Multi-modal alignment using representation codebook
Duan, J., Chen, L., Tran, S., Yang, J., Xu, Y., Zeng, B., & Chilimbi, T.

Presented by Muntasir Wahed
Motivation

- Aligning feature representations of multi-modal models
- Bridging early fusion models and late fusion models
- Improve intra-modality alignment

Contrastive Learning

- **Contrasts** every sample with all samples in the minibatch
- **Positive**: Different **views** of the same image
- **Negative**: All other samples in the minibatch

Momentum Contrast (MoCo) - Motivation

- Contrastive learning requires a large amount of negative samples
  - Large batch size - constrained by GPU memory
  - Memory bank - stale representations
- Maintain a queue of embeddings instead, evolving over time

\[ \theta_k \leftarrow m\theta_k + (1 - m)\theta_q \]
Momentum Contrast (MoCo)

(a) end-to-end

q · k

encoder q

x^q

encoder k

x^k

(b) memory bank

q · k

encoder

x^q

sampling

memory bank

x^k

(c) MoCo

q

encoder

x

query

x^q

memory

k

queue

k_0, k_1, k_2, ...

momentum encoder

x^key

x^key

x^key

x^key

x^key

k

similarity

Momentum Contrast (MoCo)

- Different view of the same image as query and key for the positive logit
- Back propagation only happens for the query
- Negative logits extracted from the queue

Algorithm 1: Pseudocode of MoCo in a PyTorch-like style.

```python
# f_q, f_k: encoder networks for query and key
# queue: dictionary as a queue of K keys (C x K)
# m: momentum
# t: temperature

f_k.params = f_q.params  # initialize
for x in loader:  # load a minibatch x with N samples
    x_q = aug(x)  # a randomly augmented version
    x_k = aug(x)  # another randomly augmented version

    q = f_q.forward(x_q)  # queries: N x C
    k = f_k.forward(x_k)  # keys: N x C
    k.detach()  # no gradient to keys

    # positive logits: N x 1
    l_pos = bmm(q.view(N, 1, C), k.view(N, C, 1))

    # negative logits: N x K
    l_neg = mm(q.view(N, C), queue.view(N, C, K))

    # logits: N x (1+K)
    logits = cat([l_pos, l_neg], dim=1)

    # contrastive loss, Eqn. (1)
    labels = zeros(N)  # positives are the 0-th
    loss = CrossEntropyLoss(logits / t, labels)

    # SGD update: query network
    loss.backward()
    update(f_q.params)

    # momentum update: key network
    f_k.params = m * f_k.params + (1 - m) * f_q.params

    # update dictionary
    enqueue(queue, k)  # enqueue the current minibatch
    dequeue(queue)  # dequeue the earliest minibatch
```

bmm: batch matrix multiplication; mm: matrix multiplication; cat: concatenation.
Scalability

Align Before Fuse (ALBEF) - Motivation

- Address the limitations of late fusion models
  - The image and text embeddings in their own spaces
  - Use of annotation-expensive and compute-expensive object detector
  - The datasets are inherently noisy, and existing pre-training objectives such as MLM may overfit
Image Text Contrastive Learning (ITC) Loss

- $g_v$ and $g_w$ are linear transformations that map the [CLS] embeddings to normalized lower-dimensional (256-d) representations.
- Two queues to store the most recent $M$ image-text representations, the normalized features denoted by $g'_v(v'_{cls})$ and $g'_w(w'_{cls})$

$$s(I, T) = g_v(v_{cls})^\top g'_w(w'_{cls}) \quad \quad \quad \quad s(T, I) = g_w(w_{cls})^\top g'_v(v'_{cls}).$$

$$p_{m}^{i2t}(I) = \frac{\exp(s(I, T_m)/\tau)}{\sum_{m=1}^{M} \exp(s(I, T_m)/\tau)}, \quad \quad \quad \quad p_{m}^{t2i}(T) = \frac{\exp(s(T, I_m)/\tau)}{\sum_{m=1}^{M} \exp(s(T, I_m)/\tau)}$$

$$\mathcal{L}_{itc} = \frac{1}{2} \mathbb{E}_{(I, T) \sim D} \left[ H(y^{i2t}(I), p^{i2t}(I)) + H(y^{t2i}(T), p^{t2i}(T)) \right]$$
Masked Language Modeling (MLM) Loss

