Geodesic Multi-Modal Mixup for Robust Fine-Tuning

Changdae Oh[1*] Junhyuk So[2*] Hoyoon Byun[1] YongTaek Lim[1]
Minchul Shin[3] Jong-June Jeon[1] Kyungwoo Song[4]

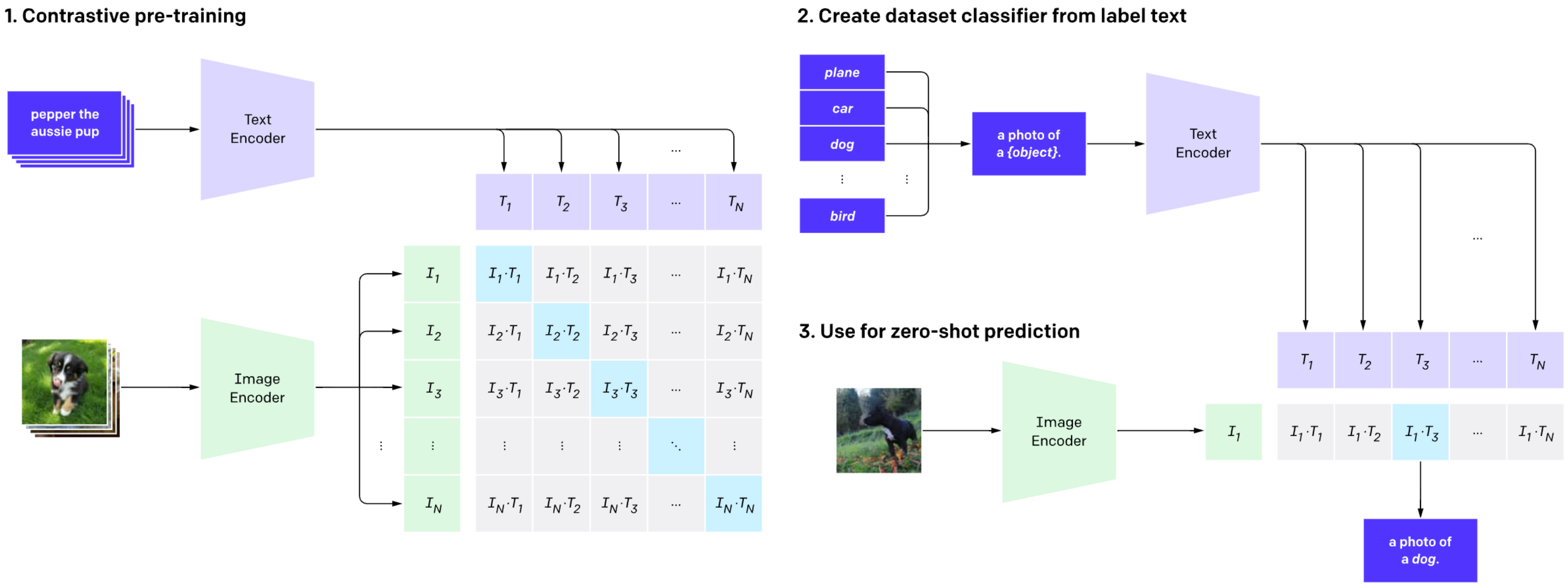
[1] University of Seoul
[2] POSTECH
[3] KAIST
[4] Yonsei University

Great Hall & Hall B1+B2 #715,
Thu 14 Dec 11:45 am – 1:45 pm

November 13, 2023

Contrastive Representation Learning

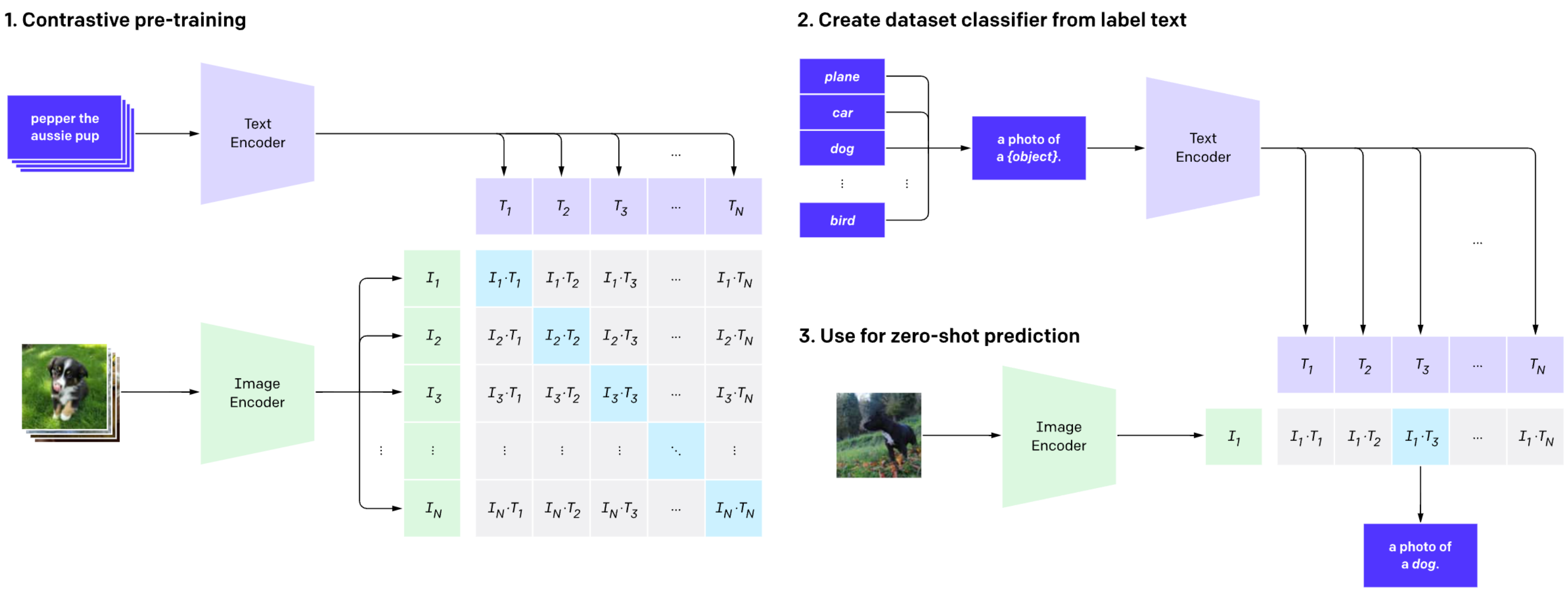
- Multi-modal Contrastive Learning
  - Contrastive Language-Image Pre-training (CLIP) popularizes the large-scale vision-language pre-training commonly equipping contrastive loss as a part of learning objective



Source: CLIP: Connecting text and images, OpenAI 2021
Learning Transferable Visual Models From Natural Language Supervision, Radford et al. ICML 2021

