CLIP embedding analysis

Semantic extraction and synthesis

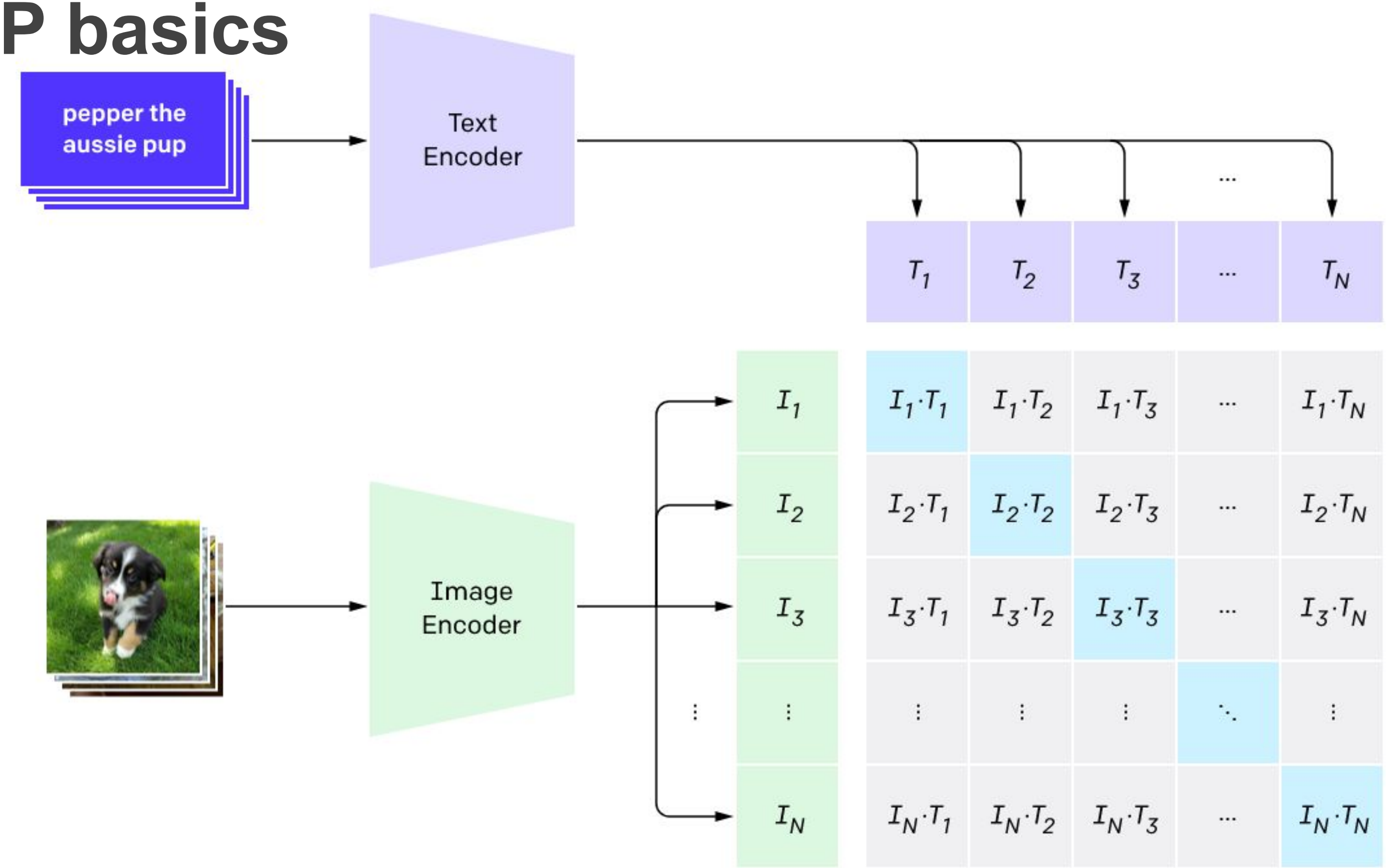
Motivation

- Multi-modality approaches are of big interest of the SOTA research.
- Dimensionality reduction is important task in ML/DL.
- Managed synthesis is key issue in generative models applications.