- Predict ground-truth labels of masked text tokens.

\[ \mathcal{L}_{mlm} = \mathbb{E}_{(I, \hat{T}) \sim D} H(y^{msk}, p^{msk}(I, \hat{T})) \]
Image Text Matching (ITM) Loss

- [CLS] token used as the joint representation of the image-text pair.
- Use a fully connected layer to predict the matching probability.

$$
\mathcal{L}_{itm} = \mathbb{E}_{(I,T) \sim D} H(y_{itm}^{itm}, p_{itm}^{itm}(I, T))
$$
ALBEF Pre-training

Training Objective

\[ \mathcal{L} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itm}} \]

Momentum Distillation

- ITC and MLM penalize all negative predictions regardless of their correctness
- Modify the loss functions to learn from pseudo-targets generated by the momentum model instead
- A weighted combination of the original loss and the KL-divergence between the model’s prediction and the pseudo-targets
Align Before Fuse (ALBEF) - Benefits

- Aligns the image and text embeddings to improve cross-modal learning
- Improves the unimodal encoders to better understand the semantic meaning of images and texts
- A common low-dimensional space to embed images and texts
  - facilitates extraction of informative samples through our contrastive hard negative mining
- Model not penalized for producing reasonable outputs different from the web annotation, resulting in more stable learning
Codebook Learning with Distillation (CODIS)

- Inspired by ALBEF
  - Consider both intra and cross modal alignment in $L_{\text{ica}}$
- Multimodal codebook learning
  - Learnable codebook for both modalities
  - Predict codebook assignment using either text or image
- Teacher-student distillation
  - Guides the codebook learning
  - Improves unimodal and cross-modal alignment
Relation to Prior Work

- A hybrid between the late-fusion and early-fusion works
  - ALBEF [1] is also doing something similar
- Codebook used by BEiT [2] and SOHO [3] to quantize the visual space
  - Contrary to them, this work quantized the join output space
- The loss function inspired by SwAV [4]
  - SwAV contrasts one view of the image with the assigned cluster of the same image
  - This paper contrasts across modalities

---

Optimal Transport

- Map one distribution to another distribution
- $n!$ combinations available for two discrete distributions consisting of $n$ items each
- Find the most optimal (with least cost) solution to this matching problem

Image credit: http://alexwilliams.info/itsneuronalblog/2020/10/09/optimal-transport/
Optimal Transport (cont.)

- Tries to minimize the optimal transport distance between prototypes and features
- Maps each feature with a prototype
- Sparse solution, with at most \((2r - 1)\) \((r = \max(N, K))\) non-zero elements

**Algorithm 2 IPOT Algorithm.**

1. **Input:** distance/similarity matrix \(Z, C, \epsilon\), probability vectors \(\mu, \nu\)
2. \(\sigma = \frac{1}{n} 1_n, T^{(1)} = 11^T\)
3. \(D_{ij} = d(z_i, c_j), A_{ij} = e^{-\frac{D_{ij}}{\epsilon}}\)
4. for \(t = 1, 2, 3\ldots\) do
5. \(Q = A \odot T^{(t)} // \odot\) is Hadamard product
6. for \(k = 1, 2, 3\ldots K\) do
7. \(\delta = \frac{\mu}{nQ_\sigma}, \sigma = \frac{\nu}{nQ^T \delta}\)
8. end for
9. \(T^{(t+1)} = \text{diag}(\delta)Q\text{diag}(\sigma)\)
10. end for
11. Return \(T\)

\[
\mathcal{L}_{ot} = \min_{T \in \Pi(u, v)} \sum_{i=1}^{N} \sum_{j=1}^{K} T_{ij} \cdot d(z_i^m, c_j) = \min_{T \in \Pi(u, v)} \langle T, D \rangle
\]
Multimodal Codebook Learning

