JOHANNES KEPLER
UNIVERSITY LINZ

# CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP

Seminar in AI
Pascal Pilz
Institute for Machine Learning

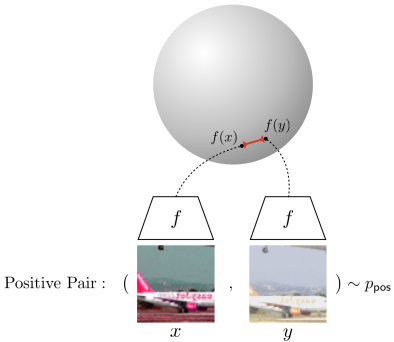
# Motivation and Setup

# Contrastive Learning and Zero-shot Transfer Learning

# Contrastive Learning and Zero-shot Transfer Learning

## Contrastive Learning:

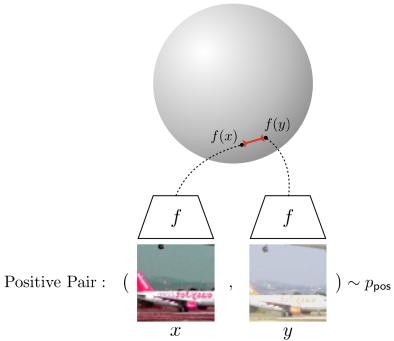
- self-supervised technique



# Contrastive Learning and Zero-shot Transfer Learning

## Contrastive Learning:

- self-supervised technique

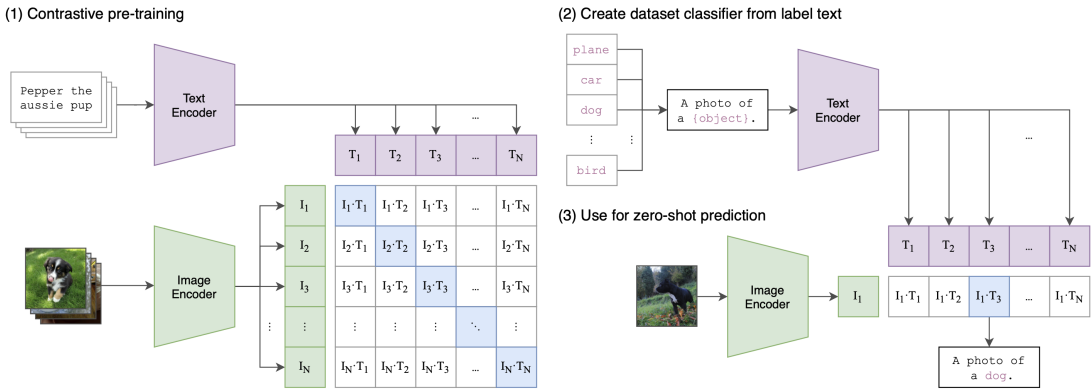



## Zero-shot Transfer Learning:

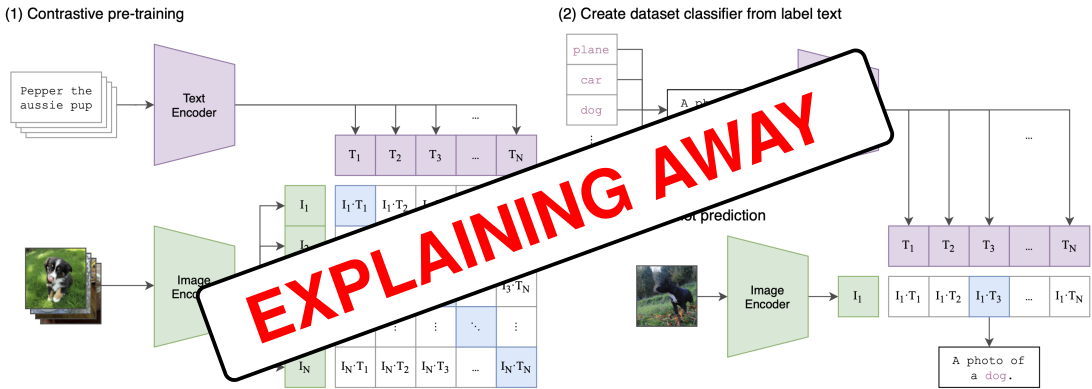
- model needs to generalize well

# CLIP (Contrastive Language-Image Pretraining) by OpenAI

# CLIP (Contrastive Language-Image Pretraining) by OpenAI



# CLIP (Contrastive Language-Image Pretraining) by OpenAI



# Explaining Away — Co-occurrences and Covariance structure

# Explaining Away — Co-occurrences and Covariance structure

# Explaining Away — Co-occurrences and Covariance structure

# Explaining Away — Co-occurrences and Covariance structure

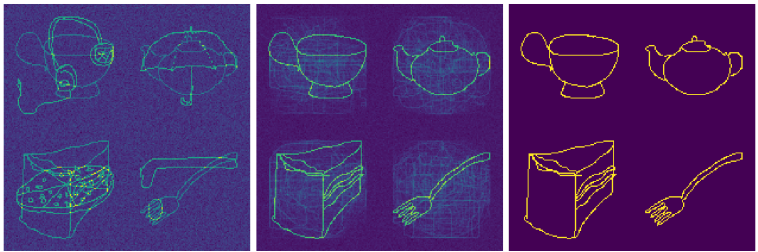