Tasks

- dimensionality reduction
- embedding clustering
- disentanglement and managed synthesis

CLIP basics

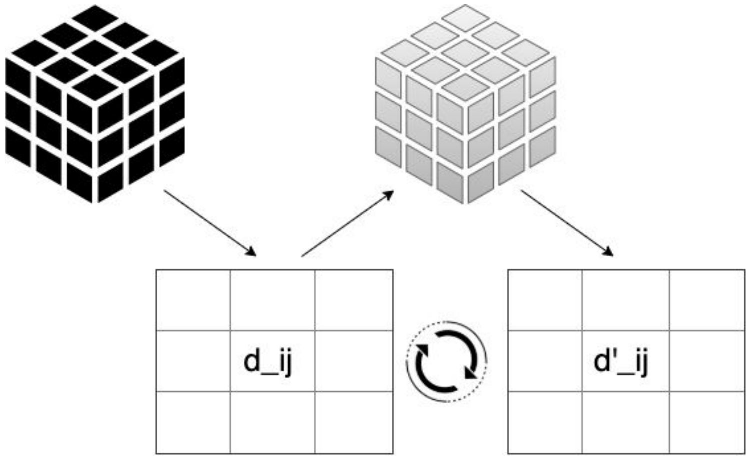


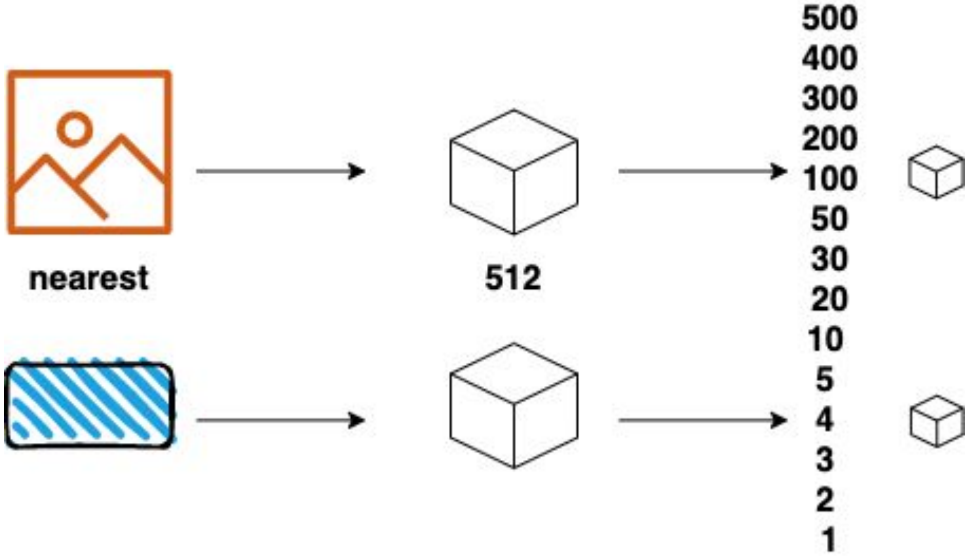
https://arxiv.org/abs/2103.00020

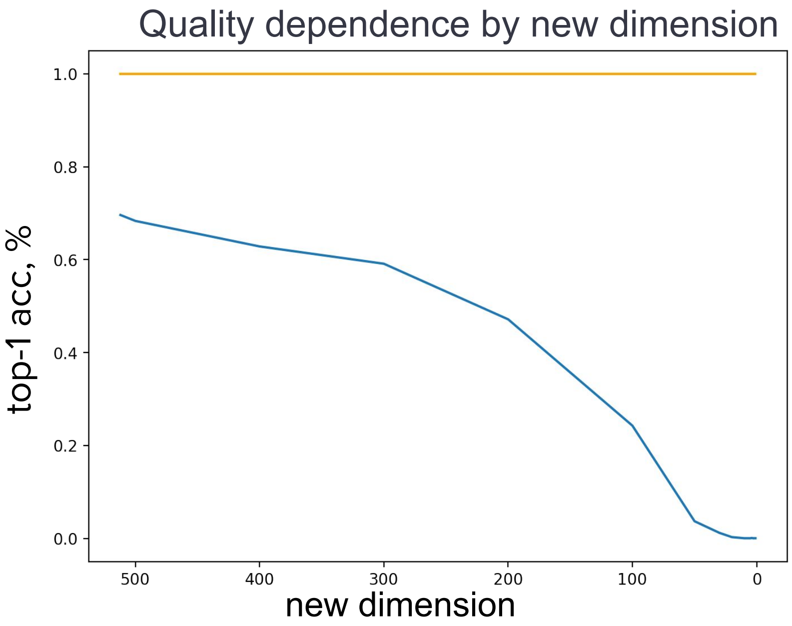
Dimensionality Reduction

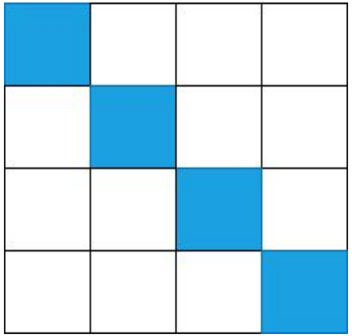
Alexey Kolosov, Ekaterina Orlova

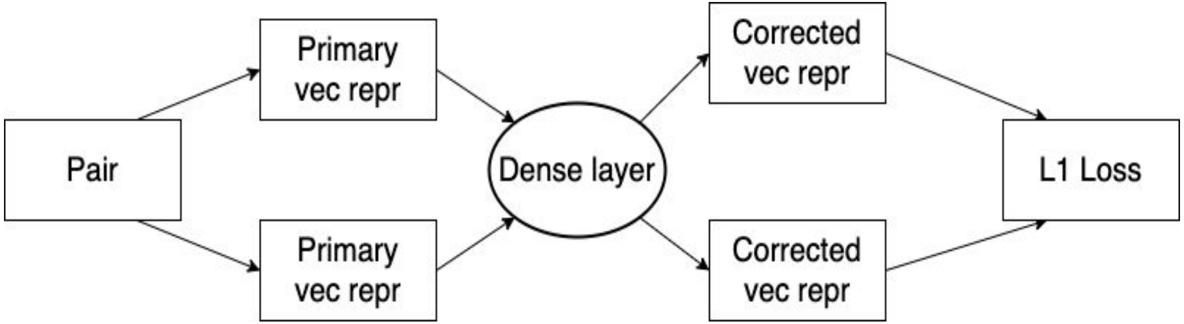
Task: Investigate dimensionality reduction methods and show that for the presented data there exist such embedding dimensionality D’ < D which doesn’t decrease embeddings correspondence quality (top-1 accuracy).