- Codebook (prototypes)
  - Encodes image and text into a joining embedding space
- Optimal Transport, T, used as ground-truth signals

\[
\begin{align*}
\mathcal{L}_{t2p}(Z_t, C, T_{i2p}) &= H(P_{t2p}, T_{i2p}), \\
\mathcal{L}_{i2p}(Z_v, C, T_{t2p}) &= H(P_{i2p}, T_{t2p}), \\
P_{t2p} &= \text{SoftMax}(Z_t C / \gamma), \\
P_{i2p} &= \text{SoftMax}(Z_v C / \gamma)
\end{align*}
\]
Codebook Loss

- Both the text-to-prototype ($L_{t2p}$) loss and image-to-prototype ($L_{i2p}$) loss chain features from both modalities
- When calculating the transport plan, use the teacher encoders
- Losses back propagated to both the codebook and the student encoders

$$
L_{code} = L_{ot}(Z^m_v, C) + L_{ot}(Z^m_t, C) \\
+ L_{t2p}(Z_t, C, T_{t2p}) + L_{i2p}(Z_v, C, T_{i2p})
$$
Teacher-student Distillation Learning

- Store features from teacher encoders $z_v^m$ and $z_t^m$ in memory queues $Q_v$ and $Q_t$.
- Pseudo negatives are sampled from the queues.
- Also use the teacher encoders to provide soft distillation targets, $y_{i2t}$, $y_{t2i}$, $y_{t2t}$, $y_{i2i}$.
- Teacher encoders are updated using momentum.

\[
p_{t2t}(T) = \exp \frac{z_t z_v^m T}{\gamma} / \sum_{z_v^{m'} \in Q_v} \exp \frac{z_t z_v^{m'} T}{\gamma}
\]

\[
p_{i2t}(I) = \exp \frac{z_v z_t^m T}{\gamma} / \sum_{z_t^{m'} \in Q_t} \exp \frac{z_v z_t^{m'} T}{\gamma}
\]

\[
p_{i2i}(I) = \exp \frac{z_v z_v^m T}{\gamma} / \sum_{z_v^{m'} \in Q_v} \exp \frac{z_v z_v^{m'} T}{\gamma}
\]

\[
p_{t2t}(T) = \exp \frac{z_t z_t^m T}{\gamma} / \sum_{z_t^{m'} \in Q_t} \exp \frac{z_t z_t^{m'} T}{\gamma}
\]

\[
f_t = \alpha f_t + (1 - \alpha) f_s, g_t = \alpha g_t + (1 - \alpha) g_s
\]
Training Objective

- Simultaneously optimize the codebook and the student encoders
- $L_{MLM}$ conditioned on both surrounding text tokens and image representations
- For $L_{itm}$, sample one negative text/image using contrastive similarity distribution.

$$L_{final} = L_{mlm} + L_{itm} + L_{ica} + L_{code}$$
Experimental Setup (Downstream Tasks)

- Image-Text Retrieval
  - Zero-shot
  - After-finetuning
- Visual Question Answering (VQA)
- Visual Reasoning (NLVR²)
- Visual Entailment (SNLI-VE)
## Experimental Results (Zero-Shot)