# Modern Hopfield Networks

# Excursus: Modern Hopfield Networks

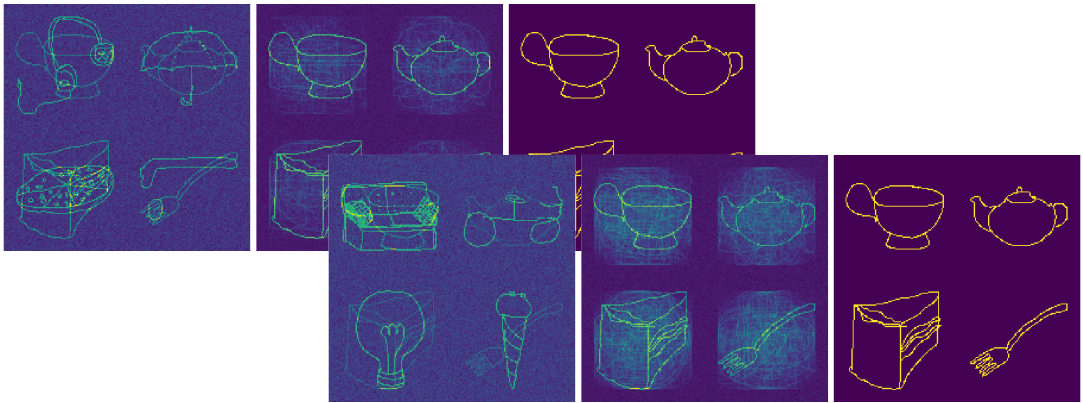
# Excursus: Modern Hopfield Networks


# Amplifying Co-occurrences and Covariance Structures

# Amplifying Co-occurrences and Covariance Structures

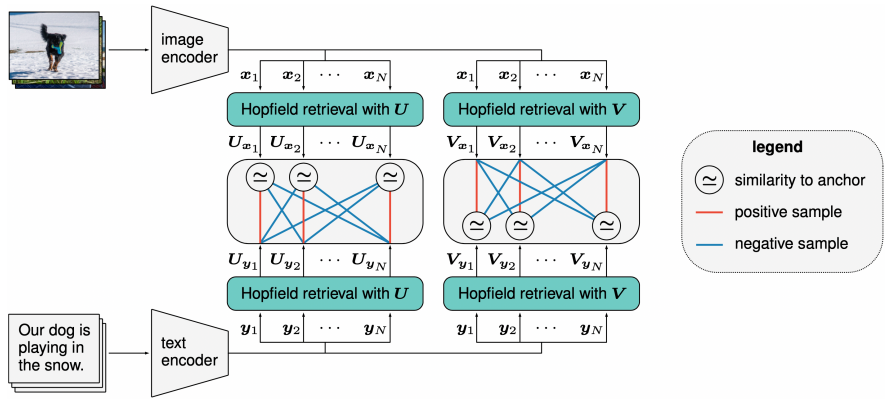


# Amplifying Co-occurrences and Covariance Structures

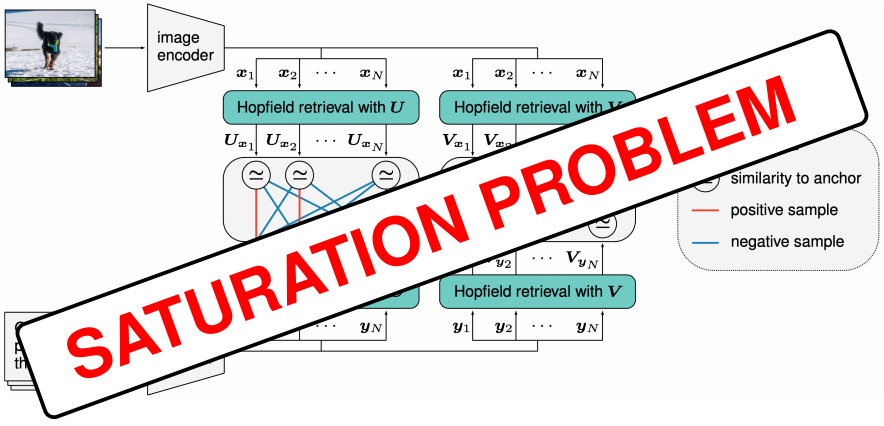


# CLOOB Architecture

# CLOOB Architecture



# CLOOB Architecture



# InfoNCE, InfoLOOB, and CLOOB

# InfoNCE (Noise Contrastive Estimation) — CLIP's Objective

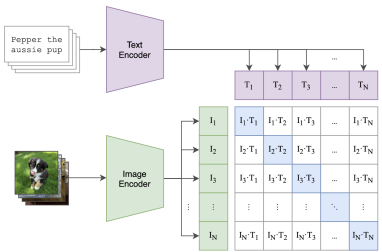
# InfoNCE (Noise Contrastive Estimation) — CLIP's Objective

$$L_{\text{InfoNCE}} = -\frac{1}{N} \ln \sum_{i=1}^{N} \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j=1}^{N} \exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_j)} - \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j=1}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}$$

# InfoNCE (Noise Contrastive Estimation) — CLIP's Objective

$$L_{\text{InfoNCE}} = -\frac{1}{N}\ln\sum_{i=1}^{N}\frac{\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_i)}{\sum_{j=1}^{N}\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_j)} - \frac{1}{N}\sum_{i=1}^{N}\ln\frac{\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_i)}{\sum_{j=1}^{N}\exp(\tau^{-1}\mathbf{x}_j^T\mathbf{y}_i)}$$

