An Empirical Study of End-to-End Video-Language Transformers with Masked Visual Modeling

Tsu-Jui Fu¹
Linjie Li²
Zhe Gan³
Kevin Lin²
William Wang¹
Lijuan Wang²
Zicheng Liu²

¹UC Santa Barbara, ²Microsoft, ³Apple
Large-scale Text-Visual Pre-training

- Masked Language Modeling (MLM): recover missing word tokens
- Visual-Text Matching (VTM): alignment between visual and textual inputs

- How to enhance the visual modality?
Mask Visual Modeling (MVM)

- MVM achieves promising results for self-supervised visual pre-training
  - MAE, BEiT, VideoMAE, ...
- In contrast, MVM even hurts performance on text-image pre-training
- How can we design effective MVM for **text-video pre-training**?

---

**Visual Pre-training**

**Text-Image Pre-training**

**Text-Video Pre-training**

---

[CVPR’22] Masked Autoencoders Are Scalable Vision Learners
Diverse Targets of MVM

- Explore various MVM targets for end-to-end VidL learning
  - **Low-level**: Pixel, HOG
  - **Semantic-level**: Depth, Flow, SIF, TVF
  - **Multi-modal**: VQ, MMF
MVM on Text-Video (WebVid-2.5M)

- **Not all MVMs** are helpful for VidL
- Only **Pixel** and **SIF** bring consistent improvement on both downstream tasks
- **SIF** gains significant advance, especially on T2V

<table>
<thead>
<tr>
<th>Pre-train</th>
<th>MVM</th>
<th>TGIF-Frame</th>
<th>DiDeMo-Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>R1</td>
</tr>
<tr>
<td>VTM+MLM</td>
<td>None</td>
<td>68.1</td>
<td>28.7</td>
</tr>
<tr>
<td>+MVM</td>
<td>Pixel</td>
<td>68.3 (+0.2)</td>
<td>29.2 (+0.5)</td>
</tr>
<tr>
<td></td>
<td>HOG</td>
<td>67.3 (-0.8)</td>
<td>26.6 (-2.1)</td>
</tr>
<tr>
<td></td>
<td>Depth</td>
<td>68.0 (-0.1)</td>
<td>27.3 (-1.4)</td>
</tr>
<tr>
<td></td>
<td>Flow</td>
<td>67.6 (-0.5)</td>
<td>30.3 (+1.6)</td>
</tr>
<tr>
<td></td>
<td>SIF</td>
<td><strong>68.8 (+0.7)</strong></td>
<td><strong>35.4 (+6.7)</strong></td>
</tr>
<tr>
<td></td>
<td>TVF</td>
<td>68.0 (-0.1)</td>
<td>32.8 (+4.1)</td>
</tr>
<tr>
<td></td>
<td>VQ</td>
<td>68.4 (+0.3)</td>
<td>28.1 (-0.6)</td>
</tr>
<tr>
<td></td>
<td>MMF</td>
<td>67.7 (-0.4)</td>
<td>29.8 (+1.1)</td>
</tr>
</tbody>
</table>
## Combination of MVM targets on Text-Video

- Joint of different MVMs is **not encouraging**
- Explicit Pixel **conflicts with** high-level SIF
- SIF+TVF cannot bring more improvement (T2V ↓)

<table>
<thead>
<tr>
<th>MVM</th>
<th>TGIF-Frame</th>
<th>DiDeMo-Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>R1</td>
</tr>
<tr>
<td>None</td>
<td>68.1</td>
<td>28.7</td>
</tr>
<tr>
<td>Pixel</td>
<td>68.3 (+0.2)</td>
<td>29.2 (+0.5)</td>
</tr>
<tr>
<td>Flow</td>
<td>67.6 (-0.5)</td>
<td>30.3 (+1.6)</td>
</tr>
<tr>
<td>SIF</td>
<td>68.8 (+0.7)</td>
<td><strong>35.4 (+6.7)</strong></td>
</tr>
<tr>
<td>TVF</td>
<td>68.0 (-0.1)</td>
<td>32.8 (+4.1)</td>
</tr>
<tr>
<td>SIF+Pixel</td>
<td>68.8 (+0.7)</td>
<td>31.8 (+3.1)</td>
</tr>
<tr>
<td>SIF+Flow</td>
<td>68.7 (+0.6)</td>
<td>34.4 (+5.7)</td>
</tr>
<tr>
<td>SIF+TVF</td>
<td><strong>69.2 (+1.1)</strong></td>
<td>33.8 (+5.1)</td>
</tr>
</tbody>
</table>
MVM on Text-Image (CC3M)

