Hi-CLIP - CONTRASTIVE LANGUAGE-IMAGE PRETRAINING WITH HIERARCHY-AWARE ATTENTION

ICLR 2023, 19 citations

CAP 6412 Group 8
Group Members

- Andrew Ballen
- Zhen Hao Sia
- Nicholas Tidwell
- Anantapadmanaabha Prasannakumar
- Abdulrahman Al Sumaih
Introduction

• Hierarchies are natural in images and text
• CLIP is unable to understand hierarchical relationships
• HiCLIP is designed with the intent of learning these
• Hierarchical understanding allows it to perform better on downstream tasks.
Introduction Cont'd

- Does not require any additional labeled data
- Directly learn hierarchy from raw image-text pair
- Inspired by NLP tree transformer, which automatically finds hierarchies
Background of Visual and Text Hierarchy

- Tree Transformer merges tokens for text
- Group Transformer merges patches for images
- Once tokens/patches are merged, they cannot be split in later layers.
- Merges are performed by an attention mask “C”
- C is then included in traditional attention scoring equation.

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_h}} \right) V
\]

\[
\text{Hierarchy Attention} = \left( C \odot \text{softmax} \left( \frac{QK^T}{\sqrt{d_h}} \right) \right) V.
\]
Background of Visual and Text Hierarchy Cont'd.
Visualization Results

Example of Unsupervised Hierarchy Induction

"a group of zebra standing next to each other on a dirt field"

(b) Language Hierarchy
Visualization Results

Example of Unsupervised Hierarchy Induction

(a) Visual Hierarchy
Hierarchy Mask, C (1D Text)

1) Neighbouring attention scores

\[
s_{i,i+1} = \frac{(t_i W'_Q) \cdot (t_{i+1} W'_K)}{\sigma_t}, \quad \text{word tokens } (t_i, t_{i+1})
\]

Two learnable key, query matrices \(W'_Q, W'_K\)

2) Per token \(t_i\) softmax

\[
p_{i,i+1}, p_{i,i-1} = \text{softmax} \left( s_{i,i+1}, s_{i,i-1} \right)
\]
Hierarchy Mask, C (1D Text)

3. Neighbouring affinity scores

neighbor pairs \((t_i, t_{i+1})\)

\[
\hat{a}_{i,i+1} = \sqrt{p_{i,i+1} \cdot p_{i+1,i}}
\]
Hierarchy Mask, C (1D Text)

IMPORTANT: Enforce non-splittable property

- prevent merged tokens from splitting in subsequent layers
- mathematical intuition:

\[ a_{i,i+1}^l \geq a_{i,i+1}^{l-1} \]

\[ a_{i,i+1}^l = a_{i,i+1}^{l-1} + (1 - a_{i,i+1}^{l-1}) \hat{a}^l_{i,i+1} \]
Hierarchy Mask, C (1D Text)

4) Tendency to merge, $C_{i,j}$

$$C_{i,j} = \prod_{k=i}^{j-1} a_{k,k+1}$$
Hierarchy Mask, C (1D Text) Example

Rick

1 = p_{0,1}
0 = p_{0,-1}

\hat{a}_{0,1}

a_{0,1} from previous layer

And

p_{1,2}

p_{1,0}

\hat{a}_{1,2}

a_{1,2}

Morty

s_{0,1}

s_{1,0}

s_{2,1}

s_{1,2}

\hat{a}_{1,2}

a_{1,2} from previous layer

p_{2,3} = 0
p_{2,1} = 1
## Hierarchy Mask, C (1D Text) Example

<table>
<thead>
<tr>
<th>Hierarchy Mask, C</th>
<th>Rick</th>
<th>And</th>
<th>Morty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rick</td>
<td>1</td>
<td>$a_{0,1}$</td>
<td>$a_{0,1} \cdot a_{1,2}$</td>
</tr>
<tr>
<td>And</td>
<td>$a_{0,1}$</td>
<td>1</td>
<td>$a_{1,2}$</td>
</tr>
<tr>
<td>Morty</td>
<td>$a_{0,1} \cdot a_{1,2}$</td>
<td>$a_{1,2}$</td>
<td>1</td>
</tr>
</tbody>
</table>
Hierarchy Mask, C (2D Images)

- text > 1 dimension
- 1-2 neighbours (left or right)