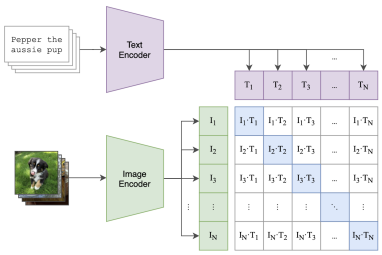
- $\mathbf{x}_i$ — image embedding
- $\mathbf{y}_i$ — text embedding



# InfoNCE (Noise Contrastive Estimation) — CLIP's Objective

$$L_{\text{InfoNCE}} = -\frac{1}{N}\ln\sum_{i=1}^{N}\frac{\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_i)}{\sum_{j=1}^{N}\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_j)} - \frac{1}{N}\sum_{i=1}^{N}\ln\frac{\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_i)}{\sum_{j=1}^{N}\exp(\tau^{-1}\mathbf{x}_j^T\mathbf{y}_i)}$$

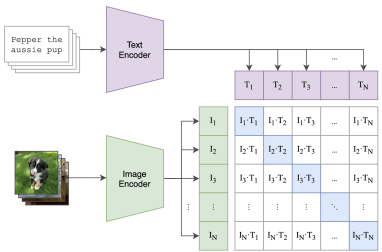
- $\mathbf{x}_i$ — image embedding
- $\mathbf{y}_i$ — text embedding
- $\tau$ — temperature, inverse entropy



# InfoNCE (Noise Contrastive Estimation) — CLIP's Objective

$$L_{\text{InfoNCE}} = -\frac{1}{N}\ln\sum_{i=1}^{N}\frac{\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_i)}{\sum_{j=1}^{N}\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_j)} - \frac{1}{N}\sum_{i=1}^{N}\ln\frac{\exp(\tau^{-1}\mathbf{x}_i^T\mathbf{y}_i)}{\sum_{j=1}^{N}\exp(\tau^{-1}\mathbf{x}_j^T\mathbf{y}_i)}$$