- **Challenging to learn** without visual implications from neighbor frames
- **Fit in static image**, which hurts video temporal
- MVM cannot work well on text-image data for VidL

<table>
<thead>
<tr>
<th>Pre-train</th>
<th>MVM</th>
<th>TGIF-Frame</th>
<th></th>
<th>DiDeMo-Retrieval</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>R1</td>
<td>R5</td>
<td>R10</td>
</tr>
<tr>
<td>VTM+MLM</td>
<td>None</td>
<td>69.8</td>
<td>36.4</td>
<td>64.3</td>
<td>74.7</td>
</tr>
<tr>
<td></td>
<td>Pixel</td>
<td>69.7 (-0.1)</td>
<td>35.8 (-0.6)</td>
<td><strong>64.4 (+0.1)</strong></td>
<td>74.9 (+0.2)</td>
</tr>
<tr>
<td></td>
<td>HOG</td>
<td>69.8</td>
<td>34.9 (-1.5)</td>
<td>64.4 (+0.1)</td>
<td>75.1 (+0.4)</td>
</tr>
<tr>
<td>+MVM</td>
<td>Depth</td>
<td>69.6 (-0.2)</td>
<td>32.3 (-4.1)</td>
<td>63.8 (-0.5)</td>
<td>74.2 (-0.5)</td>
</tr>
<tr>
<td></td>
<td>SIF</td>
<td>69.7 (-0.1)</td>
<td>31.6 (-4.8)</td>
<td>60.5 (-3.8)</td>
<td>72.5 (-2.2)</td>
</tr>
<tr>
<td></td>
<td>VQ</td>
<td>69.8</td>
<td>34.4 (-2.0)</td>
<td>62.6 (-1.7)</td>
<td>75.1 (+0.4)</td>
</tr>
<tr>
<td></td>
<td>MMF</td>
<td>69.8</td>
<td>33.6 (-2.8)</td>
<td>62.9 (-1.4)</td>
<td><strong>75.6 (+0.9)</strong></td>
</tr>
</tbody>
</table>
MVM on Text-Image & Text-Video

- Not trivial to find superior MVM combination
- **Video (SIF) + Image (None)** is our default setting

<table>
<thead>
<tr>
<th>Pre-train</th>
<th>MVM</th>
<th>TGIF-Frame</th>
<th>DiDeMo-Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WebVid</td>
<td>CC3M</td>
<td>Accuracy</td>
</tr>
<tr>
<td>VTM+MLM</td>
<td>None</td>
<td></td>
<td>69.7</td>
</tr>
<tr>
<td>+MVM</td>
<td>SIF</td>
<td>None</td>
<td>71.1 (+1.4)</td>
</tr>
<tr>
<td></td>
<td>SIF</td>
<td>Pixel</td>
<td>71.3 (+1.6)</td>
</tr>
</tbody>
</table>
SIF Extractor vs. Downstream

- Classification accuracy is crucial but **not positively correlated**
- **Similar inductive biases** is another key
- Trade-off between **informative and feasible** learning

<table>
<thead>
<tr>
<th></th>
<th>SIF</th>
<th>IN-1K</th>
<th>TGIF-Frame</th>
<th>DiDeMo-Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Train</td>
<td>Accuracy</td>
<td>Accuracy</td>
<td>R1</td>
</tr>
<tr>
<td>None</td>
<td>None</td>
<td>68.1</td>
<td>28.7</td>
<td>57.0</td>
</tr>
<tr>
<td>Res-50</td>
<td>IN-1K</td>
<td>76.1</td>
<td>67.3 (-0.8)</td>
<td>29.1 (+0.4)</td>
</tr>
<tr>
<td>Swin-T</td>
<td>IN-1K</td>
<td>81.2</td>
<td><strong>68.9 (+0.8)</strong></td>
<td>33.8 (+5.1)</td>
</tr>
<tr>
<td>DeiT</td>
<td>IN-1K</td>
<td>83.4</td>
<td>68.4 (+0.3)</td>
<td>31.4 (+2.7)</td>
</tr>
<tr>
<td>Swin-B</td>
<td>IN-1K</td>
<td>83.5</td>
<td>68.3 (+0.2)</td>
<td>34.9 (+6.2)</td>
</tr>
<tr>
<td>Swin-B</td>
<td>IN-22K</td>
<td>85.2</td>
<td>68.8 (+0.7)</td>
<td><strong>35.4 (+6.7)</strong></td>
</tr>
<tr>
<td>Swin-L</td>
<td>IN-22K</td>
<td><strong>86.3</strong></td>
<td>68.2 (+0.1)</td>
<td>33.2 (+4.5)</td>
</tr>
</tbody>
</table>
Comparison with SOTA