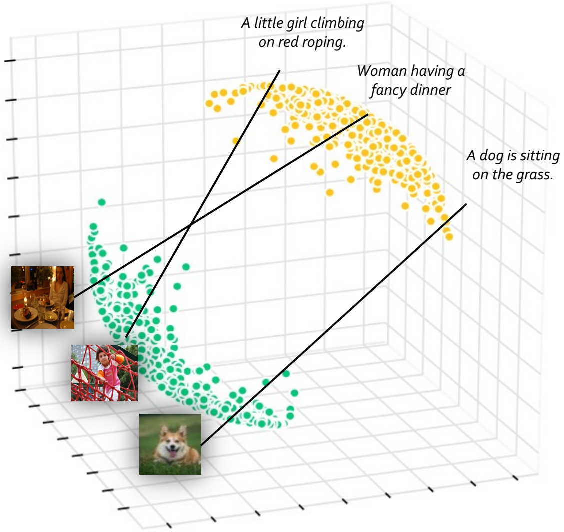
Contrastive Representation Learning

- Multi-modal Contrastive Learning
  - Contrastive Language-Image Pre-training (CLIP) popularizes the large-scale vision-language pre-training commonly equipping contrastive loss as a part of learning objective
  - Language (caption) as an alternative view of a corresponding image, and vice versa
    - ✓ align paired embeddings from two different modalities into single joint representation space



Embedding Space Analysis

- Counterintuitive observation: pre-trained CLIP has separated embedding clusters

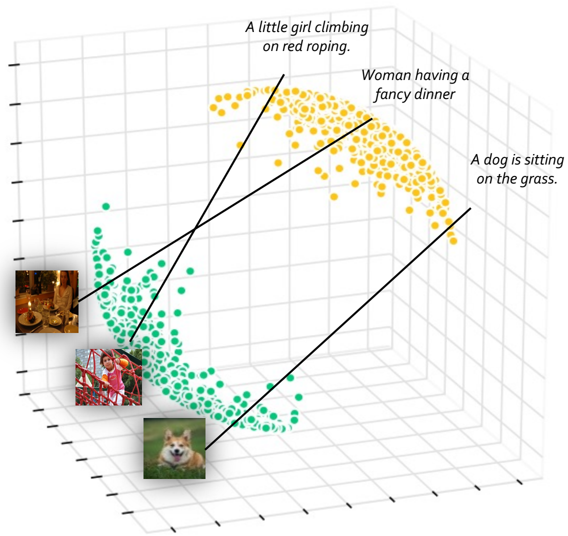


CLIP embedding visualization (DOSNES) on image-caption dataset (Flickr30k)

Source: Doubly stochastic neighbor embedding on spheres, Lu et al. Pattern Recognition Letters 2019

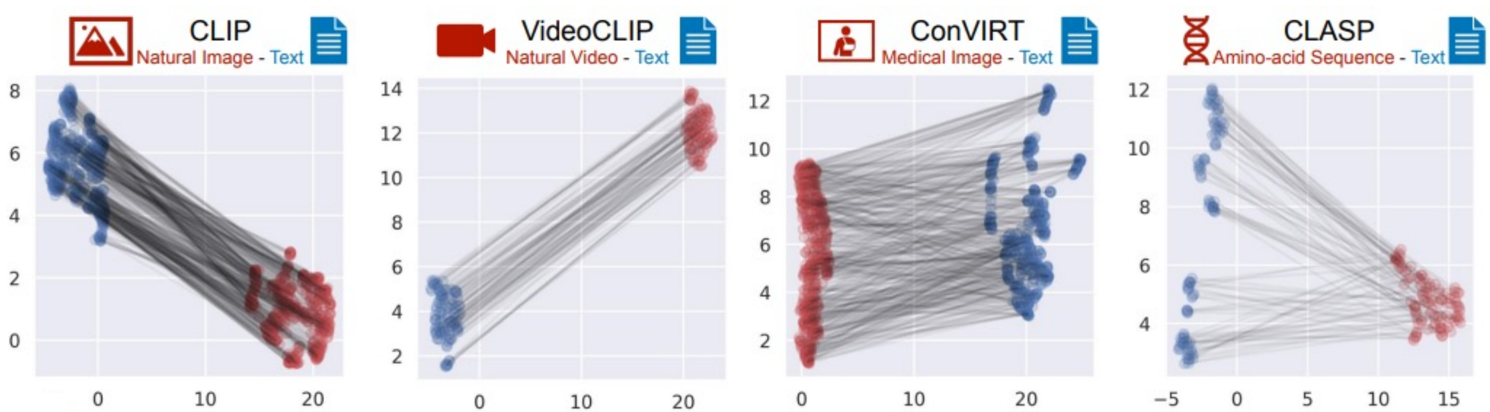
Embedding Space Analysis

- Counterintuitive observation: pre-trained CLIP has separated embedding clusters



CLIP embedding visualization (DOSNES) on image-caption dataset (Flickr30k)

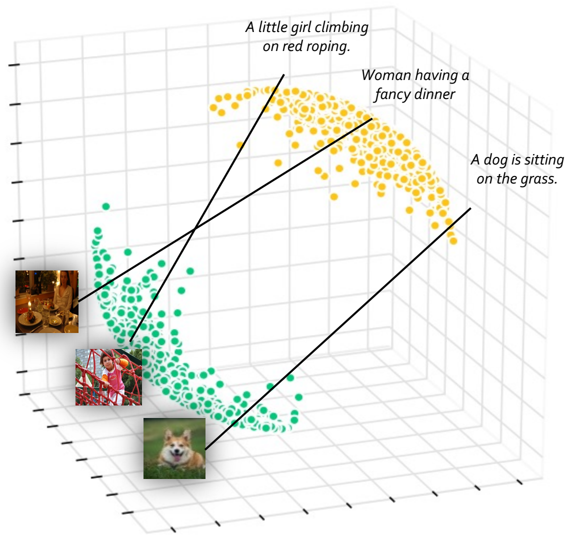
- Concurrent work (in terms of ArXiv preprint) made similar findings: Modality Gap



Source: Doubly stochastic neighbor embedding on spheres, Lu et al. Pattern Recognition Letters 2019
Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning, Liang et al. NeurIPS 2022

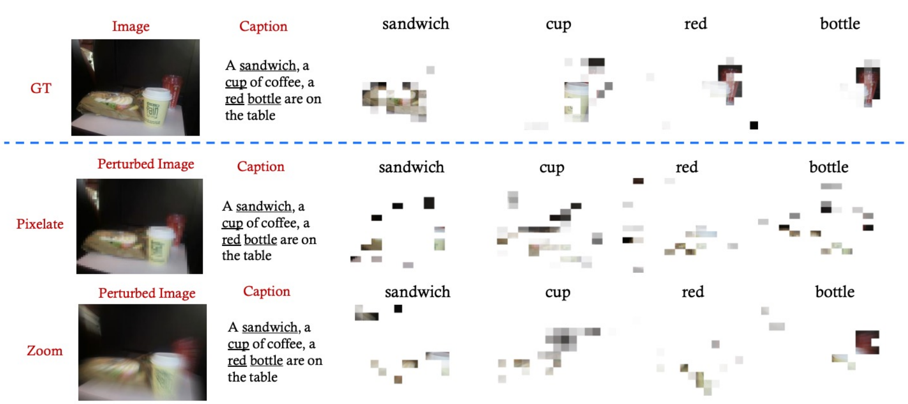
Embedding Space Analysis

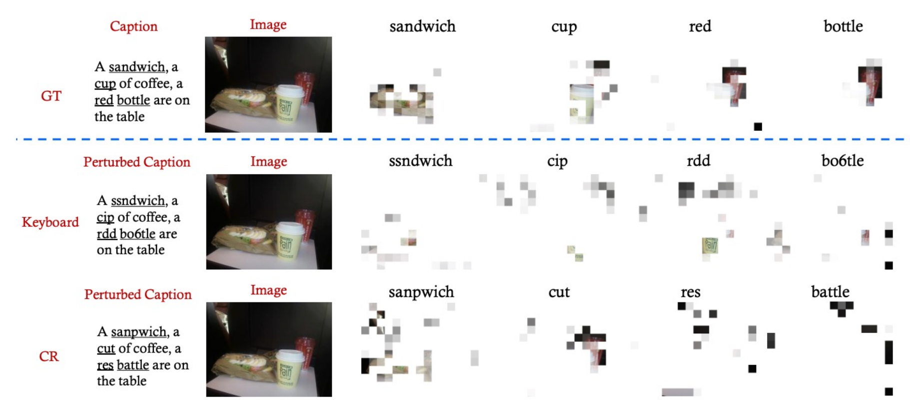
- Counterintuitive observation: pre-trained CLIP has separated embedding clusters