- $\mathbf{x}_i$ — image embedding
- $\mathbf{y}_i$ — text embedding
- $\tau$ — temperature, inverse entropy
- $\|\mathbf{x}_i\| = \|\mathbf{y}_j\| = 1 \Rightarrow \cos(\mathbf{x}_i, \mathbf{y}_j) = \mathbf{x}_i^T\mathbf{y}_j$



# InfoLOOB (Leave One Out Bound) — CLOOB's Objective

$$\mathrm{L}_{\mathrm{InfoLOOB}} = -\frac{1}{N} \ln \sum_{i=1}^{N} \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_j)} - \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}$$

# InfoLOOB (Leave One Out Bound) — CLOOB's Objective

$$L_{\text{InfoLOOB}} = -\frac{1}{N} \ln \sum_{i=1}^{N} \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_j)} - \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}$$

$$L_{\text{InfoNCE}}(\mathbf{y}_1) = -\ln \frac{\overbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}^{a}}{\underbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}_{a} + \underbrace{\sum_{j=2}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_1)}_{b}}$$

# InfoLOOB (Leave One Out Bound) — CLOOB's Objective

$$L_{\text{InfoLOOB}} = -\frac{1}{N} \ln \sum_{i=1}^{N} \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_j)} - \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}$$

$$L_{\text{InfoNCE}}(\mathbf{y}_1) = -\ln \frac{\overbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}^{a}}{\underbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}_{a} + \underbrace{\sum_{j=2}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_1)}_{b}}$$

←saturation problem

# InfoLOOB (Leave One Out Bound) — CLOOB's Objective

$$L_{\text{InfoLOOB}} = -\frac{1}{N} \ln \sum_{i=1}^{N} \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_j)} - \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \mathbf{x}_i^T \mathbf{y}_i)}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_i)}$$

$$L_{\text{InfoNCE}}(\mathbf{y}_1) = -\ln \frac{\overbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}^{a}}{\underbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}_{a} + \underbrace{\sum_{j=2}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_1)}_{b}}$$

←saturation problem

$$L_{\text{InfoLOOB}}(\mathbf{y}_1) = -\ln \frac{\overbrace{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}^{a}}{\underbrace{\sum_{j=2}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y}_1)}_{b}}$$

# CLOOB (Contrastive Leave One Out Boost)

# CLOOB (Contrastive Leave One Out Boost)

```
# image_encoder – ResNet
# text_encoder – Text Transformer

# I[n, h, w, c] – minibatch of images
# T[n, l] – minibatch of texts

# W_i[d_i, d_e] – image projection
# W_t[d_t, d_e] – text projection

# beta – inverse temperature Hopfield retrieval
# tau – temperature InfoLOOB
```

# CLOOB (Contrastive Leave One Out Boost)

```
# image_encoder – ResNet
# text_encoder – Text Transformer

# I[n, h, w, c] – minibatch of images
# T[n, l] – minibatch of texts

# W_i[d_i, d_e] – image projection
# W_t[d_t, d_e] – text projection

# beta – inverse temperature Hopfield retrieval
# tau – temperature InfoLOOB

# extract feature representations
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]
```

# CLOOB (Contrastive Leave One Out Boost)

```
# image_encoder – ResNet
# text_encoder – Text Transformer

# I[n, h, w, c] – minibatch of images
# T[n, l] – minibatch of texts

# W_i[d_i, d_e] – image projection
# W_t[d_t, d_e] – text projection

# beta – inverse temperature Hopfield retrieval
# tau – temperature InfoLOOB

# extract feature representations
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# joint multimodal embedding
x = l2_normalize(I_f @ W_i) #[n, d_e]
y = l2_normalize(T_f @ W_t) #[n, d_e]
```

# CLOOB (Contrastive Leave One Out Boost)

```
# image_encoder – ResNet
# text_encoder – Text Transformer

# I[n, h, w, c] – minibatch of images
# T[n, l] – minibatch of texts

# W_i[d_i, d_e] – image projection
# W_t[d_t, d_e] – text projection

# beta – inverse temperature Hopfield retrieval
# tau – temperature InfoLOOB

# extract feature representations
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# joint multimodal embedding
x = l2_normalize(I_f @ W_i) #[n, d_e]
y = l2_normalize(T_f @ W_t) #[n, d_e]

# Hopfield retrieval H with batch stored
# H(beta, A, B) = B.T @ softmax(beta * A @ B.T)
U_x = H(beta, x, x).T #[n, d_e]
U_y = H(beta, y, x).T #[n, d_e]
V_x = H(beta, x, y).T #[n, d_e]
V_y = H(beta, y, y).T #[n, d_e]
```

