On Composition of Transformations for Self-Supervised Learning

Mandela Patrick*, Yuki M. Asano*, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, Andrea Vedaldi

ICCV 2021

* equal contribution
Self-supervision: a pretext zoo

Images:
*Jigsaw, RotNet, Colorisation*

Videos: two more dimensions
(time and modality)
*Shuffle&Learn, L3, AVTS*

[Noroozi and Favaro, ECCV 2016; Gidaris et al., ICLR 2018; Zhang et al., ECCV 2016]
[Misra et al., ECCV 2016; Arandjelovic and Zisserman, ICCV 2017; Korbar et al., NeurIPS 2018]
Noise contrastive learning

\[ f(\text{cat}) = f(\text{cat}) \neq f(\text{dog}) \]

Key idea:
*features should encode image’s core information.*
*learn this by comparing augmentations against other images.*

Examples: NPID, MoCo, CMC, SimCLR..

[Wu et al., CVPR 2018; He et al., CVPR 2020; Tian et al., ECCV 2020; Chen et al., ICML 2020]
Most pretext tasks

\[ f(\_\_) = \text{Clockwise 90 degrees rotation} \]

e.g. RotNet

Pretext tasks = *Specifying* invariance and distinctiveness to *selected* transformations.

[Gidaris et al., ICLR 2018]
Invariance vs distinctiveness

Should the representation discount or distinguish time reversal?
GDT: Generalised data transformations

GDTs express all aspects of these methods as transformations.

Allow to specify either invariance or distinctiveness
GDT Transformations

In our work, we explore the following transformations as learning hypotheses:

1. Sample Distinctiveness
2. Time Reversal
3. Time Shift

...cross-modally.
Gains from hypotheses

Timeshift (TS)

<table>
<thead>
<tr>
<th>HMDB-51</th>
<th>Base</th>
<th>Inv (I)</th>
<th>Dist (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.4</td>
<td>58.7</td>
<td>52.4</td>
<td></td>
</tr>
</tbody>
</table>

Time reversal (TR)

<table>
<thead>
<tr>
<th>HMDB-51</th>
<th>Base</th>
<th>Dist (D)</th>
<th>Inv (I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.8</td>
<td>58.8</td>
<td>52.4</td>
<td></td>
</tr>
</tbody>
</table>

Combinations

<table>
<thead>
<tr>
<th>HMDB-51</th>
<th>Base</th>
<th>TS-D</th>
<th>TR-I</th>
<th>TS-D, TR-I</th>
</tr>
</thead>
<tbody>
<tr>
<td>58.7</td>
<td>58.8</td>
<td>61.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SOTA video action retrieval and few-shot learning results

Retrieval @1: UCF-101

- SpeedNet: 13
- VSP: 24.6
- SeLaVi: 52
- CoCLR: 53.3
- GDT: 62.8

1-Shot learning @UCF-101

- Random: 2.3
- 3DRotNet: 15
- GDT: 26.7

[Benaim et al, CVPR 2020; Cho et al., ArXiv; Asano et al., NeurIPS 2020; Han et al., NeurIPS 2020]

[Jing and Tian, arXiv]
SOTA fine-tuning video-action recognition results

HMDB-51

<table>
<thead>
<tr>
<th>Method</th>
<th>AVTS</th>
<th>AVID</th>
<th>MMV</th>
<th>XDC</th>
<th>GDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>61.6</td>
<td>64.7</td>
<td>68.3</td>
<td>68.9</td>
<td>72.8</td>
</tr>
</tbody>
</table>

UCF-101

<table>
<thead>
<tr>
<th>Method</th>
<th>AVTS</th>
<th>MMV</th>
<th>AVID</th>
<th>XDC</th>
<th>GDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>89</td>
<td>91.1</td>
<td>91.5</td>
<td>95.5</td>
<td>95.2</td>
</tr>
</tbody>
</table>

[Bruno et al., NeurIPS 2018; Morgado et al., CVPR 2021; Alayrac et al., NeurIPS 2020; Alwassel et al., NeurIPS 2020]