Neural MDS

Alexey Kolosov

Problems statements

1. COCO, isometric, val
2. COCO, isotonic, val

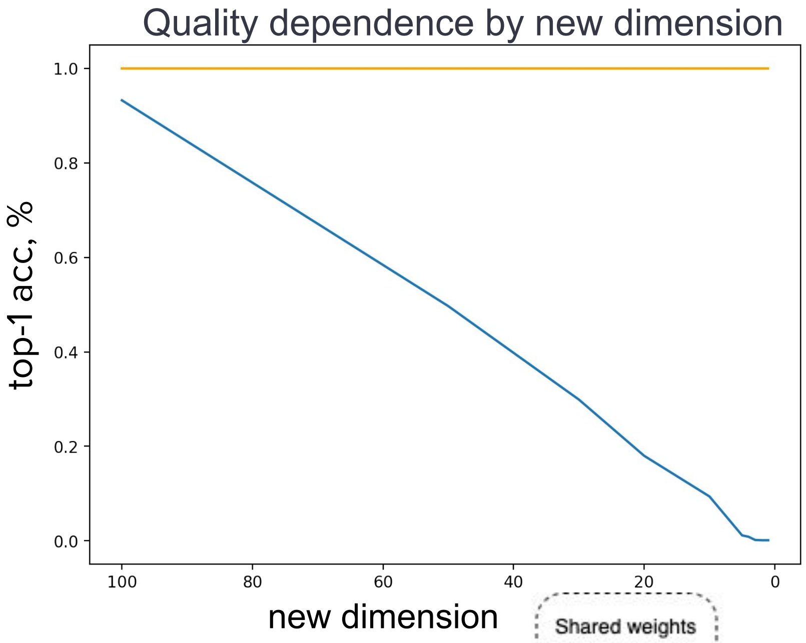
$$d(i,j) = e(g(i), g(j))$$

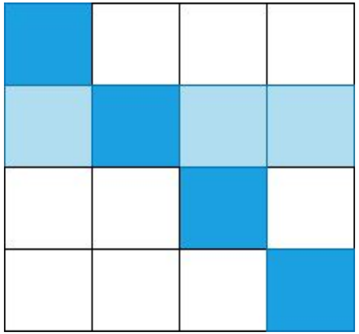
Top-1 accuracy

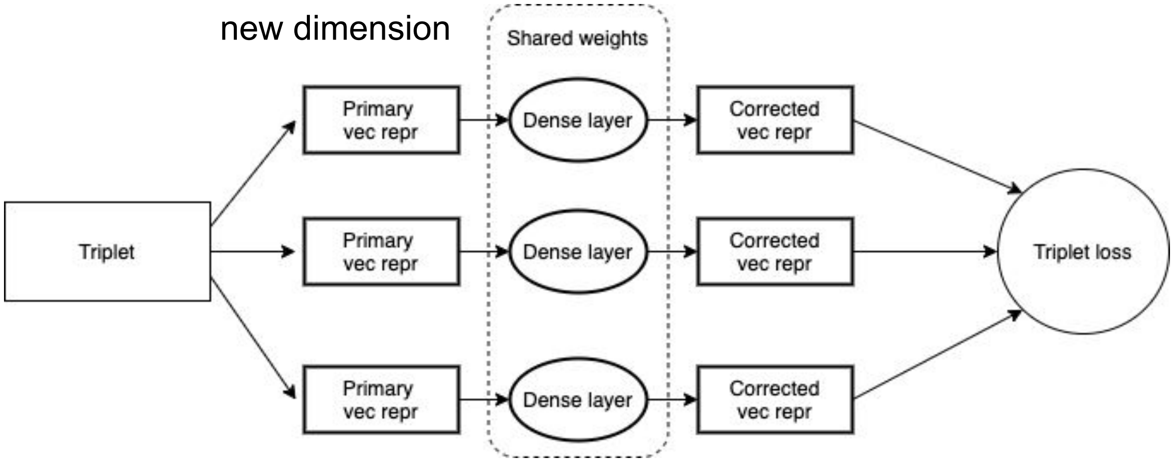
COCO, isometric, val

5000 pairs, 200 epochs

for 512 dim - 1639 pairs







Neural MDS

Alexey Kolosov

Isotonic problem result

$$d(i,j) < d(k,l) \Rightarrow$$
$$e(g(i),g(j)) < e(g(k),g(l))$$

Top-1 accuracy

COCO, isotonic, val

8000000 quads, 200 epochs

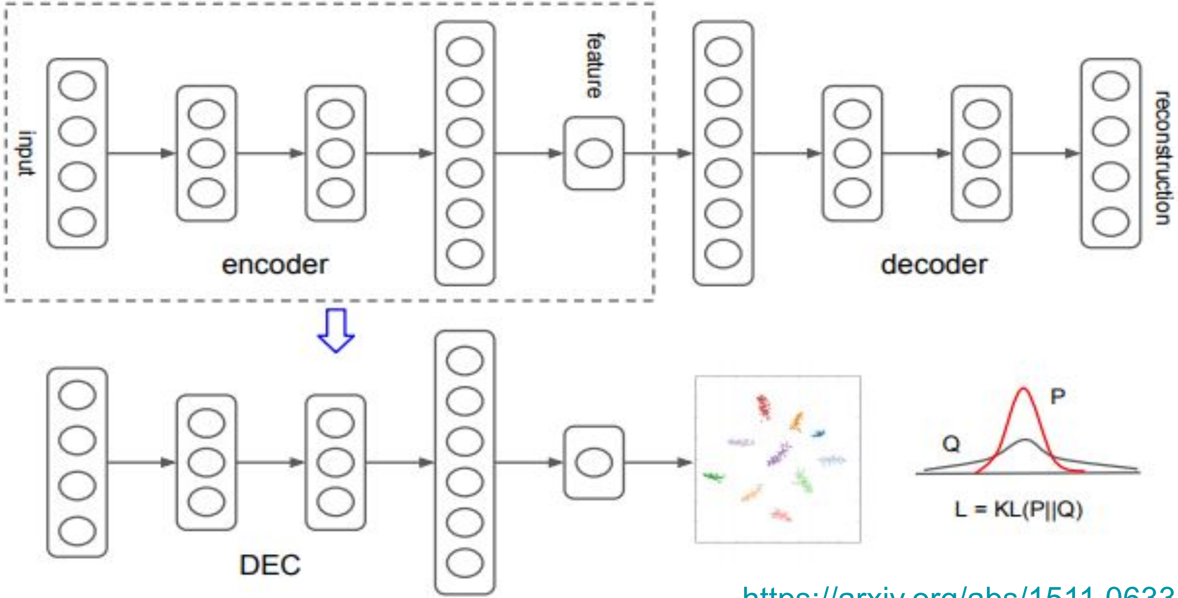
for 512 dim - 1639 pairs

Embeddings clusterization

Abdullaeva Uma, Anna Dmitrienko, Sergey Skorik, Anna Rudenko

Task: Investigate the clusterization methods and show that there exist clusters in the embeddings data. Perform the visualization of these data clusters.

Deep embedded clustering (DEC) model



https://arxiv.org/abs/1511.06335

Metrics

Let "S"- set of n element

$X = \{X1, X2, \ldots, Xn\}$ – the division into classes

$\mathrm{Y} = \{\mathrm{Y1}, \mathrm{Y2}, \ldots, \mathrm{Yn}\}$ - the resulting division into clusters

$X \backslash Y$	Y_1	Y_2	$\ldots$	Y_s	Sums
X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	a_r
Sums	b_1	b_2	$\ldots$	b_s	n

$$p_{ij} = \frac{n_{ij}}{n}, p_i = \frac{a_i}{n}, p_j = \frac{b_j}{n}$$

$$\overbrace{ARI}^{\text{Adjusted Index}} = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected Index}}}$$