# CLOOB (Contrastive Leave One Out Boost)

```
# image_encoder – ResNet
# text_encoder – Text Transformer

# I[n, h, w, c] – minibatch of images
# T[n, l] – minibatch of texts

# W_i[d_i, d_e] – image projection
# W_t[d_t, d_e] – text projection

# beta – inverse temperature Hopfield retrieval
# tau – temperature InfoLOOB

# extract feature representations
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# joint multimodal embedding
x = l2_normalize(I_f @ W_i) #[n, d_e]
y = l2_normalize(T_f @ W_t) #[n, d_e]

# Hopfield retrieval H with batch stored
# H(beta, A, B) = B.T @ softmax(beta * A @ B.T)
U_x = H(beta, x, x).T #[n, d_e]
U_y = H(beta, y, x).T #[n, d_e]
V_x = H(beta, x, y).T #[n, d_e]
V_y = H(beta, y, y).T #[n, d_e]

# normalize retrievals
U_x = l2_normalize(U_x) #[n, d_e]
U_y = l2_normalize(U_y) #[n, d_e]
V_x = l2_normalize(V_x) #[n, d_e]
V_y = l2_normalize(V_y) #[n, d_e]
```

# CLOOB (Contrastive Leave One Out Boost)

```
# image_encoder – ResNet
# text_encoder – Text Transformer

# I[n, h, w, c] – minibatch of images
# T[n, l] – minibatch of texts

# W_i[d_i, d_e] – image projection
# W_t[d_t, d_e] – text projection

# beta – inverse temperature Hopfield retrieval
# tau – temperature InfoLOOB

# extract feature representations
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# joint multimodal embedding
x = l2_normalize(I_f @ W_i) #[n, d_e]
y = l2_normalize(T_f @ W_t) #[n, d_e]

# Hopfield retrieval H with batch stored
# H(beta, A, B) = B.T @ softmax(beta * A @ B.T)
U_x = H(beta, x, x).T #[n, d_e]
U_y = H(beta, y, x).T #[n, d_e]
V_x = H(beta, x, y).T #[n, d_e]
V_y = H(beta, y, y).T #[n, d_e]

# normalize retrievals
U_x = l2_normalize(U_x) #[n, d_e]
U_y = l2_normalize(U_y) #[n, d_e]
V_x = l2_normalize(V_x) #[n, d_e]
V_y = l2_normalize(V_y) #[n, d_e]

# loss: info_loob(tau, anchors, samples)
# samples contain pos. and neg. embeddings
loss_i = info_loob(tau, U_x, U_y)
loss_t = info_loob(tau, V_y, V_x)
loss = (loss_i + loss_t) * tau
```

# Experiments

# Experiments

# Experiments

- Conceptual Captions (CC) Pretraining
  - rich textual description
  - only 2.9 million images

# Experiments

- Conceptual Captions (CC) Pretraining
  - rich textual description
  - only 2.9 million images
- Yahoo Flickr Create Commons (YFCC) Pretraining
  - less rich textual description
  - 15 million images

# Experiments

- Conceptual Captions (CC) Pretraining
  - rich textual description
  - only 2.9 million images
- Yahoo Flickr Create Commons (YFCC) Pretraining
  - less rich textual description
  - 15 million images
- Both tested on seven image classification datasets

# Experiments

- Conceptual Captions (CC) Pretraining
  - rich textual description
  - only 2.9 million images
- Yahoo Flickr Create Commons (YFCC) Pretraining
  - less rich textual description
  - 15 million images
- Both tested on seven image classification datasets
- Ablation studies
  - Modern Hopfield networks
  - InfoLOOB