CLIP embedding visualization (DOSNES) on image-caption dataset (Flickr30k)

- This may be vulnerable to unexpected perturbations or out-of-distribution samples





Source: Are Multimodal Models Robust to Image and Text Perturbations?, arXiv 2022

Embedding Space Analysis

- ***Uniformity-Alignment*** (Wang & Isola 2020)
  - quantitative measurement of representation quality

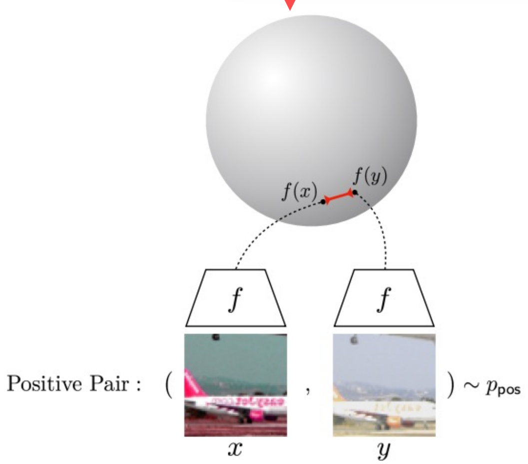
Source: Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, Wang and Isola ICML 2020

Embedding Space Analysis

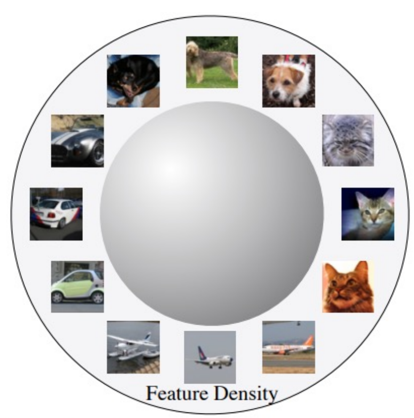
- ***Uniformity-Alignment*** (Wang & Isola 2020)
  - quantitative measurement of representation quality
  - Contrastive loss asymptotically maximizes uniformity and alignment

Theorem 1 (Asymptotics of $\mathcal{L}_{\text{contrastive}}$). *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M \to \infty} \mathcal{L}_{\text{contrastive}}(f; \tau, M) - \log M = - \frac{1}{\tau} \mathop{\mathbb{E}}_{(x,y) \sim p_{\text{pos}}} \left[f(x)^{\mathsf{T}} f(y) \right] + \mathop{\mathbb{E}}_{x \sim p_{\text{data}}} \left[\log \mathop{\mathbb{E}}_{x^- \sim p_{\text{data}}} \left[e^{f(x^-)^{\mathsf{T}} f(x)/\tau} \right] \right] \quad (2)$$



Alignment: Similar samples have similar features.



Uniformity: Preserve maximal information.

Source: Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, Wang and Isola ICML 2020

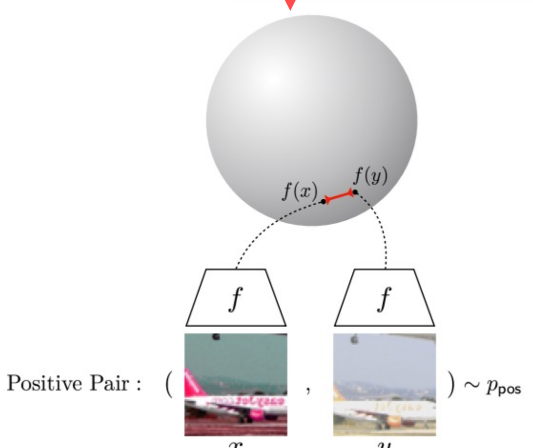
Embedding Space Analysis

- ***Uniformity-Alignment*** (Wang & Isola 2020)
  - quantitative measurement of representation quality
  - Contrastive loss asymptotically maximizes uniformity and alignment

Theorem 1 (Asymptotics of $\mathcal{L}_{\text{contrastive}}$). *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

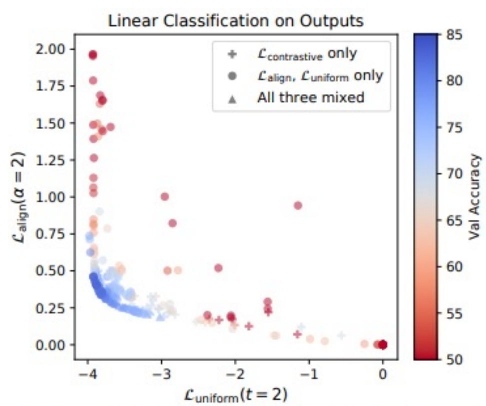
$$\lim_{M\to\infty} \mathcal{L}_{\text{contrastive}}(f;\tau,M) - \log M = - \frac{1}{\tau} \mathop{\mathbb{E}}_{(x,y)\sim p_{\text{pos}}} \left[f(x)^\intercal f(y) \right] + \mathop{\mathbb{E}}_{x\sim p_{\text{data}}} \left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}} \left[e^{f(x^-)^\intercal f(x)/\tau} \right] \right] \quad (2)$$

Unif-Align have strong correlation with downstream task performances



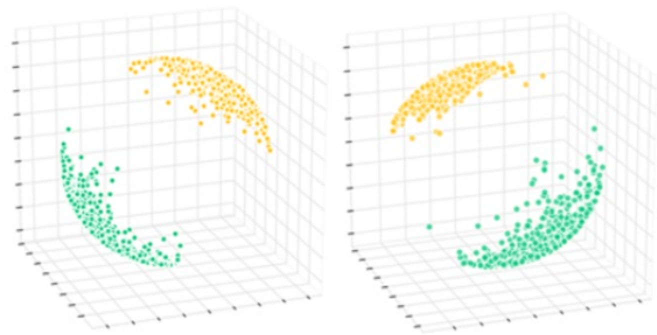
Alignment: Similar samples have similar features.

Uniformity: Preserve maximal information.

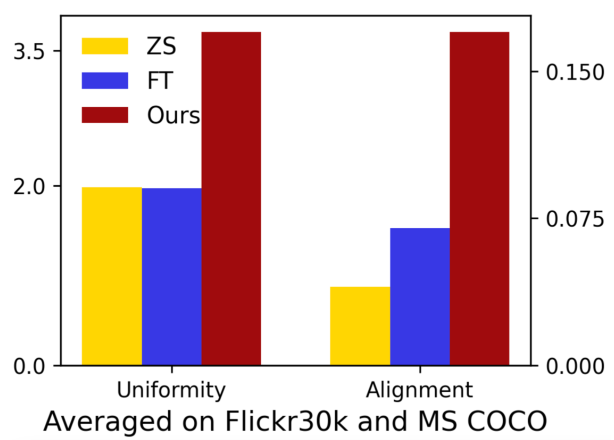


Embedding Space Analysis

- CLIP has limited uniformity-alignment and retains its bipartite embedding structure whether being fine-tuned or not!
- This may constrict the transferability and robustness of the representation



pre-trained CLIP fine-tuned CLIP



Source:

Understanding the Fine-Tuning of CLIP

- Why does CLIP preserve its bipartite structure (so called modality gap) and fail to increase uniformity-alignment during fine-tuning?