Rick ➔ And ➔ Morty
Hierarchy Mask, C (2D Images)

- images > 2 dimensions
- 2-4 neighbours (left, right, up, down)
Hierarchy Mask, C (2D Images)

1) Neighbouring attention scores

\[
S(i,j),(i',j') = \frac{(t_{i,j} W_{Q''}) \cdot (t_{i',j'} W_{K''})}{\sigma_v}
\]

- 2-4 neighbours for each patch \( t_{i,j} \) (up, down, left, right)

\((i',j') \in \{(i + \delta, j + \eta); \delta, \eta \in \{-1, +1\}\} \equiv A\)
Hierarchy Mask, C (2D Images)

2) Per patch \(( t_{i,j} )\) softmax

- 4 values for 4 neighbours

\[
\{ p_{(i,j),(i',j')} \} = \text{softmax}(\{ s_{(i,j),(i',j')} ; (i',j') \in A \})
\]

- if no neighbour in given direction > softmax = 0
3. Neighbouring affinity scores

- for neighbouring patches $t_{i,j}$ and $t_{i',j'}$:

$$\hat{a}_{(i,j),(i',j')} = \sqrt{p(i,j),(i',j') \cdot p(i',j'),(i,j)}$$
Hierarchy Mask, C (2D Images)

IMPORTANT: Enforce non-splittable property

- prevent merged patches from splitting in subsequent layers
- mathematical intuition:

\[ a_{(i,j),(i',j')}^l \geq a_{(i,j),(i',j')}^{l-1} \]

\[ a_{(i,j),(i',j')}^l = a_{(i,j),(i',j')}^{l-1} + \left( 1 - a_{(i,j),(i',j')}^{l-1} \right) \hat{a}_{(i,j),(i',j')}^l \]
Hierarchy Mask, C (2D Images)

4) Tendency to merge, \( C_{(i_1,j_1),(i_2,j_2)} = \max(C_1, C_2) \)

- vertical, then horizontal traversal

\[
C_1 = \prod_{n=i_1}^{i_2-1} a_{(n,j_1),(n+1,j_1)} \prod_{m=j_1}^{j_2-1} a_{(i_2,m),(i_2,m+1)}
\]

- horizontal, then vertical traversal

\[
C_2 = \prod_{m=j_1}^{j_2-1} a_{(i_1,m),(i_1,m+1)} \prod_{n=i_1}^{i_2-1} a_{(n,j_2),(n+1,j_2)}
\]
Experimental Settings

Pretraining Datasets
● YFCC15M Dataset
● Custom 30M Dataset = YFCC15M+CC3M+CC12M

Downstream Datasets
● 11 visual recognition datasets under zero-shot setting:
  ○ ImageNet, C10, C100, StanfordCars, Caltech101, Flowers102, SUN397, DTD, FGVAircraft, OxfordPets, Food101
Implementation Details