# CLIP and CLOOB: Results

CC — mean accuracy over 5 runs

| Dataset | CLIP RN-50 | CLOOB RN-50 | CLIP* RN-50 | CLOOB* RN-50 |
|---|---|---|---|---|
| Birdsnap | $2.26 \pm 0.20$ | $\mathbf{3.06 \pm 0.30}$ | $2.8 \pm 0.16$ | $\mathbf{3.24 \pm 0.31}$ |
| Country211 | $0.67 \pm 0.11$ | $0.67 \pm 0.05$ | $0.7 \pm 0.04$ | $0.73 \pm 0.05$ |
| Flowers102 | $12.56 \pm 0.38$ | $13.45 \pm 1.19$ | $13.32 \pm 0.43$ | $14.36 \pm 1.17$ |
| GTSRB | $7.66 \pm 1.07$ | $6.38 \pm 2.11$ | $8.96 \pm 1.70$ | $7.03 \pm 1.22$ |
| UCF101 | $20.98 \pm 1.55$ | $22.26 \pm 0.72$ | $21.63 \pm 0.65$ | $\mathbf{23.03 \pm 0.85}$ |
| Stanford Cars | $0.91 \pm 0.10$ | $\mathbf{1.23 \pm 0.10}$ | $0.99 \pm 0.16$ | $\mathbf{1.41 \pm 0.32}$ |
| ImageNet | $20.33 \pm 0.28$ | $\mathbf{23.97 \pm 0.15}$ | $21.3 \pm 0.42$ | $\mathbf{25.67 \pm 0.22}$ |
| ImageNet V2 | $20.24 \pm 0.50$ | $\mathbf{23.59 \pm 0.15}$ | $21.24 \pm 0.22$ | $\mathbf{25.49 \pm 0.11}$ |

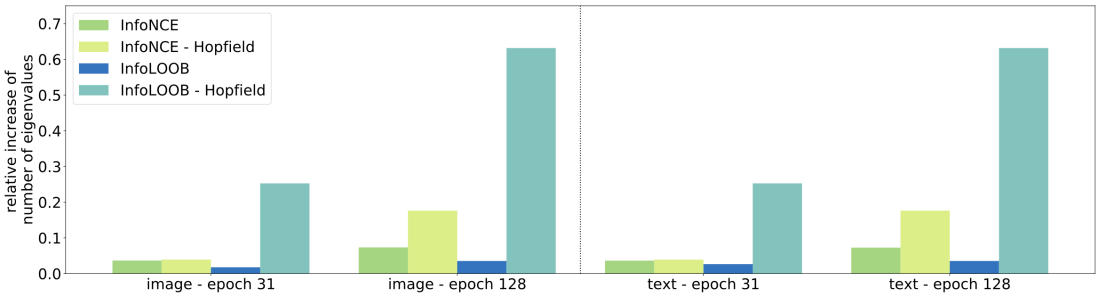
Bold: statistically significant

YFCC — one run

| | RN-50 | | RN-101 | | RN-50x4 | |
|---|---|---|---|---|---|---|
| Dataset | CLIP | CLOOB | CLIP | CLOOB | CLIP | CLOOB |
| Birdsnap | 21.8 | **28.9** | 22.6 | **30.3** | 20.8 | **32.0** |
| Country211 | 6.9 | **7.9** | 7.8 | **8.5** | 8.1 | **9.3** |
| Flowers102 | 48.0 | **55.1** | 48.0 | **55.3** | 50.1 | **54.3** |
| GTSRB | 7.9 | **8.1** | 7.4 | **11.6** | 9.4 | **11.8** |
| UCF101 | **27.2** | 25.3 | 28.6 | **28.8** | 31.0 | **31.9** |
| Stanford Cars | 3.7 | **4.1** | 3.8 | **5.5** | 3.5 | **6.1** |
| ImageNet | 34.6 | **35.7** | 35.3 | **37.1** | 37.7 | **39.0** |
| ImageNet V2 | 33.4 | **34.6** | 34.1 | **35.6** | 35.9 | **37.3** |

Bold: higher value

# Results Ablation Study



# Conclusion

# Critical Assessment

# Critical Assessment

- Reproducibility
  - hyperparameters, experiments well-defined ✓
  - code, datasets public ✓

# Critical Assessment

- Reproducibility
  - hyperparameters, experiments well-defined ✓
  - code, datasets public ✓
- Thoroughness
  - two theorems, both proven ✓
  - all assumptions listed ✓