<table>
<thead>
<tr>
<th>Method</th>
<th>MSCOCO (5K)</th>
<th></th>
<th></th>
<th>Flickr30K (1K)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text Retrieval</td>
<td></td>
<td></td>
<td>Image Retrieval</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R@1  R@5  R@10</td>
<td></td>
<td></td>
<td>R@1  R@5  R@10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageBERT [36]</td>
<td>44.0  71.2  80.4</td>
<td>32.3  59.0  70.2</td>
<td>70.7  90.2  94.0</td>
<td>54.3  79.6  87.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unicoder-VL [24]</td>
<td>-  -  -</td>
<td>-  -  -</td>
<td>64.3  85.8  92.3</td>
<td>48.4  76.0  85.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNITER [8]</td>
<td>-  -  -</td>
<td>-  -  -</td>
<td>80.7  95.7  98.0</td>
<td>66.2  88.4  92.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViLT [22]</td>
<td>56.5  82.6  89.6</td>
<td>40.4  70.0  81.1</td>
<td>73.2  93.6  96.5</td>
<td>55.0  82.5  89.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIP [37]</td>
<td>58.4  81.5  88.1</td>
<td>37.8  62.4  72.2</td>
<td>88.0  98.7  99.4</td>
<td>68.7  90.6  95.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALIGN [21]</td>
<td>58.6  83.0  89.7</td>
<td>45.6  69.8  78.6</td>
<td>88.6  98.7  99.7</td>
<td>75.7  93.8  96.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALBEF 4M [25]</td>
<td>68.6  89.5  94.7</td>
<td>50.1  76.4  84.5</td>
<td>90.5  98.8  99.7</td>
<td>76.8  93.7  96.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>71.5</strong></td>
<td><strong>91.1</strong></td>
<td><strong>95.5</strong></td>
<td><strong>53.9</strong></td>
<td><strong>79.5</strong></td>
<td><strong>87.1</strong></td>
</tr>
</tbody>
</table>
### Experimental Results (Finetuning)

<table>
<thead>
<tr>
<th>Method</th>
<th>MSCOCO (5K) Text Retrieval</th>
<th>MSCOCO (5K) Image Retrieval</th>
<th>Flickr30K (1K) Text Retrieval</th>
<th>Flickr30K (1K) Image Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
</tr>
<tr>
<td>ImageBERT [36]</td>
<td>66.4</td>
<td>89.8</td>
<td>94.4</td>
<td>50.5</td>
</tr>
<tr>
<td>UNITER [8]</td>
<td>65.7</td>
<td>88.6</td>
<td>93.8</td>
<td>52.9</td>
</tr>
<tr>
<td>VILLA [14]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OSCAR [28]</td>
<td>70.0</td>
<td>91.1</td>
<td>95.5</td>
<td>54.0</td>
</tr>
<tr>
<td>ViLT [22]</td>
<td>61.5</td>
<td>86.3</td>
<td>92.7</td>
<td>42.7</td>
</tr>
<tr>
<td>UNIMO [27]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SOHO [20]</td>
<td>66.4</td>
<td>88.2</td>
<td>93.8</td>
<td>50.6</td>
</tr>
<tr>
<td>ALBEF 4M [25]</td>
<td>73.1</td>
<td>91.4</td>
<td>96.0</td>
<td>56.8</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>75.3</strong></td>
<td><strong>92.6</strong></td>
<td><strong>96.6</strong></td>
<td><strong>58.7</strong></td>
</tr>
</tbody>
</table>

R@1, R@5, and R@10 denote the recall at 1, 5, and 10, respectively.
# Experimental Results (VQA, NVLR$^2$, SNLI-VE)

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA</th>
<th>NLVR$^2$</th>
<th>SNLI-VE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>test-dev</td>
<td>test-std</td>
<td>dev</td>
</tr>
<tr>
<td>VisualBERT [26]</td>
<td>70.80</td>
<td>71.00</td>
<td>67.40</td>
</tr>
<tr>
<td>LXMERT [43]</td>
<td>72.42</td>
<td>72.54</td>
<td>74.90</td>
</tr>
<tr>
<td>12-in-1 [32]</td>
<td>73.15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UNITER [8]</td>
<td>72.70</td>
<td>72.91</td>
<td>77.18</td>
</tr>
<tr>
<td>ViLT [22]</td>
<td>70.94</td>
<td>-</td>
<td>75.24</td>
</tr>
<tr>
<td>OSCAR [28]</td>
<td>73.16</td>
<td>73.44</td>
<td>78.07</td>
</tr>
<tr>
<td>VILLA [14]</td>
<td>73.59</td>
<td>73.67</td>
<td>78.39</td>
</tr>
<tr>
<td>ALBEF 4M [25]</td>
<td>74.54</td>
<td>74.70</td>
<td>80.24</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>74.86</strong></td>
<td><strong>74.97</strong></td>
<td><strong>80.50</strong></td>
</tr>
</tbody>
</table>
## Ablation Studies