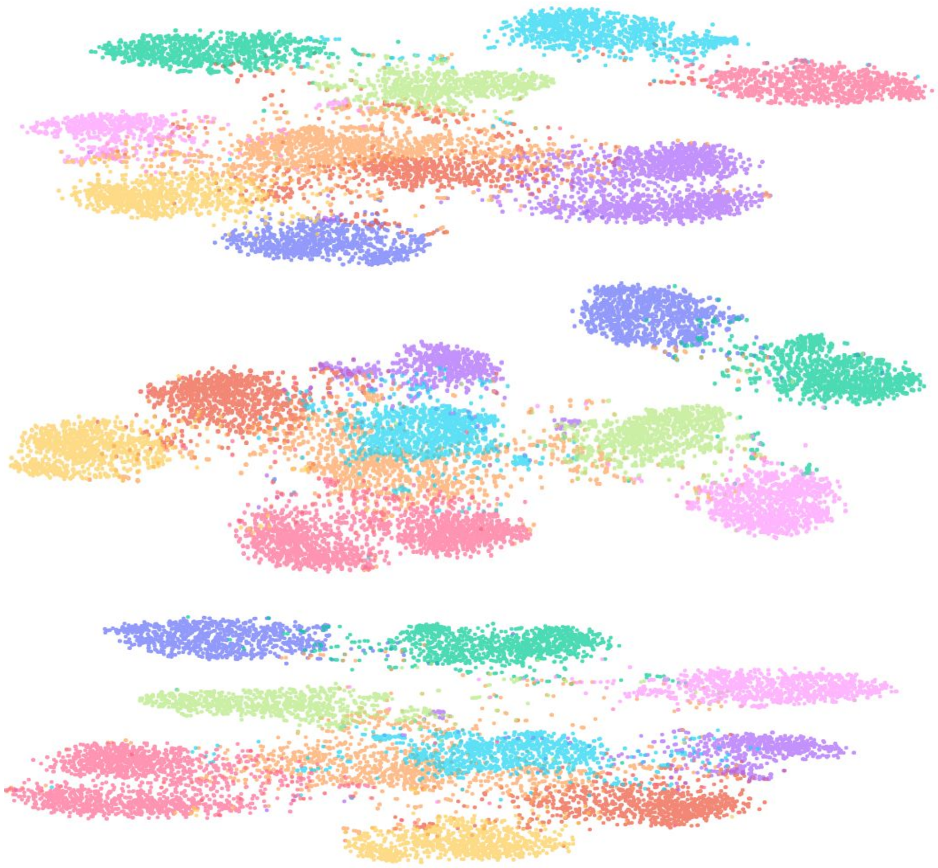
$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]} \qquad \mathrm{RI} = \frac{a+b}{C_2^{n_{samples}}}$$