- Vision encoder > Group Transformer
- Text encoder > Tree Transformer
- Image size is 224 x 224; input text sequence padded to 77;
- Embedding size is fixed at 512
- Same training hyperparameters for all models
- Scaling factor of Hierarchy-aware attention set to 256 for both Group and Tree Transformer
## Zero-shot Setting Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>C10</th>
<th>C100</th>
<th>F101</th>
<th>Pets</th>
<th>Flow.</th>
<th>SUN</th>
<th>Cars</th>
<th>DTD</th>
<th>Cal.</th>
<th>Air.</th>
<th>IN</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>15M</td>
<td>63.7</td>
<td>33.2</td>
<td>34.6</td>
<td>20.1</td>
<td>50.1</td>
<td>35.7</td>
<td>2.6</td>
<td>15.5</td>
<td>59.9</td>
<td>1.2</td>
<td>32.8</td>
<td>31.8</td>
</tr>
<tr>
<td>SLIP</td>
<td>15M</td>
<td>50.7</td>
<td>25.5</td>
<td>33.3</td>
<td>23.5</td>
<td>49.0</td>
<td>34.7</td>
<td>2.8</td>
<td>14.4</td>
<td>59.9</td>
<td>1.7</td>
<td>34.3</td>
<td>30.0</td>
</tr>
<tr>
<td>FILIP</td>
<td>15M</td>
<td>65.5</td>
<td>33.5</td>
<td>43.1</td>
<td>24.1</td>
<td>52.7</td>
<td>50.7</td>
<td>3.3</td>
<td>24.3</td>
<td>68.8</td>
<td>3.2</td>
<td>39.5</td>
<td>37.2</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>15M</td>
<td>74.1</td>
<td>46.0</td>
<td>51.2</td>
<td>37.8</td>
<td>60.9</td>
<td>50.6</td>
<td>4.5</td>
<td>23.1</td>
<td>67.4</td>
<td>3.6</td>
<td>40.5</td>
<td>41.8 (+10.0)</td>
</tr>
<tr>
<td>DeCLIP</td>
<td>15M</td>
<td>66.7</td>
<td>38.7</td>
<td>52.5</td>
<td>33.8</td>
<td>60.8</td>
<td>50.3</td>
<td>3.8</td>
<td>27.7</td>
<td>74.7</td>
<td>2.1</td>
<td>43.2</td>
<td>41.3</td>
</tr>
<tr>
<td>DeFILIP</td>
<td>15M</td>
<td>70.1</td>
<td>46.8</td>
<td>54.5</td>
<td>40.3</td>
<td>63.7</td>
<td>52.4</td>
<td>4.6</td>
<td>30.2</td>
<td>75.0</td>
<td>3.3</td>
<td>45.0</td>
<td>44.2</td>
</tr>
<tr>
<td>HiDeCLIP</td>
<td>15M</td>
<td>65.1</td>
<td>39.4</td>
<td>56.3</td>
<td>43.6</td>
<td>64.1</td>
<td>55.4</td>
<td>5.4</td>
<td>34.0</td>
<td>77.0</td>
<td>4.6</td>
<td>45.9</td>
<td>44.6 (+3.3)</td>
</tr>
<tr>
<td>CLIP</td>
<td>30M</td>
<td>77.3</td>
<td>48.1</td>
<td>59.1</td>
<td>58.5</td>
<td>58.2</td>
<td>52.6</td>
<td>17.7</td>
<td>28.0</td>
<td>80.8</td>
<td>3.2</td>
<td>48.8</td>
<td>48.4</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>30M</td>
<td>77.6</td>
<td>56.2</td>
<td>63.9</td>
<td>65.6</td>
<td>62.5</td>
<td>60.7</td>
<td>22.2</td>
<td>38.0</td>
<td>82.4</td>
<td>5.5</td>
<td>52.9</td>
<td>53.4 (+5.0)</td>
</tr>
<tr>
<td>DeCLIP</td>
<td>30M</td>
<td>84.0</td>
<td>57.1</td>
<td>67.3</td>
<td>71.7</td>
<td>65.0</td>
<td>62.5</td>
<td>23.0</td>
<td>39.5</td>
<td>86.1</td>
<td>5.3</td>
<td>55.3</td>
<td>56.1</td>
</tr>
<tr>
<td>HiDeCLIP</td>
<td>30M</td>
<td>80.4</td>
<td>54.2</td>
<td>68.9</td>
<td>73.5</td>
<td>66.1</td>
<td>65.2</td>
<td>26.8</td>
<td>44.2</td>
<td>87.8</td>
<td>7.2</td>
<td>56.9</td>
<td>57.4 (+1.3)</td>
</tr>
</tbody>
</table>
## Results on Downstream Tasks