Understanding the Fine-Tuning of CLIP

- Why does CLIP preserve its bipartite structure (so called modality gap) and poor uniformity-alignment during fine-tuning?

- Our arguments:
  - By assuming vanishing temperature τ, $\mathcal{L}_{\text{CLIP}}$ converges to triplet loss with zero-margin same as the negative relative **alignment**

Alignment $:= -\mathbb{E}_{(x_i,y_i)}\left[\|f(x_i)-g(y_i)\|_2^2 - \min_{k\neq i}\|f(x_i)-g(y_k)\|_2^2\right]$

$$
\begin{aligned}
C(I,T;\theta) &= \lim_{\tau\to 0^+}\frac{1}{M}\sum_{i=1}^{M} -\log\frac{\exp((I_i\cdot T_i)/\tau)}{\sum_{j=1}^{M}\exp((I_i\cdot T_j)/\tau)} \\
&= \lim_{\tau\to 0^+}\frac{1}{M}\sum_{i=1}^{M} -\exp((I_i\cdot T_i)/\tau) + \log\left[\exp((I_i\cdot T_i)/\tau) + \sum_{j\neq i}\exp((I_i\cdot T_j)/\tau)\right] \\
&= \lim_{\tau\to 0^+}\frac{1}{M}\sum_{i=1}^{M} \log\left[1 + \sum_{j\neq i}\exp((I_i\cdot T_j)-(I_i\cdot T_i)/\tau)\right] \\
&= \lim_{\tau\to 0^+}\frac{1}{M}\sum_{i=1}^{M} \log\left[1 + \sum_{j\in\mathcal{J}(i,I,T)}\exp((I_i\cdot T_j)-(I_i\cdot T_i)/\tau)\right] \\
&\qquad\qquad (\text{where } \mathcal{J}(i,I,T) := \{j \mid (I_i\cdot T_j) > (I_i\cdot T_i)\}) \\
&= \lim_{\tau\to 0^+}\frac{1}{M}\sum_{i=1}^{M}\frac{1}{\tau}\max\left[\max_j(I_i\cdot T_j)-(I_i\cdot T_i), 0\right]
\end{aligned}
$$

Source:

Understanding the Fine-Tuning of CLIP

- Why does CLIP preserve its bipartite structure (so called modality gap) and poor uniformity-alignment during fine-tuning?

- Our arguments:
  - By assuming vanishing temperature τ, $\mathcal{L}_{\text{CLIP}}$ converges to triplet loss with zero-margin same as the negative relative **alignment**
  - Lack of hard negative samples to encourage alignment further

$$\text{Alignment} := -\mathbb{E}_{(x_i, y_i)}\left[\|f(x_i) - g(y_i)\|_2^2 - \min_{k \neq i} \|f(x_i) - g(y_k)\|_2^2\right]$$

$$C(I, T; \theta) = \lim_{\tau \to 0^+} \frac{1}{M} \sum_{i=1}^{M} -\log \frac{\exp((I_i \cdot T_i)/\tau)}{\sum_{j=1}^{M} \exp((I_i \cdot T_j)/\tau)}$$

$$= \lim_{\tau \to 0^+} \frac{1}{M} \sum_{i=1}^{M} -\exp((I_i \cdot T_i)/\tau) + \log\left[\exp((I_i \cdot T_i)/\tau) + \sum_{j \neq i} \exp((I_i \cdot T_j)/\tau)\right]$$

$$= \lim_{\tau \to 0^+} \frac{1}{M} \sum_{i=1}^{M} \log\left[1 + \sum_{j \neq i} \exp((I_i \cdot T_j) - (I_i \cdot T_i)/\tau)\right]$$

$$= \lim_{\tau \to 0^+} \frac{1}{M} \sum_{i=1}^{M} \log\left[1 + \sum_{j \in \mathcal{J}(i, I, T)} \exp((I_i \cdot T_j) - (I_i \cdot T_i)/\tau\right]$$

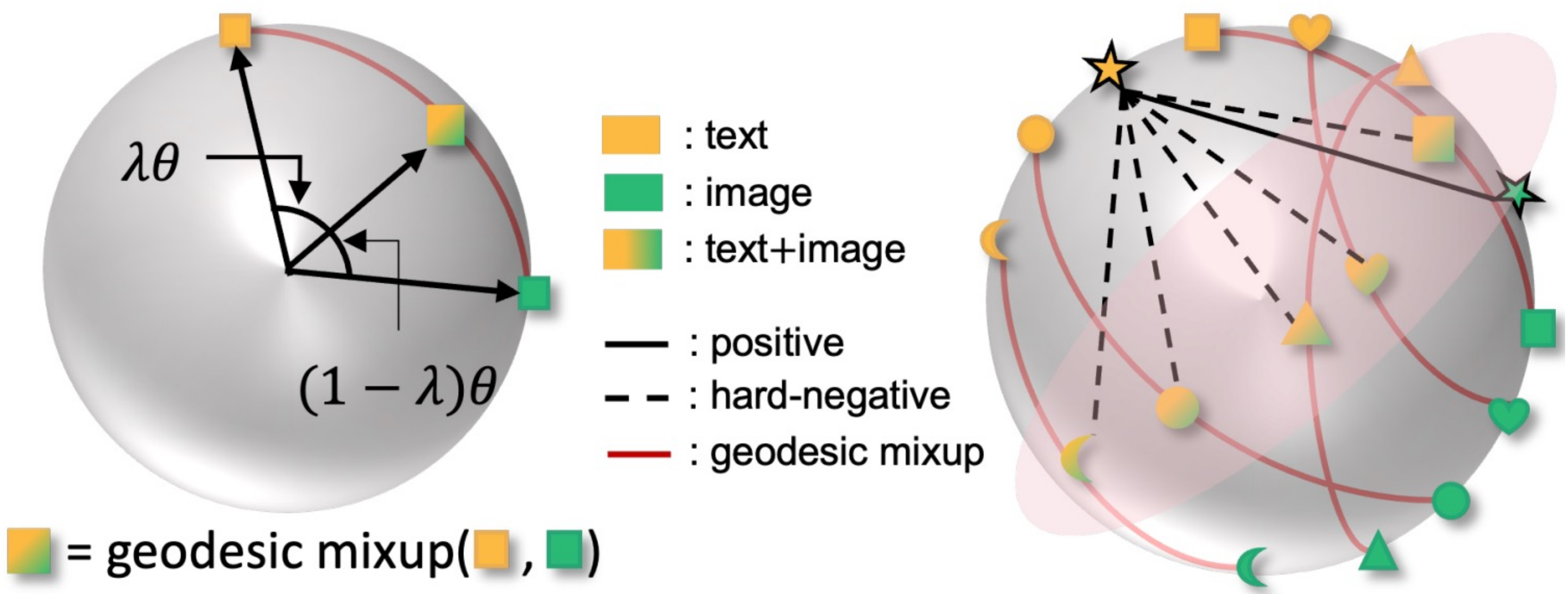
$(\text{where } \mathcal{J}(i, I, T) : \ldots \cdot T_j) > (I_i \cdot T_i)\})$

$$= \lim_{\tau \to 0^+} \frac{1}{M} \sum_{i=1}^{M} \frac{1}{\tau} \max\left[\max_j (I_i \cdot T_j) - (I_i \cdot T_i), 0\right]$$

