Two-stage Ranking Strategy for Ad-hoc Video Search

Fangming Zhou\textsuperscript{1}, Changqiao Wu\textsuperscript{2}, Xiaofeng Guo\textsuperscript{2}, Haofan Wang\textsuperscript{2}, Jincan Deng\textsuperscript{2}, Debing Zhang\textsuperscript{2}

\textsuperscript{1}Renmin University of China
\textsuperscript{2}MMU, Kuaishou Technology

TRECVID 2021 workshop
2021-12
Ad-hoc Video Search

• **ESSENCE**: text-video matching in retrieval scenario (a hot topic)

Previous works

• RUC_AIMC3 at TRECVID 2020 [1]:
  • Two-branch model
  • Addition of irCSN feature

• RUCMM at TRECVID 2020 [2]:
  • Multi-space & multi-loss strategy
  • Addition of C3d feature

Two-stage Ranking Strategy

• **Stage - I : Keyword-based Rank**
  - Model architecture
  - Training data
  - Visual features and textual encoders

• **Stage - II : Fine-grained Re-rank**
  - Frame-level matching
  - Weighted sum of two stages as final similarity
  - Reasonableness of re-ranking
Two-stage Ranking Strategy

- **Stage - I**: Keyword-based Rank
- **Model architecture**
  - SEA: sentence encoder assembly
  - Multi-space architecture
  - Learning $k$ common space for $k$ sentence encoders
  - Combined loss:
    $$loss = \sum_{i=1}^{k} loss_i(\text{sentence}, \text{video})$$

[1] Li et al., SEA: Sentence encoder assembly for video retrieval by textual queries. TMM, 2021.
Two-stage Ranking Strategy

• Stage - I: Keyword-based Rank
  • Model architecture

SEA -> SEA++

![Diagram showing the SEA and SEA++ models with encoders and feature vectors.](Image)
Two-stage Ranking Strategy

- Stage - I : Keyword-based Rank
  - Model architecture - SEA++ model
    - ✓ Individual common space for each combination.
    - ✓ First stage similarity:
      \[ S_{first} = \sum_{i=1}^{m} \sum_{j=1}^{n} S_{space, i, j} \]
    - ✓ Only the first \( K \) videos sorted according to \( S_{first} \) will be passed to the Stage - II.
Two-stage Ranking Strategy

- Stage - I: Keyword-based Rank
  - Model architecture - SEA++ model
Two-stage Ranking Strategy

- Stage - I : Keyword-based Rank
  - Training data
    ✓ Concepts in different datasets should be complementary.
    ✓ Training data should be similar to V3C1.

<table>
<thead>
<tr>
<th>Training data</th>
<th>Num of video/image</th>
<th>Num of sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSR-VTT</td>
<td>10k</td>
<td>200k</td>
</tr>
<tr>
<td>TGIF</td>
<td>100k</td>
<td>124k</td>
</tr>
<tr>
<td>VATEX</td>
<td>32k</td>
<td>349k</td>
</tr>
<tr>
<td>MSCOCO</td>
<td>123k</td>
<td>616k</td>
</tr>
</tbody>
</table>
Two-stage Ranking Strategy

- Stage - I: Keyword-based Rank
- Visual features and textual encoders
  - Visual features: ResNeXt101, irCSN, CLIP, timesformer
  - Textual encoders: Bag-of-word, word2vec (keyword-based)

<table>
<thead>
<tr>
<th>Model</th>
<th>Feature</th>
<th>TV19</th>
<th>TV20</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEA(BoW, w2v)</td>
<td>Resnext+irCSN</td>
<td>0.167</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>Resnext+irCSN+CLIP</td>
<td>0.185</td>
<td>0.327</td>
</tr>
<tr>
<td></td>
<td>Resnext+irCSN+CLIP+timesformer</td>
<td>0.191</td>
<td>0.332</td>
</tr>
</tbody>
</table>

Ablation experiment on visual feature
Two-stage Ranking Strategy

• Stage II: Fine-grained Re-rank
  • Frame-level matching
  • Weighted sum of two stages as final similarity
  • Reasonableness of re-ranking
Two-stage Ranking Strategy

- Stage - II: Fine-grained Re-rank
  - **Frame-level matching**
    - ✓ We use out-of-box CLIP [1] as frame-level matching model.
    - ✓ Only Top-K videos sorted in Stage I are considered.
    - ✓ Second stage similarity:
      \[
      S_{\text{second}} = \max(I_1 \cdot T, \ldots, I_n \cdot T)
      \]
      \[
      I_i = \text{ImageEncoder(frame}_i\text{)}
      \]
      \[
      T = \text{TextEncoder(query)}
      \]

Two-stage Ranking Strategy

• Stage - II: Fine-grained Re-rank
  • Weighted sum of two stages as final similarity
    \[ S(Q,V) = \begin{cases} 0, & \text{if } S_{first} < S_{threshold} \\ w_1 \cdot S_{first} + w_2 \cdot S_{second}, & \text{if } S_{first} \geq S_{threshold} \end{cases} \]

• Reasonableness of re-ranking
  ✓ Keyframe is enough in most cases.
  ✓ The sentence semantics is considered (versus previous keyword-based method).