<table>
<thead>
<tr>
<th>Objective functions</th>
<th>Text Retrieval</th>
<th>MSCOCO (5K)</th>
<th>Image Retrieval</th>
<th>Flickr30K (1K)</th>
<th>Text Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>a: MLM+ITM+ITC (cross align)</td>
<td>68.60</td>
<td>89.50</td>
<td>94.70</td>
<td>50.10</td>
<td>76.40</td>
</tr>
<tr>
<td>b: MLM+ITM+ITC (intra + cross)</td>
<td>69.86</td>
<td>89.48</td>
<td>94.42</td>
<td>50.52</td>
<td>77.02</td>
</tr>
<tr>
<td>a + codebook (teacher feature)</td>
<td>70.74</td>
<td>89.54</td>
<td>94.88</td>
<td>51.39</td>
<td>77.86</td>
</tr>
<tr>
<td>b + codebook (student feature)</td>
<td>71.12</td>
<td>89.62</td>
<td>94.78</td>
<td>51.40</td>
<td>77.42</td>
</tr>
<tr>
<td>b + codebook (teacher feature)</td>
<td><strong>71.10</strong></td>
<td><strong>90.60</strong></td>
<td><strong>95.10</strong></td>
<td><strong>52.10</strong></td>
<td><strong>78.00</strong></td>
</tr>
</tbody>
</table>
## Ablation Studies

<table>
<thead>
<tr>
<th></th>
<th>TR@1</th>
<th>TR@5</th>
<th>TR@10</th>
<th>IR@1</th>
<th>IR@5</th>
<th>IR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBEF</td>
<td>55.70</td>
<td>81.92</td>
<td>88.78</td>
<td>41.08</td>
<td>69.01</td>
<td>78.86</td>
</tr>
<tr>
<td>0.5x codebook</td>
<td>58.66</td>
<td>83.9</td>
<td>90.64</td>
<td>43.74</td>
<td>72.10</td>
<td>81.58</td>
</tr>
<tr>
<td>2.0x codebook</td>
<td>59.02</td>
<td>84.46</td>
<td>91.06</td>
<td>43.62</td>
<td>71.69</td>
<td>81.12</td>
</tr>
<tr>
<td>3K codewords</td>
<td>58.96</td>
<td>84.28</td>
<td>90.98</td>
<td>44.66</td>
<td>72.31</td>
<td>81.68</td>
</tr>
<tr>
<td>500 codewords</td>
<td>55.52</td>
<td>81.68</td>
<td>89.28</td>
<td>41.53</td>
<td>68.75</td>
<td>78.43</td>
</tr>
<tr>
<td>Ours</td>
<td>59.38</td>
<td>84.04</td>
<td>91.20</td>
<td>44.71</td>
<td>72.63</td>
<td>81.69</td>
</tr>
</tbody>
</table>
Qualitative Analysis

“A person does a trick on a skateboard while a man takes a picture”

“a giraffe walking through trees on a sunny day”
Strengths

- Proposes intra-modal alignment to further improve cross-modal alignment
  - Ablation studies show that it improves the performance significantly
- The proposed teacher-student distillation framework works well
  - the slowly evolving teacher encoder helps the training process
- Strong results across multiple experiments against state-of-the-art baselines
- GRAD-CAM visualization is very interesting
Weaknesses

- Updating all the encoders simultaneously
  - Can lead to unpredictable oscillations
- Various issues with optimal transport
  - Why optimal transport instead of a simpler clustering algorithm?
  - Not clear if each codebook has only one image and vice versa
- Issues with notation.
  - Assumes too much about reader’s prior knowledge.
  - Prior concepts used in the paper not explained properly
  - Missing notations for the algorithm for Optimal Transport
- Some minor errors in the tables
Future Works

- Instead of aligning the embeddings in a single layer, we can experiment with aligning them over multiple layers.
  - This might have the effect of aligning the embeddings at different semantic levels.
- Using the codebooks, we can sample hard negatives for the $L_{itm}$ loss.
Discussion

- What is the reason for using optimal transport?
- Why do you think the intra-modal alignment is helping improve the results?