I+T)	85.8	68.7	54.6	24.6	45.2	61.2	56.7
	DINOv2 [53] (I)	91.4	73.0	69.2	24.5	70.4	53.6	63.7
	Ours (I)	92.0	75.9	69.2	**31.7**	70.4	55.1	65.7
	Ours (I+T)	**92.2**	**76.9**	**69.4**	**31.7**	**70.7**	**61.9**	**67.1**

Thank You!

Questions: yqwu24@cse.cuhk.edu.hk