# Critical Assessment

- Reproducibility
  - hyperparameters, experiments well-defined ✓
  - code, datasets public ✓
- Thoroughness
  - two theorems, both proven ✓
  - all assumptions listed ✓
- Fairness of comparison
  - as faithful to original CLIP as possible ✓
  - CLIP's 400 million image dataset not public

# Critical Assessment

- Reproducibility
  - hyperparameters, experiments well-defined ✓
  - code, datasets public ✓
- Thoroughness
  - two theorems, both proven ✓
  - all assumptions listed ✓
- Fairness of comparison
  - as faithful to original CLIP as possible ✓
  - CLIP's 400 million image dataset not public

Paper fulfills all NeurIPS check marks

# Summary

➔ CLIP

# Summary

➔ CLIP

✗ explaining away

# Summary

➔ CLIP

✗ explaining away

➔ Modern Hopfield networks

# Summary

➔ CLIP

✗ explaining away

➔ Modern Hopfield networks

✗ saturation problem

# Summary

➔ CLIP

✗ explaining away

➔ Modern Hopfield networks

✗ saturation problem

➔ CLOOB

# Summary

➔ CLIP

✗ explaining away

➔ Modern Hopfield networks

✗ saturation problem

➔ CLOOB

✓ Modern Hopfield networks

# Summary

➔ CLIP
- ✗ explaining away

➔ Modern Hopfield networks
- ✗ saturation problem

➔ CLOOB
- ✓ Modern Hopfield networks
- ✓ InfoLOOB

# Summary

➔ CLIP

✗ explaining away

➔ Modern Hopfield networks

✗ saturation problem

➔ CLOOB

✓ Modern Hopfield networks

✓ InfoLOOB

➔ ☺

# Summary

➔ CLIP

✗ explaining away

➔ Modern Hopfield networks

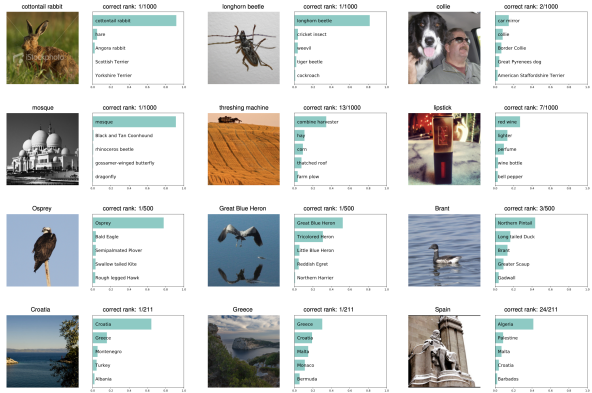
✗ saturation problem

➔ CLOOB

✓ Modern Hopfield networks

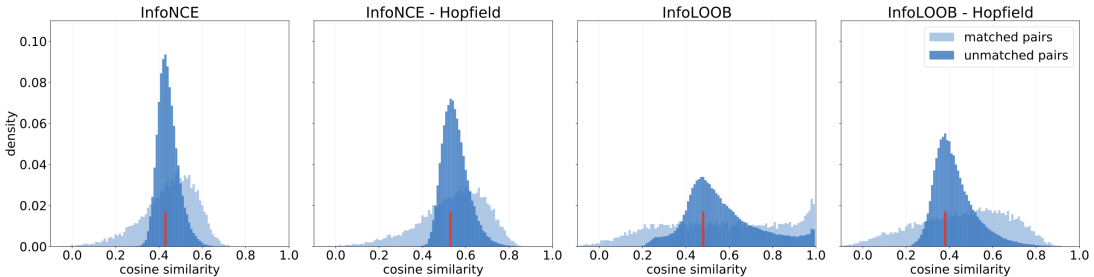
✓ InfoLOOB

➔ ☺



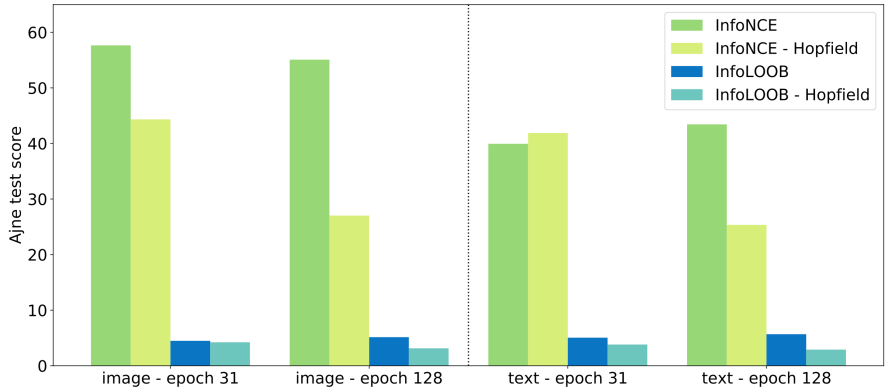
# Additional Material

# Distribution of Cosine Similarity




Distribution of cosine similarity of matched pairs and of the 10 unmatched pairs that have the highest similarity score with anchor

# Ajne Test


InfoNCE
InfoNCE - Hopfield
InfoLOOB
InfoLOOB - Hopfield
Ajne test score
0
10
20
30
40
50
60
image - epoch 31
image - epoch 128
text - epoch 31
text - epoch 128

# InfoLOOB – Further details

| | $\mathbf{u}_i$ - stored image emb. | $\mathbf{v}_i$ - stored text emb. |