$$\mathrm{NMI}(U, V) = \frac{\mathrm{MI}(U, V)}{\mathrm{mean}(H(U), H(V))}$$

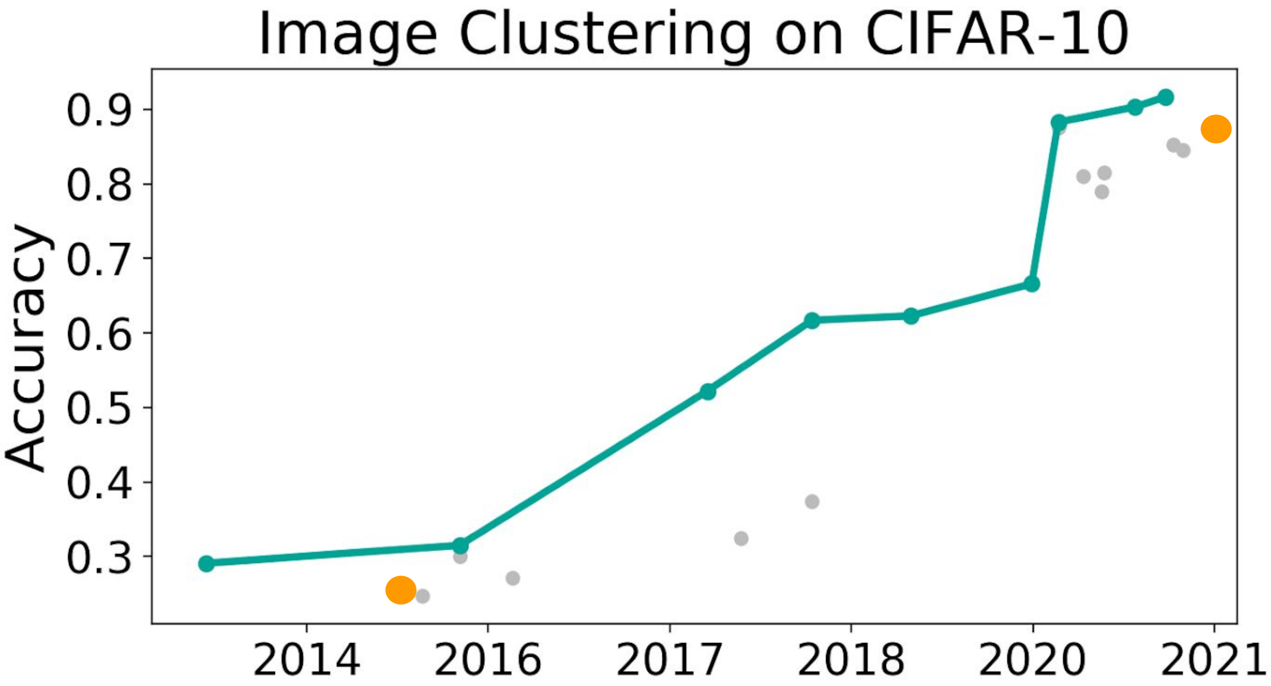
Results

	AMI	ARI	FMI	NMI
K-Means	0.72	0.63	0.67	0.72
Auto - encoder	0.74	0.67	0.70	0.76
DEC	**0.82**	**0.77**	**0.79**	**0.82**

Visualizing clusters using T-SNE



Benchmark



https://paperswithcode.com/sota/image-clustering-on-cifar-10

Advantages of DEC model

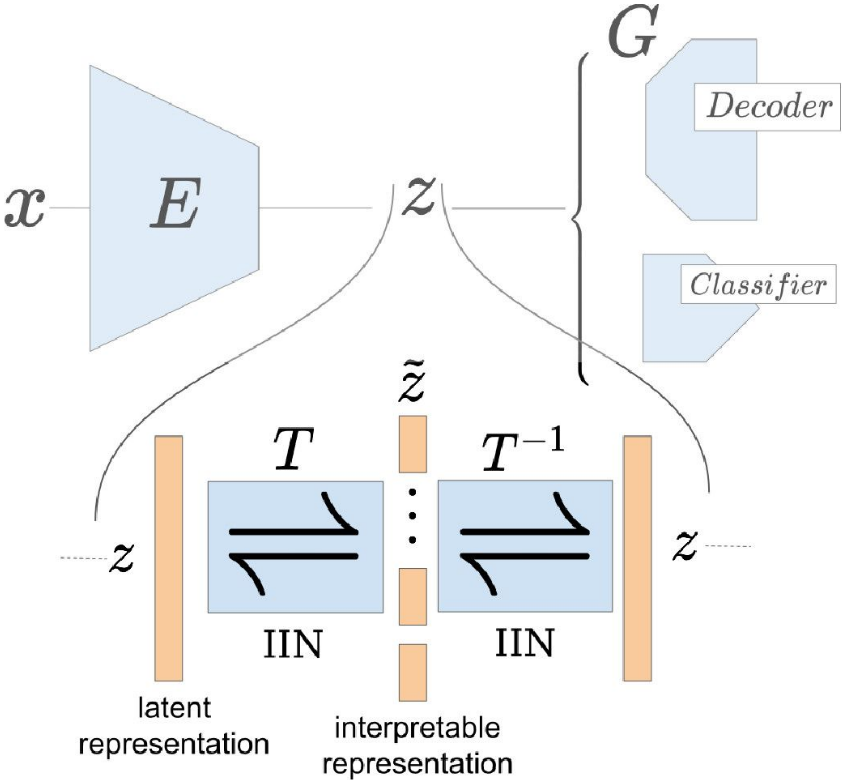
- DEC method is linear in the number of data points and scales gracefully to large datasets

- DEC employs deep neural networks to perform non-linear embedding that is necessary for more complex data

- CLIP + DEC show SOTA results in clustering

Disentanglement

Ekaterina Orlova, Anna Dmitrienko, Sergey Skorik, Anna Rudenko, Abdullaeva Uma



Embedding in latent space:

$z = E(x) \in \mathbb{R}^{H \times W \times C}$

Invertible Interpretation Network:

$T(z) = \tilde{z}$

Modified latent vector z:

$z \rightarrow z^{\star} := T^{-1}(T(z)^{\star})$

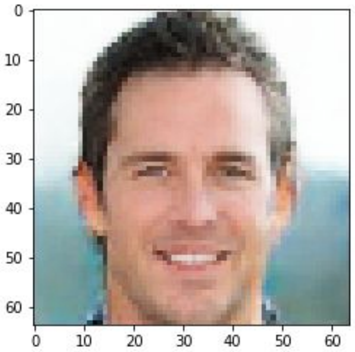
Loss function:

$\mathcal{L} = \sum_{F=1}^{K} \mathbb{E}_{(x^a, x^b) \sim p(x^a, x^b | F)} \ell \left(E\left(x^a\right), E\left(x^b\right) \mid F \right),$

where l – per-example loss

https://arxiv.org/pdf/2004.13166.pdf

Process



E

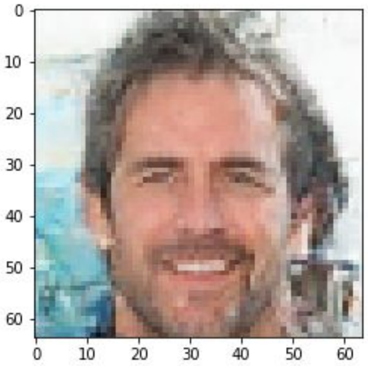
T

-1,99
-1,65
1,94
-0,82
-0,82
-1,06
-1,23
1,06
0,32
2,52
-0,28
-0,40
0,18

-1,99
-1,65
1,94
-0,82
-0,82
-1,06
-1,23
1,06
0,32
2,52
-0,28
-3,00
0,18

D

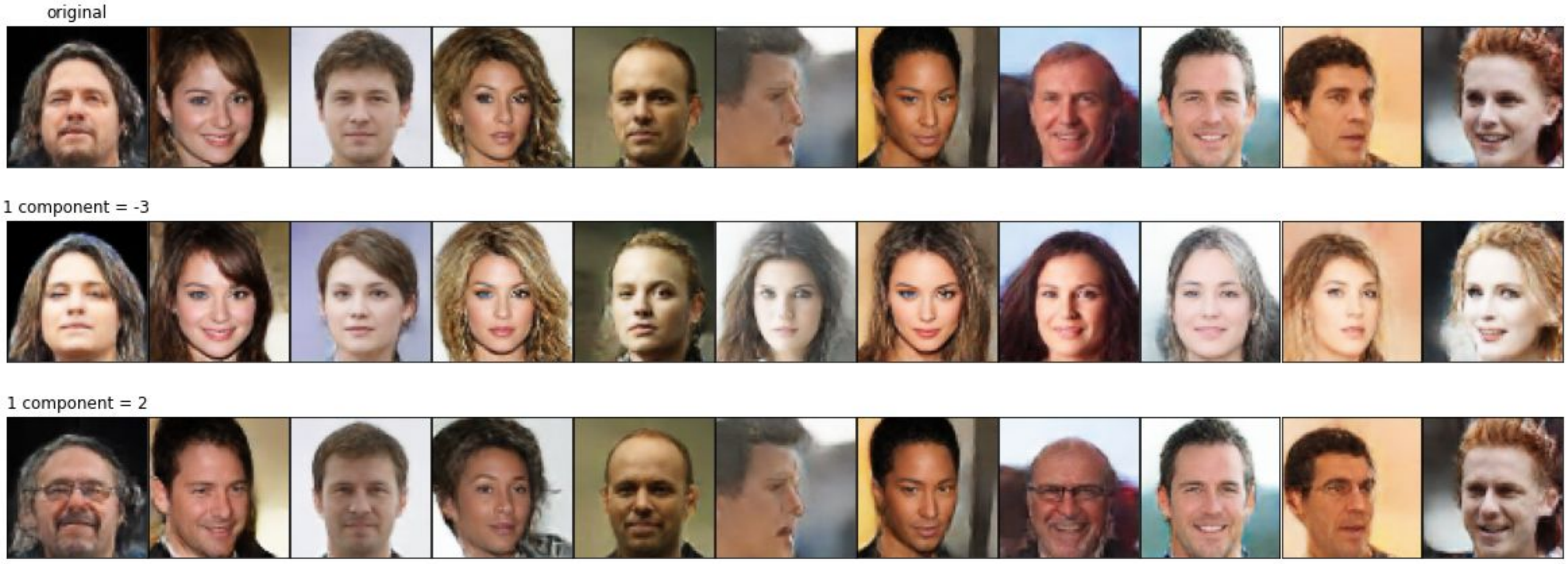
T-1



SelebA - Glasses

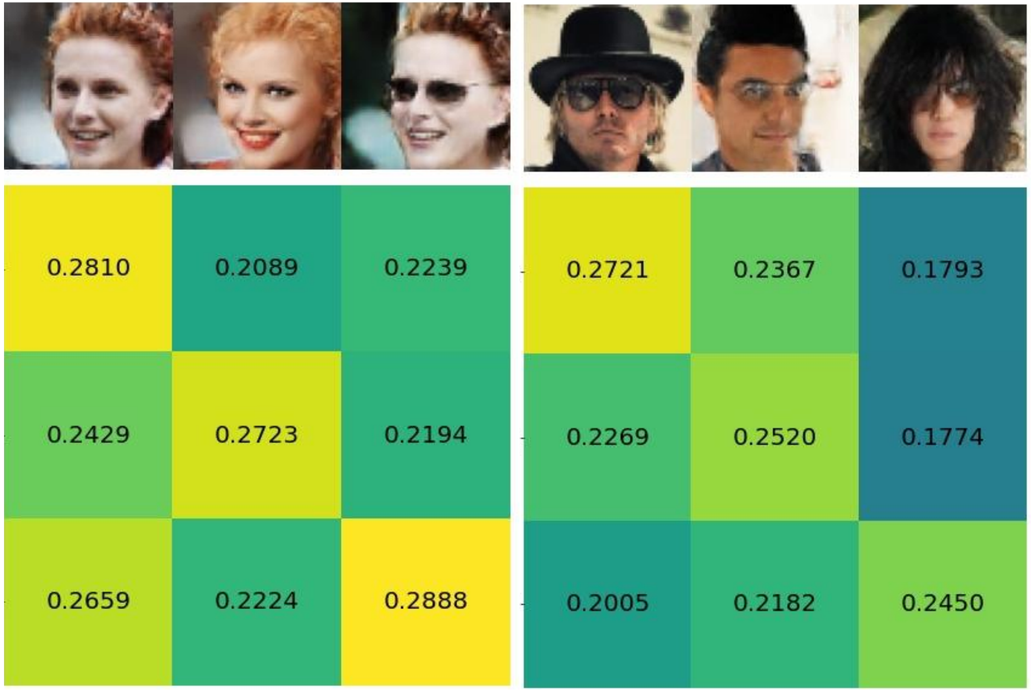
SelebA - Race

SelebA - Sex

