Correlation between Alignment-Uniformity and Performance of Dense Contrastive Representations

Jong Hak Moon[1], Wonjae Kim[2], Edward Choi[1]

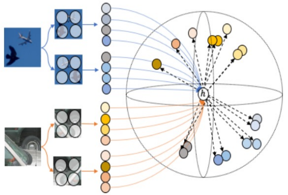
KAIST[1], NAVER AI[2]

paper, code

Motivation & Contribution

- ✓ Unsupervised CL has developed to achieve superior performance by extending the feature representation from instance- to dense-level, but the properties of contrastive representation have not yet been studied.
- ✓ We **analyze** the theoretical ideas of **dense CL** through the lens of **alignment and uniformity on the hypersphere** and introduces a new scalar metric that summarizes the **correlation between alignment-uniformity and downstream performance.**
- ✓ We **discover the core principle** in **constructing a positive pair** of dense features and empirically proved its validity.

Alignment-Uniformity of DenseCL



✓ Quantifying correlation.

Kendall's tau correlation metric:
Min-max norm $(sum\ (L_a, L_u), P_{task})$.

Correlation value varies between -1 and +1, with a value close to 0 indicating a weak correlation.

Negative correlation indicate that pretraining is useful.

✓ *Dense CL with InfoNCE loss.*

$$L^{DenseCont} = -\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} \log \frac{e^{sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)})/\tau}}{\sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \sum_q^{HW} \mathbb{1}_{[k \neq i] \times [\substack{q \neq p \\ k = j}]} e^{sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(k,q)})/\tau}}, \quad sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

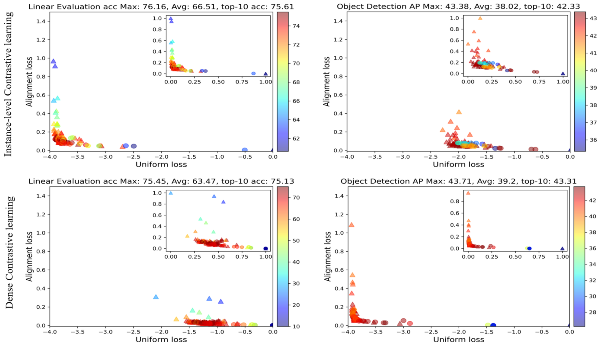
$$\begin{aligned} L^{DenseCont} &= -\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} \Big(sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)})/\tau - \log \sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \sum_{q}^{HW} \mathbb{1}_{[k \neq i] \times [\substack{q \neq p \\ k = j}]} e^{sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(k,q)})/\tau} \Big) \\ &= \underbrace{-\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)})/\tau}_{\text{alignment property}} + \underbrace{\frac{1}{\mathcal{N}} \sum_{i}^{\mathcal{N}} \frac{1}{HW} \sum_{p}^{HW} \log \Big(\sum_{k \overset{\text{i.i.d}}{\sim} 2\mathcal{N}} \sum_{q}^{HW} \mathbb{1}_{[k \neq i] \times [\substack{q \neq p \\ k = j}]} e^{sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(k,q)})/\tau} \Big)}_{\text{distribution to be uniform property}} \end{aligned}$$

$$\boxed{L_a \triangleq -\frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} sim(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)}),} \quad \boxed{L_u \triangleq log \frac{1}{\mathcal{N}} \sum_{i,j \overset{\text{i.i.d}}{\sim} \mathcal{B}} \frac{1}{HW} \sum_{p}^{HW} G(\mathbf{h}_{(i,p)}, \mathbf{h}_{(j,p)})}$$

Experiments & Results

- ✓ How does the alignment-uniformity property of dense CL correlate with the performance of object detection and linear evaluation?
- ✓ How different is the behavior of dense feature representations on single or multi-object datasets?
- ✓ How effective is the index-wise matching strategy in terms of different augmentation techniques?

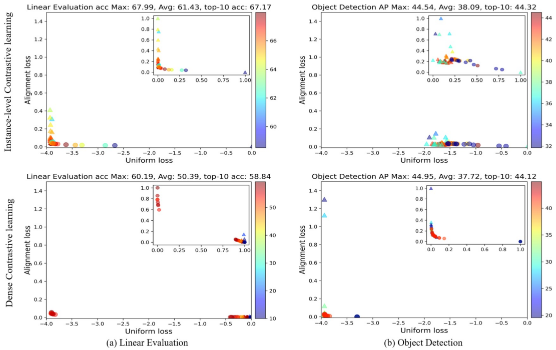
Single-object pretraining and fine-tuning



(a) Linear Evaluation (b) Object Detection

Pretraining	Loss	Instance-level Evaluation: linear evaluation(Acc) exp	max	Avg	top10	correlation τ	Dense-level Evaluation: object detection(AP) exp	max	Avg	top10	correlation τ
Instance	L_a & L_u	70	76.16	64.39	75.56	-0.50	70	40.37	37.21	40.14	-0.31
	$L_{InfoNCE}$	30	75.47	71.99	74.97	-0.07	30	43.38	40.17	42.33	-0.41
	total	100	76.16	66.51	75.61	-0.45	100	43.38	38.02	42.33	-0.41
Dense	L_a & L_u	70	75.45	64.61	75.01	-0.19	70	43.44	38.99	43.19	-0.22
	$L_{InfoNCE}$	30	75.12	60.85	74.18	-0.01	30	43.71	39.63	42.80	-0.54
	total	100	75.45	63.47	75.13	-0.32	100	43.71	39.2	43.31	-0.12
Random init		1	28.04	-	-	1	31.93	-	-		

Multi-object pretraining and fine-tuning



(a) Linear Evaluation (b) Object Detection

Pretraining	Loss	Instance-level Evaluation: linear evaluation(Acc) exp	max	Avg	top10	correlation τ	Dense-level Evaluation: object detection(AP) exp	max	Avg	top10	correlation τ
Instance	L_a & L_u	40	67.58	59.48	66.75	-0.54	40	44.54	38.27	43.94	-0.23
	$L_{InfoNCE}$	20	67.99	65.05	66.63	-0.67	20	44.51	37.77	42.83	-0.03
	total	60	67.99	61.43	67.17	-0.67	60	44.54	38.09	44.32	-0.13
Dense	L_a & L_u	40	60.19	53.41	58.64	-0.21	40	44.71	36.99	42.90	-0.41
	$L_{InfoNCE}$	20	59.29	46.36	57.30	-0.1	20	44.95	38.69	42.89	-0.54
	total	60	60.19	50.39	58.84	-0.01	60	44.95	37.72	44.12	-0.21
Random init		1	28.04	-	-	-	1	31.93	-	-	-

Confusing positive samples in Dense CL

Single-object dataset *Multi-object dataset*

Pretraining	Instance-level Evaluation: linear evaluation(Acc) exp	max	Avg	top10	correlation τ	Dense-level Evaluation: object detection(AP) exp	max	Avg	top10	correlation τ
Single-object	46	69.94	57.60	68.74	-0.35	46	43.06	40.01	42.78	-0.65
Multi-object	12	54.89	40.30	44.80	-0.15	12	32.49	32.03	32.12	0.03
Random init	1	28.04	-	-	-	1	31.93	-	-	-

positive pairs should share **mutually agreeable information** in multi-object datasets.