<table>
<thead>
<tr>
<th>Models</th>
<th>TV19</th>
<th>TV20</th>
<th>TV21</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.211</td>
<td>0.362</td>
<td>0.340</td>
</tr>
<tr>
<td>Baseline + re-rank</td>
<td>0.241</td>
<td>0.360</td>
<td>0.349</td>
</tr>
</tbody>
</table>

[*] We set \( w_1 = 0.2 \) \( w_2 = 0.8 \) in our experiments.
Overall Pipeline

Stage-1: Keyword-based Rank
- Query Q
- Sentence
- Text Feature 1
- Text Feature M
- Video Feature 1
- Video Feature N
- Video V
- Video
- sum of similarities

Stage-2: Fine-grained Re-rank
- Text Encoder
- Image Encoder
- $T$: text encoding
- $I$: image encoding
- $S$: similarity
- $S_{first}$
- $S_{second}$
- Frame 1
- Frame 2
- Frame n
- max of similarities

Submitted Runs

- Our final submitted runs as followed:
  - run 3: single SEA++ model

<table>
<thead>
<tr>
<th>Submissions</th>
<th>TV19</th>
<th>TV20</th>
<th>TV21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winner in 2019</td>
<td>0.163</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Winner in 2020</td>
<td>0.359</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>run3</td>
<td>0.206</td>
<td>0.354</td>
<td>0.332</td>
</tr>
</tbody>
</table>
Submitted Runs

• Our final submitted runs as followed:
  • run 3: single SEA++ model
  • run 2: model ensemble

<table>
<thead>
<tr>
<th>Submissions</th>
<th>TV19</th>
<th>TV20</th>
<th>TV21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winner in 2019</td>
<td>0.163</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Winner in 2020</td>
<td>-</td>
<td>0.359</td>
<td>-</td>
</tr>
<tr>
<td>run3</td>
<td>0.206</td>
<td>0.354</td>
<td>0.332</td>
</tr>
<tr>
<td>run2</td>
<td>0.211</td>
<td><strong>0.362</strong></td>
<td>0.340</td>
</tr>
</tbody>
</table>

1,2 indicates different ensemble strategy
Submitted Runs

- Our final submitted runs as followed:
  - run 3: single SEA++ model
  - run 2: model ensemble\(^1\)
  - run 1: model ensemble\(^1\) + re-rank

<table>
<thead>
<tr>
<th>Submissions</th>
<th>TV19</th>
<th>TV20</th>
<th>TV21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winner in 2019</td>
<td>0.163</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Winner in 2020</td>
<td>-</td>
<td>0.359</td>
<td>-</td>
</tr>
<tr>
<td>run3</td>
<td>0.206</td>
<td>0.354</td>
<td>0.332</td>
</tr>
<tr>
<td>run2</td>
<td>0.211</td>
<td><strong>0.362</strong></td>
<td>0.340</td>
</tr>
<tr>
<td>run1(primary)</td>
<td><strong>0.241</strong></td>
<td>0.360</td>
<td><strong>0.349</strong></td>
</tr>
</tbody>
</table>

\(^1\) indicates different ensemble strategy
Submitted Runs

- Our final submitted runs as followed:
  - run 3: single SEA++ model
  - run 2: model ensemble
  - run 1: model ensemble + re-rank
  - run 4: model ensemble + re-rank

<table>
<thead>
<tr>
<th>Submissions</th>
<th>TV19</th>
<th>TV20</th>
<th>TV21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winner in 2019</td>
<td>0.163</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Winner in 2020</td>
<td>-</td>
<td>0.359</td>
<td>-</td>
</tr>
<tr>
<td>run3</td>
<td>0.206</td>
<td>0.354</td>
<td>0.332</td>
</tr>
<tr>
<td>run2</td>
<td>0.211</td>
<td><strong>0.362</strong></td>
<td>0.340</td>
</tr>
<tr>
<td>run1 (primary)</td>
<td><strong>0.241</strong></td>
<td>0.360</td>
<td><strong>0.349</strong></td>
</tr>
<tr>
<td>run4</td>
<td>0.239</td>
<td>0.358</td>
<td><strong>0.349</strong></td>
</tr>
</tbody>
</table>

1,2 indicates different ensemble strategy
Take-home Message

1) We propose an improved video retrieval model, namely SEA++, which built a solid backbone for our best run.
Take-home Message

1) We propose an improved video retrieval model, namely SEA++, which built a solid backbone for our best run.

2) Re-ranking by CLIP is an effective method to gain higher performance.
THANKS!

• Contact with us:
  fangming_zhou@ruc.edu.cn,
  wuchangqiao@kuaishou.com,
  zhangdebing@kuaishou.com