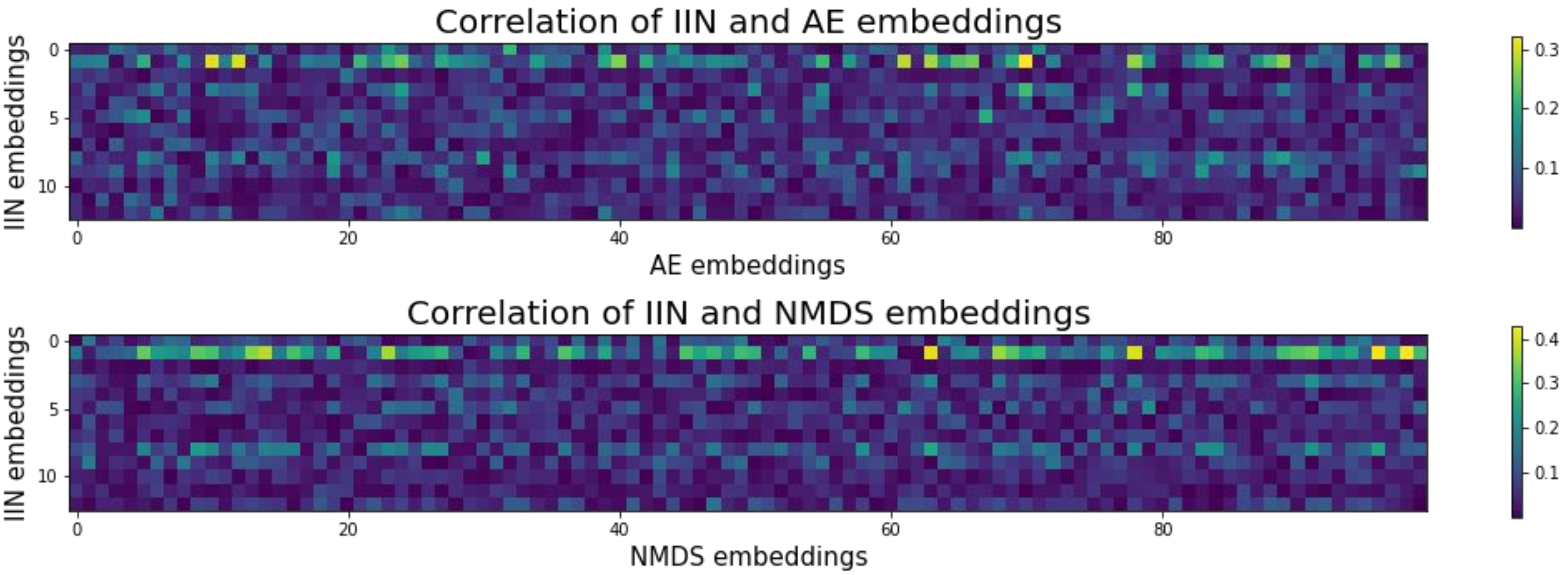


Semantic analysis by CLIP

Cosine similarity between text and image features



Correlation of semantic features and embeddings



Conclusions

- Proposed new method for **dimensionality reduction** of CLIP embeddings
- Proposed and evaluated new method for **clusterization**, close to SOTA
- Interpreted **latent representations** of various VAE

What's next?

- Automatic data augmentation with text descriptions
- Improving managed image synthesis

https://github.com/nihao88/SemanticSirius2021