There is no incentive to enforce the alignment without hard negatives

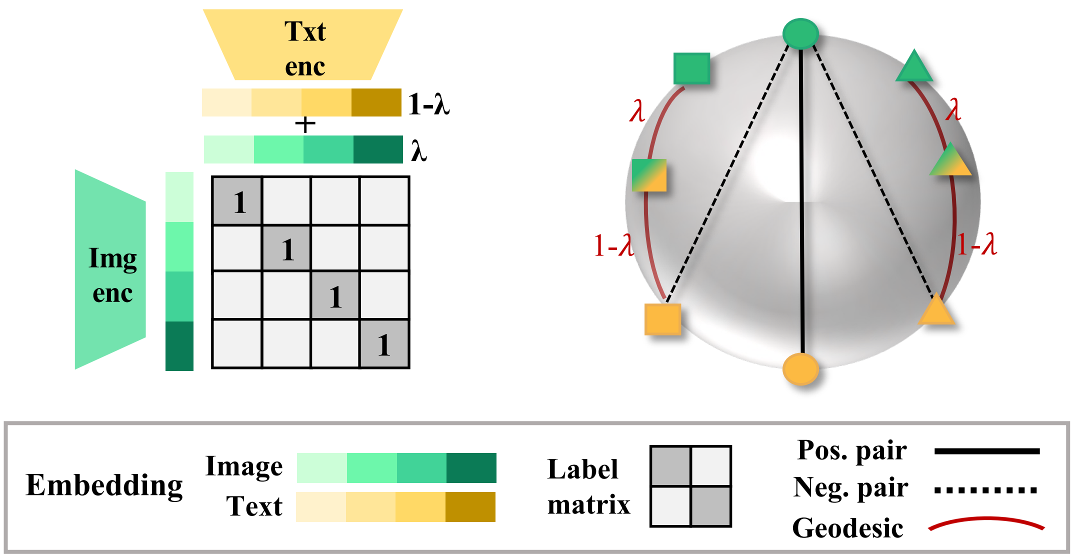
Robust Fine-Tuning with Geodesic Multi-Modal Mixup

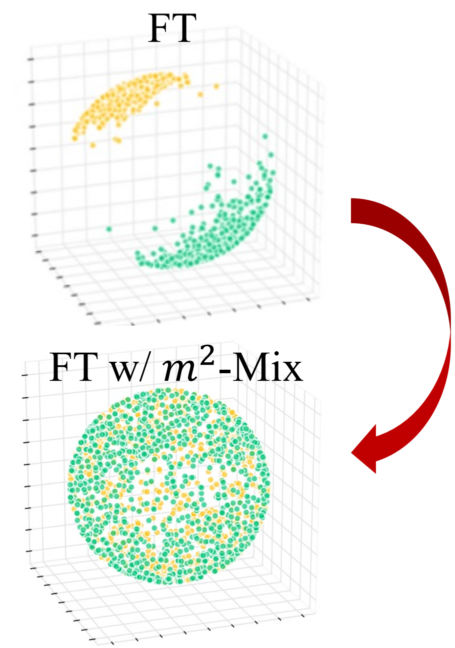
- **Geodesic Multi-Modal Mixup**, m^2-Mix
  - Mixes the heterogeneous embeddings from two modalities (i.e., image and text)
  - Use that mixtures as virtual hard negatives for the contrastive loss



Robust Fine-Tuning with Geodesic Multi-Modal Mixup

- **Geodesic Multi-Modal Mixup**, m^2-Mix
  - Mixes the heterogeneous embeddings from two modalities (i.e., image and text)
  - Use that mixtures as virtual hard negatives for the contrastive loss





$$m_\lambda(\vec{a}, \vec{b}) = \vec{a}\frac{\sin(\lambda\theta)}{\sin(\theta)} + \vec{b}\frac{\sin((1-\lambda)\theta)}{\sin(\theta)}, \quad \text{where } \theta = \cos^{-1}(\vec{a}\cdot\vec{b}) \text{ and } \lambda \sim \text{Beta}(\alpha, \alpha)$$

ensure that the mixture embeddings lie on the hypersphere

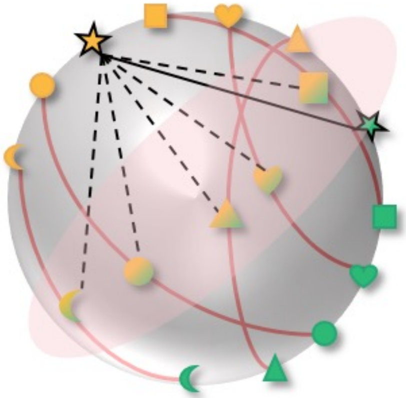
Source:

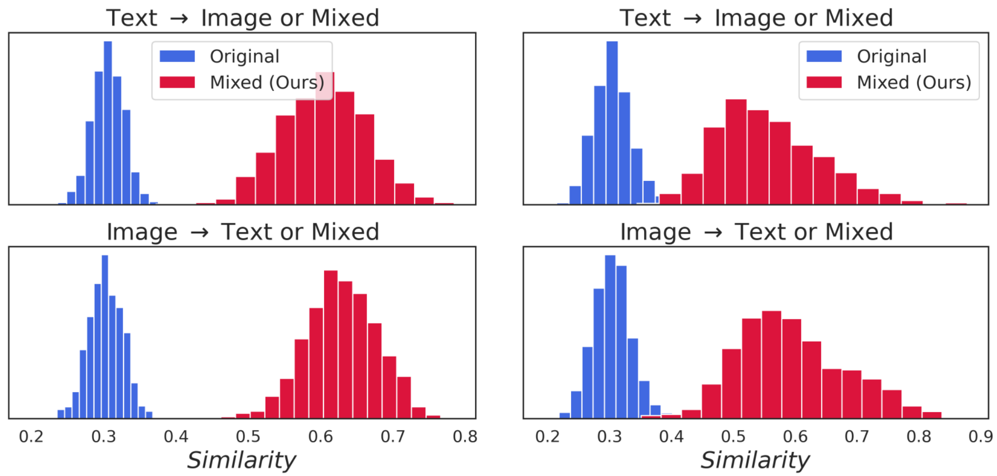
Understanding the Fine-Tuning of CLIP with m^2-Mix

- **Hard negative generation with m^2-Mix**

Theorem 4.1 (Hardness of m^2-Mixed samples). *Let's assume that two random variables x_1 and x_2 follow the $M_d(\mu_1, \kappa)$ and $M_d(\mu_2, \kappa)$, von Mises–Fisher distribution with mean direction μ_1, μ_2 and concentration parameter κ in $\mathbb{R}^d$, respectively. Let $\widetilde{x} = x_1 + x_2$ and d $= 2$. Then, $D_{KL}(p(x_1)||p(\widetilde{x})) \leq D_{KL}(p(x_1)||p(x_2))$ for sufficiently large κ.*