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Text Retrieval</th>
<th></th>
<th>Image Retrieval</th>
<th></th>
<th>RSUM</th>
<th></th>
<th>VQA (test-dev)</th>
<th></th>
<th>SNLI (val+test)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>Y/N</td>
<td>Num.</td>
<td>Other</td>
<td>All</td>
</tr>
<tr>
<td>CLIP</td>
<td>15M</td>
<td>21.4</td>
<td>44.7</td>
<td>56.4</td>
<td>13.7</td>
<td>32.4</td>
<td>42.9</td>
<td>211.5</td>
<td>67.3</td>
<td>30.5</td>
<td>32.7</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>15M</td>
<td>34.2</td>
<td>60.3</td>
<td>70.9</td>
<td>20.6</td>
<td>43.8</td>
<td>55.3</td>
<td><strong>285.1</strong></td>
<td>69.4</td>
<td>33.7</td>
<td>37.2</td>
</tr>
<tr>
<td>DeCLIP</td>
<td>15M</td>
<td>29.1</td>
<td>55.2</td>
<td>66.6</td>
<td>19.0</td>
<td>41.2</td>
<td>53.1</td>
<td>264.2</td>
<td>70.3</td>
<td>34.9</td>
<td>36.9</td>
</tr>
<tr>
<td>HiDeCLIP</td>
<td>15M</td>
<td>38.7</td>
<td>64.4</td>
<td>74.8</td>
<td>23.9</td>
<td>48.2</td>
<td>60.1</td>
<td><strong>310.1</strong></td>
<td>72.4</td>
<td>36.1</td>
<td>40.9</td>
</tr>
<tr>
<td>CLIP</td>
<td>30M</td>
<td>34.8</td>
<td>63.3</td>
<td>73.9</td>
<td>23.3</td>
<td>46.9</td>
<td>58.6</td>
<td>300.8</td>
<td>69.7</td>
<td>34.8</td>
<td>37.8</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>30M</td>
<td>43.9</td>
<td>69.1</td>
<td>78.8</td>
<td>27.0</td>
<td>51.8</td>
<td>62.9</td>
<td><strong>333.5</strong></td>
<td>72.2</td>
<td>36.1</td>
<td>40.9</td>
</tr>
<tr>
<td>DeCLIP</td>
<td>30M</td>
<td>41.3</td>
<td>68.8</td>
<td>79.3</td>
<td>25.6</td>
<td>50.7</td>
<td>62.3</td>
<td>328.0</td>
<td>71.3</td>
<td>35.4</td>
<td>39.7</td>
</tr>
<tr>
<td>HiDeCLIP</td>
<td>30M</td>
<td>48.6</td>
<td>74.1</td>
<td>82.7</td>
<td>29.6</td>
<td>54.9</td>
<td>66.3</td>
<td><strong>356.2</strong></td>
<td>73.3</td>
<td>37.0</td>
<td>42.5</td>
</tr>
</tbody>
</table>
## Linear Probe Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>C10</th>
<th>C100</th>
<th>F101</th>
<th>Pets</th>
<th>Flow.</th>
<th>SUN</th>
<th>Cars</th>
<th>DTD</th>
<th>Cal.</th>
<th>Air.</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP-ViT-B/32</td>
<td>15M</td>
<td>86.5</td>
<td>64.7</td>
<td>69.2</td>
<td>64.6</td>
<td>90.6</td>
<td>66.0</td>
<td>24.9</td>
<td>61.3</td>
<td>79.1</td>
<td>23.1</td>
<td>63.0</td>
</tr>
<tr>
<td>HiCLIP-ViT-B/32</td>
<td>15M</td>