|---|---|---|
| $\mathbf{x}_i$ - image query emb. | $\mathbf{U}_{\mathbf{x}_i}$ - image-retr. image emb. | $\mathbf{V}_{\mathbf{x}_i}$ - image-retr. text emb. |
| $\mathbf{y}_i$ - text query emb. | $\mathbf{U}_{\mathbf{y}_i}$ - text-retr. image emb. | $\mathbf{V}_{\mathbf{y}_i}$ - text-retr. text emb. |

$$L_{\text{InfoLOOB}} = -\frac{1}{N} \ln \sum_{i=1}^{N} \frac{\exp(\tau^{-1} \mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_i})}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{U}_{\mathbf{x}_i}^T \mathbf{U}_{\mathbf{y}_j})} - \frac{1}{N} \sum_{i=1}^{N} \ln \frac{\exp(\tau^{-1} \mathbf{V}_{\mathbf{x}_i}^T \mathbf{V}_{\mathbf{y}_i})}{\sum_{j \neq i}^{N} \exp(\tau^{-1} \mathbf{V}_{\mathbf{x}_j}^T \mathbf{V}_{\mathbf{y}_i})}$$

# Gradients of InfoLOOB and InfoNCE

$$\frac{\partial}{\partial \mathbf{y}} \mathsf{L}_{\text{InfoNCE}}(\mathbf{y}) = \frac{\partial}{\partial \mathbf{y}} - \ln \frac{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}{\sum_{j=1}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y})} = -\tau^{-1} \mathbf{y}^T \mathbf{x}_1 + \tau^{-1} \operatorname{lse}(\tau^{-1}, \mathbf{X}^T \mathbf{y})$$

$$\frac{\partial}{\partial \mathbf{y}} \mathsf{L}_{\text{InfoLOOB}}(\mathbf{y}) = \frac{\partial}{\partial \mathbf{y}} - \ln \frac{\exp(\tau^{-1} \mathbf{x}_1^T \mathbf{y}_1)}{\sum_{j \neq 1}^{N} \exp(\tau^{-1} \mathbf{x}_j^T \mathbf{y})} = -\tau^{-1} \mathbf{y}^T \mathbf{x}_1 + \tau^{-1} \operatorname{lse}(\tau^{-1}, \tilde{\mathbf{X}}^T \mathbf{y})$$

$$\frac{\partial}{\partial \mathbf{y}} \mathsf{L}_{\text{InfoNCE}}(\mathbf{y}) = -\tau^{-1}(1 - p_1)(\mathbf{x}_1 - \tilde{\mathbf{X}} \operatorname{softmax}(\tau^{-1} \tilde{\mathbf{X}}^T \mathbf{y})) = (1 - p_1) \frac{\partial}{\partial \mathbf{y}} \mathsf{L}_{\text{InfoLOOB}}(\mathbf{y})$$

$$\operatorname{lse}(\beta, \boldsymbol{\alpha}) = \beta^{-1} \log \left( \sum_{i=1}^{N} \exp(\beta \alpha_i) \right)$$

$$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N), \quad \tilde{\mathbf{X}} = (\mathbf{x}_2, \ldots, \mathbf{x}_N)$$