- Corresponds to our intuition (supported by empirical results)





Initial epoch Last epoch

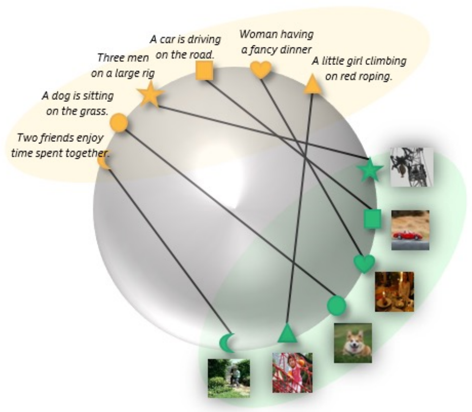
Understanding the Fine-Tuning of CLIP with m^2-Mix

- **Contrastive Loss with m^2-Mix converges to negative uniformity**, so complements the uniformity which is lack in $\mathcal{L}_{CLIP}$

$$
\begin{aligned}
C_{m^2\text{-Mix}}(I,T;\theta) &= \lim_{\tau\to 0^+} \frac{1}{M}\sum_{i=1}^{M} -\log \frac{\exp((I_i\cdot T_i)/\tau)}{\sum_{j=1}^{M}\exp\left((I_i\cdot mix(I_i,T_j)/\tau\right)} \qquad (3)\\
&= \lim_{\tau\to 0^+} \frac{1}{M}\sum_{i=1}^{M} -((I_i\cdot T_i)/\tau) + \log\left[\exp\left((I_i\cdot T_i)/\tau\right) + \sum_{j\neq i}\exp\left((I_i\cdot mix(I_i,T_j))/\tau\right)\right]\\
&= \lim_{\tau\to 0^+} \frac{1}{M}\sum_{i=1}^{M} \log\left[1+\sum_{j\neq i}\exp\left((I_i\cdot mix(I_i,T_j) - (I_i\cdot T_i))/\tau\right)\right]\\
&= \lim_{\tau\to 0^+} \frac{1}{M}\sum_{i=1}^{M} \log\left[1+\sum_{j\neq i}\exp\left(I_i\cdot mix(I_i,T_j)/\tau\right)\right]\\
&\qquad\qquad \text{(by assuming } (I_i\cdot mix(I_i,T_j)) > (I_i\cdot T_i) \text{ for all } j\neq i)\\
&= \lim_{\tau\to 0^+} \frac{1}{M}\sum_{i=1}^{M} \log\sum_{j\neq i}\exp\left(I_i\cdot mix(I_i,T_j)/\tau\right)\\
&\simeq \lim_{\tau\to 0^+} -\text{Uniformity}(I, mix(I_i,T_j);\theta) \qquad \text{(for sufficiently large } M)
\end{aligned}
$$

Understanding the Fine-Tuning of CLIP with m^2-Mix

- By equipping our $\mathcal{L}_{m^2\text{–Mix}}$ with $\mathcal{L}_{CLIP}$, we can expect:

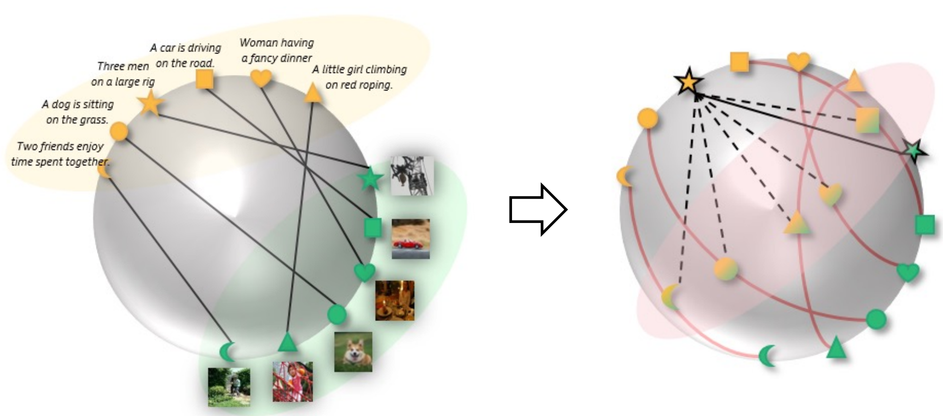


Source:

Understanding the Fine-Tuning of CLIP with m^2-Mix

- By equipping our $\mathcal{L}_{m^2\text{–Mix}}$ with $\mathcal{L}_{CLIP}$, we can expect:
  - **Enhanced alignment** through hard-negative-based contrastive learning

Theorem 4.1 (Hardness of m^2-Mixed samples). *Let's assume that two random variables x_1 and x_2 follow the $M_d(\mu_1, \kappa)$ and $M_d(\mu_2, \kappa)$, von Mises–Fisher distribution with mean direction μ_1, μ_2 and concentration parameter κ in $\mathbb{R}^d$, respectively. Let $\widetilde{x} = x_1 + x_2$ and $d = 2$. Then, $D_{KL}(p(x_1)||p(\widetilde{x})) \leq D_{KL}(p(x_1)||p(x_2))$ for sufficiently large κ.*



Source:

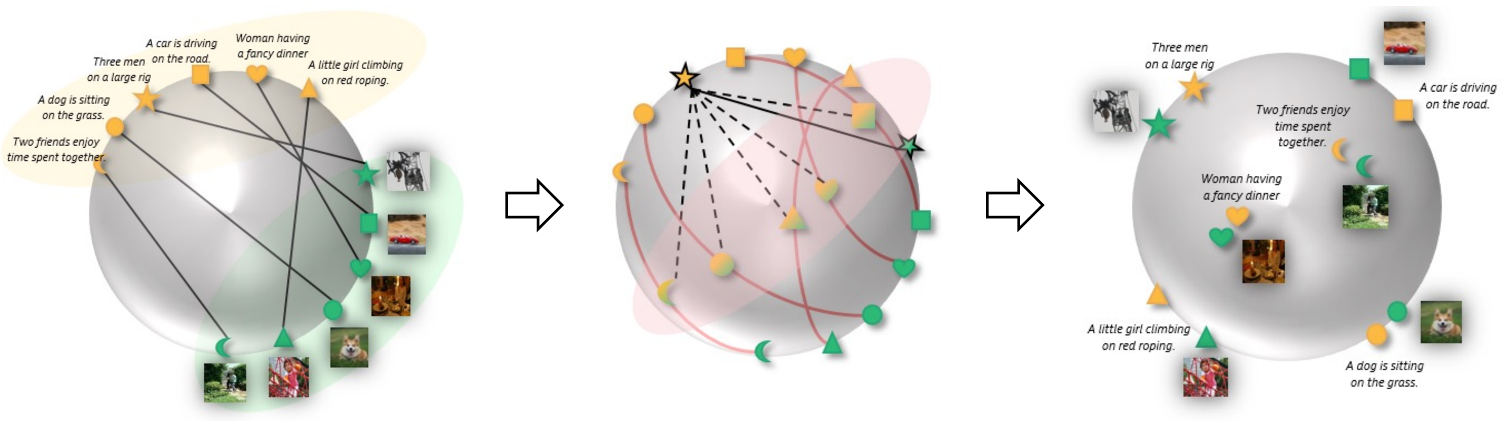
Understanding the Fine-Tuning of CLIP with m^2-Mix