<td>89.5</td>
<td>71.1</td>
<td>73.5</td>
<td>70.6</td>
<td>91.9</td>
<td>68.8</td>
<td>30.8</td>
<td>63.9</td>
<td>84.8</td>
<td>27.4</td>
<td><strong>67.2 (+4.2)</strong></td>
</tr>
<tr>
<td>DeCLIP-ViT-B/32</td>
<td>15M</td>
<td>89.2</td>
<td>69.0</td>
<td>75.4</td>
<td>72.2</td>
<td>94.4</td>
<td>71.6</td>
<td>31.0</td>
<td>68.8</td>
<td>87.9</td>
<td>27.6</td>
<td>68.7</td>
</tr>
<tr>
<td>HiDeCLIP-ViT-B/32</td>
<td>15M</td>
<td>88.1</td>
<td>70.7</td>
<td>77.6</td>
<td>75.5</td>
<td>95.6</td>
<td>72.2</td>
<td>36.0</td>
<td>70.1</td>
<td>90.0</td>
<td>32.6</td>
<td><strong>70.8 (+2.1)</strong></td>
</tr>
<tr>
<td>CLIP-ViT-B/16</td>
<td>15M</td>
<td>88.5</td>
<td>66.4</td>
<td>77.2</td>
<td>69.3</td>
<td>94.1</td>
<td>69.8</td>
<td>29.0</td>
<td>65.2</td>
<td>82.4</td>
<td>25.5</td>
<td>66.7</td>
</tr>
<tr>
<td>HiCLIP-ViT-B/16</td>
<td>15M</td>
<td>89.1</td>
<td>70.4</td>
<td>81.0</td>
<td>75.3</td>
<td>95.2</td>
<td>72.5</td>
<td>36.4</td>
<td>68.7</td>
<td>86.4</td>
<td>32.3</td>
<td><strong>70.7 (+4.0)</strong></td>
</tr>
<tr>
<td>DeCLIP-ViT-B/16</td>
<td>15M</td>
<td>88.7</td>
<td>69.5</td>
<td>83.0</td>
<td>74.3</td>
<td>97.3</td>
<td>74.4</td>
<td>36.9</td>
<td>70.9</td>
<td>89.8</td>
<td>32.2</td>
<td>71.7</td>
</tr>
<tr>
<td>HiDeCLIP-ViT-B/16</td>
<td>15M</td>
<td>88.8</td>
<td>70.3</td>
<td>84.3</td>
<td>80.6</td>
<td>97.1</td>
<td>75.1</td>
<td>42.5</td>
<td>74.3</td>
<td>90.7</td>
<td>38.3</td>
<td><strong>74.2 (+2.5)</strong></td>
</tr>
<tr>
<td>CLIP-ViT-B/32</td>
<td>30M</td>
<td>92.0</td>
<td>74.7</td>
<td>78.8</td>
<td>80.7</td>
<td>93.7</td>
<td>72.6</td>
<td>55.9</td>
<td>71.4</td>
<td>88.6</td>
<td>29.7</td>
<td>73.8</td>
</tr>
<tr>
<td>HiCLIP-ViT-B/32</td>
<td>30M</td>
<td>92.8</td>
<td>75.8</td>
<td>80.5</td>
<td>81.3</td>
<td>94.4</td>
<td>73.6</td>
<td>59.4</td>
<td>72.2</td>
<td>90.3</td>
<td>33.6</td>
<td><strong>75.4 (+1.6)</strong></td>
</tr>
<tr>
<td>DeCLIP-ViT-B/32</td>
<td>30M</td>
<td>93.1</td>
<td>76.9</td>
<td>82.0</td>
<td>82.7</td>
<td>96.0</td>
<td>74.9</td>
<td>59.8</td>
<td>74.5</td>
<td>92.6</td>
<td>32.7</td>
<td>76.5</td>
</tr>
<tr>
<td>HiDeCLIP-ViT-B/32</td>
<td>30M</td>
<td>92.7</td>
<td>75.6</td>
<td>82.9</td>
<td>83.3</td>
<td>95.7</td>
<td>75.6</td>
<td>62.8</td>
<td>74.5</td>
<td>92.0</td>
<td>35.8</td>
<td><strong>77.1 (+0.6)</strong></td>
</tr>
</tbody>
</table>
# Ablation Studies