- Video Question Answering (VideoQA)

<table>
<thead>
<tr>
<th>Method</th>
<th>#Pre-train</th>
<th>TGIF</th>
<th>MSRVT</th>
<th>LSMDC</th>
<th>MSVD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Act.</td>
<td>Trans.</td>
<td>Frame</td>
<td>MC</td>
</tr>
<tr>
<td>ClipBERT</td>
<td>0.2M</td>
<td>82.8</td>
<td>87.8</td>
<td>60.3</td>
<td>88.2</td>
</tr>
<tr>
<td>ALRPO</td>
<td>5M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>JustAsk</td>
<td>69M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MERLOT</td>
<td>180M</td>
<td>94.0</td>
<td>96.2</td>
<td>69.5</td>
<td>90.9</td>
</tr>
<tr>
<td>VIOLET</td>
<td>186M</td>
<td>92.5</td>
<td>95.7</td>
<td>68.9</td>
<td>91.9</td>
</tr>
<tr>
<td>All-in-One</td>
<td>283M</td>
<td>95.5</td>
<td>94.7</td>
<td>66.3</td>
<td>92.3</td>
</tr>
<tr>
<td>VIOLETv2</td>
<td>5M</td>
<td>94.8</td>
<td>99.0</td>
<td>72.8</td>
<td>97.6</td>
</tr>
</tbody>
</table>
Comparison with SOTA

- Text-to-Video Retrieval (T2V)

<table>
<thead>
<tr>
<th>Method</th>
<th>#Pre-train</th>
<th>MARVTT</th>
<th></th>
<th>DiDeMo</th>
<th></th>
<th>LSMDC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R1</td>
<td>R5</td>
<td>R10</td>
<td>R1</td>
<td>R5</td>
<td>R10</td>
</tr>
<tr>
<td>ClipBERT</td>
<td>0.2M</td>
<td>22.0</td>
<td>46.8</td>
<td>59.9</td>
<td>20.4</td>
<td>48.0</td>
<td>60.8</td>
</tr>
<tr>
<td>Frozen</td>
<td>5M</td>
<td>31.0</td>
<td>59.5</td>
<td>70.5</td>
<td>31.0</td>
<td>59.8</td>
<td>72.8</td>
</tr>
<tr>
<td>ALPRO</td>
<td>5M</td>
<td>33.9</td>
<td>60.7</td>
<td>73.2</td>
<td>35.9</td>
<td>67.5</td>
<td>78.8</td>
</tr>
<tr>
<td>B-Former</td>
<td>5M</td>
<td><strong>37.6</strong></td>
<td>64.8</td>
<td>75.1</td>
<td>37.0</td>
<td>62.2</td>
<td>73.9</td>
</tr>
<tr>
<td>All-in-One</td>
<td>138M</td>
<td>37.9</td>
<td>68.1</td>
<td>77.1</td>
<td>32.7</td>
<td>61.4</td>
<td>73.5</td>
</tr>
<tr>
<td>VIOLET</td>
<td>186M</td>
<td>34.5</td>
<td>63.0</td>
<td>73.4</td>
<td>32.6</td>
<td>62.8</td>
<td>74.7</td>
</tr>
<tr>
<td>Clip4Clip</td>
<td>400M</td>
<td>42.1</td>
<td>71.9</td>
<td>81.4</td>
<td>43.4</td>
<td>70.2</td>
<td>80.6</td>
</tr>
<tr>
<td>VIOLETv2</td>
<td>5M</td>
<td>37.2</td>
<td><strong>64.8</strong></td>
<td><strong>75.8</strong></td>
<td><strong>47.9</strong></td>
<td><strong>76.5</strong></td>
<td><strong>84.1</strong></td>
</tr>
</tbody>
</table>
Summary

- Explore various MVM targets for VidL learning
  - Low-level: **Pixel**, HOG
  - Semantic-level: Depth, Flow, **SIF**, TVF
  - Multi-modal: VQ, MMF

- Best setting should be **Text-Video (SIF) + Text-Image (None)**
  - Not trivial to find superior combination of MVM

- Features extractor is also crucial
  - Classification accuracy is **not always positively correlated**
  - **Similar inductive biases** is the key
  - Trade-off between **informative and feasible** learning