- By equipping our $\mathcal{L}_{m^2-\mathrm{Mix}}$ with $\mathcal{L}_{CLIP}$, we can expect:
  - **Enhanced alignment** through hard-negative-based contrastive learning

Theorem 4.1 (Hardness of m^2-Mixed samples). *Let's assume that two random variables x_1 and x_2 follow the $M_d(\mu_1, \kappa)$ and $M_d(\mu_2, \kappa)$, von Mises–Fisher distribution with mean direction μ_1, μ_2 and concentration parameter κ in $\mathbb{R}^d$, respectively. Let $\tilde{x} = x_1 + x_2$ and $d = 2$. Then, $D_{KL}(p(x_1)||p(\tilde{x})) \leq D_{KL}(p(x_1)||p(x_2))$ for sufficiently large κ.*

  - Approximately **maximizing uniformity and alignment, simultaneously**

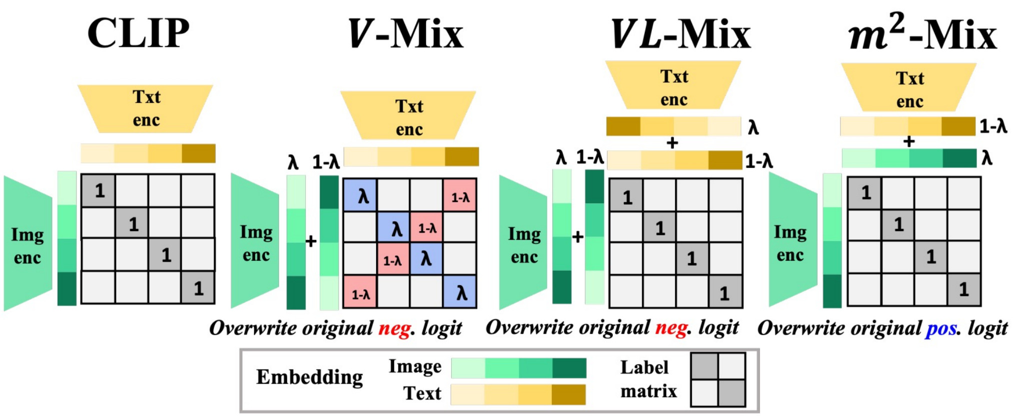
Proposition 4.2 (Limiting behavior of $\mathcal{L}_{\mathrm{CLIP}}$ with m^2-Mix). *For sufficiently large M, as the temperature of contrastive loss $\tau \to 0^+$, the $\mathcal{L}_{CLIP}$ and $\mathcal{L}_{m^2\text{-}Mix}$ converges to the triplet loss with zero-margin (i.e., corresponding to negative Alignment) and negative Uniformity, respectively. That is: $\lim_{\tau \to 0^+} \mathcal{L}_{CLIP} + \mathcal{L}_{m^2\text{-}Mix} \simeq -(Alignment + Uniformity)$*

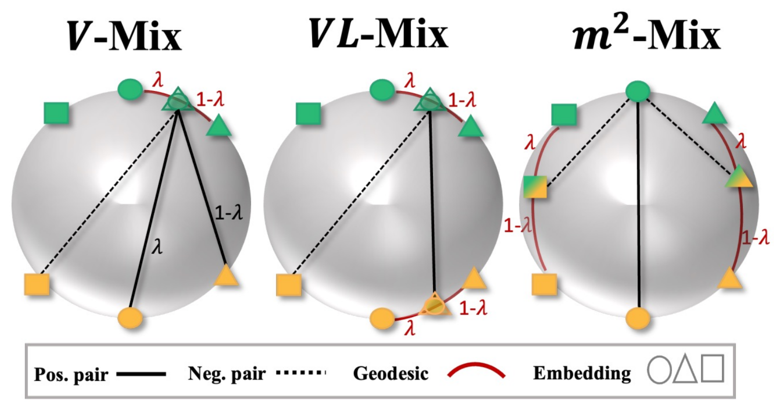


Source:

Uni-Modal Mixups for Multi-Modal Contrastive Learning

- Representation learning can be further robustified with uni-modal Mixups





- Complete learning objective, m^3-Mix (multiple multi-modal Mixup)

$$\mathcal{L}_{m^3\text{-Mix}} = \mathcal{L}_{CLIP} + \mathcal{L}_{m^2\text{-Mix}} + \mathcal{L}_{uni\text{-Mix}} + \mathcal{L}_{VL\text{-Mix}}$$

Source:

Key Results

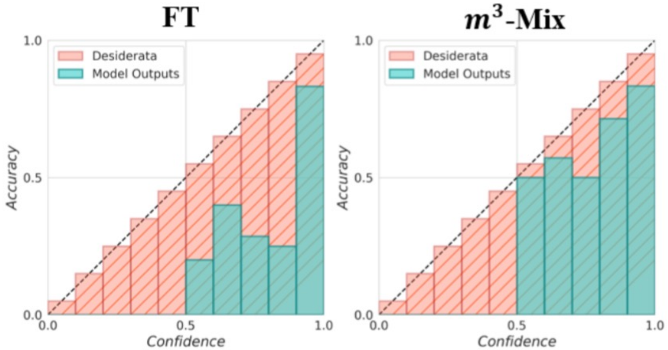
- Cross-modal retrieval (left: CLIP, right: BERT-RN50)

	Flickr30k				MS COCO			
	i→t		t→i		i→t		t→i	
	R1	R5	R1	R5	R1	R5	R1	R5
ZS	71.1	90.4	68.5	88.9	31.9	56.9	28.5	53.1
ES [27]	71.8	90.0	68.5	88.9	31.9	56.9	28.7	53.0
FT	81.2	95.4	80.7	95.8	36.7	63.6	36.9	63.9
FT ($\tau = 0.05$)	82.4	95.1	82.1	95.7	40.2	68.2	41.6	**69.9**
FT ($\tau = 0.10$)	75.7	93.9	78.0	92.9	34.2	62.7	36.7	64.2
i-Mix [50]	72.3	91.7	69.0	91.1	34.0	63.0	34.6	62.2
Un-Mix [51]	78.5	95.4	74.1	91.8	38.8	66.2	33.4	61.0
m^3-Mix	82.3	**95.9**	82.7	**96.0**	**41.0**	**68.3**	39.9	67.9
m^3-Mix ($\tau = 0.05$)	**82.7**	95.7	**82.8**	95.5	40.4	67.9	**42.0**	68.8

	Flickr30k			
	i → t		t → i	
	R1	R5	R1	R5
ZS	0.1	0.4	0.1	0.2
ES [27]	0.1	0.5	0.2	0.2
FT	28.7	61.7	26.7	59.4
FT ($\tau = 0.05$)	31.5	64.2	29.2	61.4
FT ($\tau = 0.10$)	30.0	62.7	30.1	60.6
i-Mix [50]	27.6	60.3	27.1	60.7
Un-Mix [51]	31.5	64.3	29.2	61.2
m^3-Mix	31.9	62.6	30.3	61.0
m^3-Mix ($\tau = 0.05$)	**32.5**	**64.7**	**30.4**	**63.4**

- Expected calibration error on retrieval recall

Metric	Task	ZS	FT	m^3-Mix
ECE (↓)	i → t	1.90	2.26	**1.54**
	t → i	1.88	2.00	**1.58**



Robust representation with better uniformity-alignment contributes to enhance calibration as well as improve recall

Source:

Key Results

- Few-shot adaptation (left) and zero-shot transfer (right)