Influence of patch granularity and dataset scale

<table>
<thead>
<tr>
<th>Method</th>
<th>Encoder</th>
<th>Data</th>
<th>ImageNet Acc.</th>
<th>11 Datasets Avg.</th>
<th>COCO Rsum</th>
<th>VQA Acc.</th>
<th>SNLI Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>ViT-B/32</td>
<td>15M</td>
<td>32.8</td>
<td>31.8</td>
<td>211.5</td>
<td>46.7</td>
<td>62.5</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>ViT-B/32</td>
<td>15M</td>
<td>40.5</td>
<td>41.8</td>
<td>285.1</td>
<td>50.1</td>
<td>67.7</td>
</tr>
<tr>
<td>DeCLIP</td>
<td>ViT-B/32</td>
<td>15M</td>
<td>43.2</td>
<td>41.3</td>
<td>264.2</td>
<td>50.4</td>
<td>66.1</td>
</tr>
<tr>
<td>HiDeCLIP</td>
<td>ViT-B/32</td>
<td>15M</td>
<td>45.9</td>
<td>44.6</td>
<td>310.1</td>
<td>53.3</td>
<td>70.5</td>
</tr>
<tr>
<td>CLIP</td>
<td>ViT-B/16</td>
<td>15M</td>
<td>39.3</td>
<td>35.5</td>
<td>245.0</td>
<td>48.8</td>
<td>63.8</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>ViT-B/16</td>
<td>15M</td>
<td>45.2</td>
<td>44.9</td>
<td>313.9</td>
<td>51.2</td>
<td>69.0</td>
</tr>
<tr>
<td>DeCLIP</td>
<td>ViT-B/16</td>
<td>15M</td>
<td>48.2</td>
<td>43.7</td>
<td>290.3</td>
<td>51.5</td>
<td>67.3</td>
</tr>
<tr>
<td>HiDeCLIP</td>
<td>ViT-B/16</td>
<td>15M</td>
<td>51.1</td>
<td>48.3</td>
<td>339.6</td>
<td>54.4</td>
<td>71.3</td>
</tr>
<tr>
<td>CLIP</td>
<td>ViT-B/32</td>
<td>30M</td>
<td>48.8</td>
<td>48.4</td>
<td>300.8</td>
<td>50.6</td>
<td>66.9</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>ViT-B/32</td>
<td>30M</td>
<td>52.9</td>
<td>53.4</td>
<td>333.5</td>
<td>53.2</td>
<td>70.1</td>
</tr>
<tr>
<td>DeCLIP</td>
<td>ViT-B/32</td>
<td>30M</td>
<td>55.3</td>
<td>56.1</td>
<td>328.0</td>
<td>52.2</td>
<td>69.0</td>
</tr>
<tr>
<td>HiDeCLIP</td>
<td>ViT-B/32</td>
<td>30M</td>
<td>56.9</td>
<td>57.4</td>
<td>356.2</td>
<td>54.6</td>
<td>72.5</td>
</tr>
</tbody>
</table>
## Ablation Studies

### Use of G-Trans and T-trans

<table>
<thead>
<tr>
<th>Method</th>
<th>Encoder</th>
<th>G-Trans</th>
<th>T-Trans</th>
<th>ImageNet Acc.</th>
<th>11 Datasets Avg.</th>
<th>Text Retrieval</th>
<th>Image Retrieval</th>
<th>COCO Rsum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
</tr>
<tr>
<td>CLIP</td>
<td>ViT-B/32</td>
<td>-</td>
<td>-</td>
<td>32.8</td>
<td>31.8</td>
<td>21.4</td>
<td>44.7</td>
<td>56.4</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>ViT-B/32</td>
<td>-</td>
<td>✓</td>
<td>37.1</td>
<td>38.4</td>
<td>28.7</td>
<td>53.8</td>
<td>65.7</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>ViT-B/32</td>
<td>✓</td>
<td>-</td>
<td>36.2</td>
<td>35.3</td>
<td>22.9</td>
<td>47.5</td>
<td>59.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td><strong>40.5</strong></td>
<td><strong>41.8</strong></td>
<td>34.2</td>
<td>60.3</td>
<td>70.9</td>
</tr>
<tr>
<td>CLIP</td>
<td>ViT-B/16</td>
<td>-</td>
<td>-</td>
<td>39.3</td>
<td>35.5</td>
<td>26.1</td>
<td>52.0</td>
<td>64.6</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>ViT-B/16</td>
<td>-</td>
<td>✓</td>
<td>40.4</td>
<td>39.6</td>
<td>32.8</td>
<td>58.5</td>
<td>69.7</td>
</tr>
<tr>
<td>HiCLIP</td>
<td>ViT-B/16</td>
<td>✓</td>
<td>-</td>
<td>42.2</td>
<td>37.7</td>
<td>28.5</td>
<td>53.2</td>
<td>65.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td><strong>45.2</strong></td>
<td><strong>44.9</strong></td>
<td>39.0</td>
<td>65.7</td>
<td>76.4</td>
</tr>
</tbody>
</table>
Visualization Results

Visualization of learned feature space via t-SNE

YFCC-15M
Visualization Results

Visualization of learned feature space via t-SNE

30M
Limitations

- Vague architecture specification
  - Number of hierarchical attention layers/encoder blocks in Group Transformer not mentioned
- Implementation is unavailable
- Computing C adds more computation
Conclusion

- Hierarchy-aware attention in CLIP > increase performance (zero-shot classification, VQA, VE)
- Better performance at the cost of more computation

Future plans:

- Scale visual encoder to validate scalability.
- Explore other methods to introduce hierarchy other than dot product.
- Explore multimodal fusion/cross attention between image and text.
- Scale up pretraining dataset.
Thank you