Method	Pets	SVHN	CLEVR	Avg.
ZS	87.49	13.63	20.70	40.61
FT	89.37	45.00	53.49	62.62
FT w/ V-Mix	89.45	44.61	53.93	62.66
FT w/ L-Mix	89.43	48.42	53.91	63.92
FT w/ VL-Mix	89.56	45.22	53.75	62.84
FT w/ m^2-Mix	90.05	46.24	53.60	63.29
m^3-Mix	90.16	54.84	53.85	66.28
m^3-Mix ($\tau = 0.05$)	90.49	60.90	53.95	68.45
WiSE-FT [10]	91.80	35.04	41.93	56.25
WiSE-FT w/ m^3-Mix	**92.51**	58.55	47.11	66.06
LP-FT [11]	89.92	44.91	53.62	62.82
LP-FT w/ m^3-Mix	91.03	**64.24**	**55.20**	**70.16**
MaPLe [64]	90.87	47.62	43.05	60.51
MaPLe w/ m^3-Mix	91.14	52.72	45.20	63.02

Method	IN	IN-V2	IN-A	IN-R	IN-S	Avg.
ZS	62.06	54.80	29.63	66.02	40.82	50.67
FT	65.44	55.35	20.07	58.16	34.50	46.70
FT w/ V-Mix	66.00	56.19	20.85	60.50	34.97	47.70
FT w/ L-Mix	65.96	55.95	20.57	60.54	35.25	47.65
FT w/ VL-Mix	66.24	56.70	21.36	61.07	35.11	48.10
FT w/ m^2-Mix	67.04	57.39	20.05	59.28	35.31	47.81
m^3-Mix	67.08	57.55	20.80	60.96	35.86	48.45
m^3-Mix ($\tau = 0.05$)	68.40	58.51	22.17	62.28	37.62	49.80
WiSE-FT [10]	69.00	59.66	28.01	64.84	41.05	52.51
WiSE-FT w/ m^3-Mix	**69.65**	**60.71**	29.16	66.75	42.19	**53.69**
LP-FT [11]	68.22	58.40	25.57	63.36	38.04	50.72
LP-FT w/ m^3-Mix	68.62	59.17	25.85	65.14	38.78	51.51
MaPLe [64]	65.59	58.44	32.49	68.13	42.53	53.44
MaPLe w/ m^3-Mix	65.76	58.16	**32.52**	**68.20**	**42.67**	53.46

- Geodesic Mixup vs Linear Mixup

Temperature (τ)	m^3-Mix type linear	m^3-Mix type geodesic
0.01	48.36	**48.45**
0.05	48.48	**49.80**
0.10	45.20	**46.41**

- The proposed Mixups largely boost few-shot adaption and zero-shot transfer performances
- m^3-Mix is a flexible plug-in method that provides complementary benefits to recent fine-tuning methods
- Geodesic Mixup is more favorable to contrastive learning with normalized embedding

Source:

Key Results

- Multi-modal sentiment classification under modality missing

(a) MulT (b) GMC (c) GMC+m^2-Mix

	Test-time Observed Modalities																			
	Full (T+V+A)		T			V			A			T+V			T+A			V+A		
	acc.	unif.	acc.	align.	unif.	acc.	align.	unif.	acc.	align.	unif.	acc.	align.	unif.	acc.	align.	unif.	acc.	align.	unif.
MulT [86]	80.5	0.99	60.0	-	1.03	53.9	-	2.07	52.7	-	0.62	57.8	-	1.27	58.8	-	0.77	54.6	-	1.36
GMC [84]	80.1	3.06	78.5	0.20	3.03	**64.7**	0.17	3.01	66.0	0.09	3.03	77.0	0.07	2.94	77.4	0.08	3.00	67.3	0.05	2.98
GMC+m^2-Mix	**80.5**	**3.18**	**78.9**	**0.23**	**3.17**	64.2	**0.19**	**3.15**	**66.2**	**0.12**	**3.15**	**77.8**	**0.08**	**3.08**	**77.9**	**0.09**	**3.08**	**67.4**	**0.06**	**3.10**

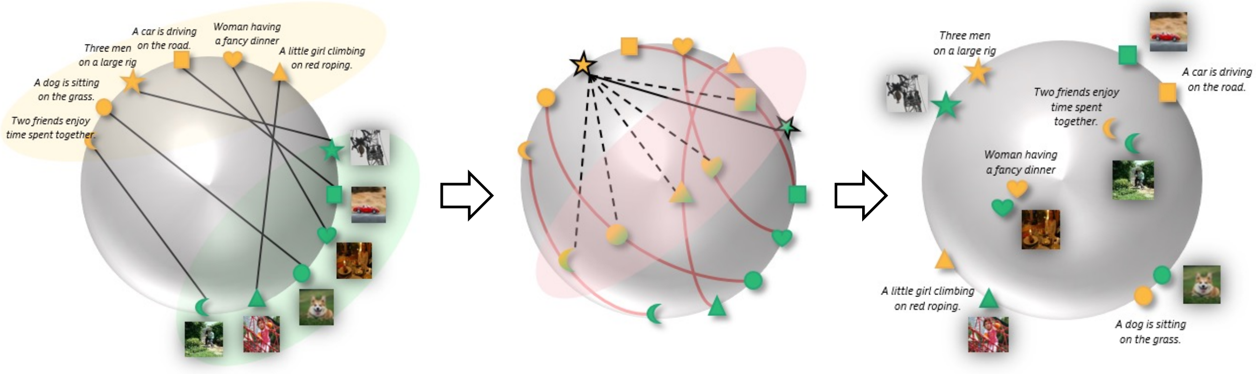
- Image captioning with Contrastive Captioner (CoCa, Yu et al. 2022)

Method	Metrics				
	BLEU@4	METEOR	ROUGE-L	CIDEr	SPICE
ZS	7.2	12.4	26.3	35.2	9.3
Cap	36.0	29.4	57.3	125.1	23.1
CL + Cap	35.7	29.3	57.1	124.9	23.0
CL w/ $\mathcal{L}_{m^2\text{-Mix}}$ + Cap	**36.3**	**29.5**	**57.5**	**125.6**	**23.2**

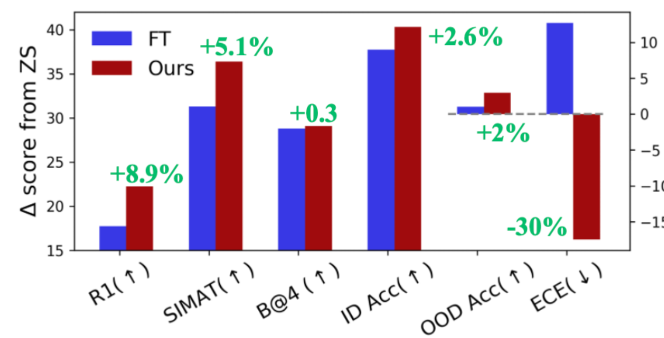
Source: CoCa: Contrastive Captioners are Image-Text Foundation Models, Yu et al. TMLR 2022

thanks!

paper code



Highlights



- Observation
  - CLIP has **Image-versus-text** separated embeddings with limited *uniformity-alignment*
- Problem Define
  - Poor uniformity-alignment **may limit transferability and robustness of learned embedding**
  - Naïve fine-tuning can not mitigate above issue, so how can we address this?
- Our Approach
  - Contrastive Learning with ***Geodesic Multi-Modal Mixup***