(Un)likelihood Learning for Interpretable Embedding

Jiaxin Wu\(^1\), Zhijian Hou\(^1\), Zhixin Ma\(^2\) and Chong-Wah Ngo\(^2\)
\(^1\)City University of Hong Kong
\(^2\)Singapore Management University

TRECVID 2021 Workshop
Outline

- Interpretable embedding and overlooked issue
- Unlikelihood learning for interpretable embedding
- Submitted runs and analysis
- Summary
Interpretable embedding (Dual-task model)

- Main idea: equip embedding search with interpretability.

Wu and Ngo, Interpretable Embedding for Ad-hoc Video Search, ACM MM, 2020
Overlooked Issue: Inconsistent Interpretation

• Contrary concepts are simultaneously decoded for visual embeddings
• Hurt representation learning and retrieval performances
How to generate consistent interpretation

• Two “supervisors” (Likelihood and Unlikelihood)

Ground truth:
A truck is being driven on a road.
A truck drives on a road at daytime.

Prior knowledge:
daytime | night
outdoors | indoors

Don’t do that!

Likelihood supervisor

Concept Decoder

Unlikelihood supervisor
Likelihood learning

• Goal: recover the concepts in the annotated label.
• Obstacles: sparse and incomplete label.
• Propose class-sensitive BCE loss.

\[
\text{Loss}_{\text{BCE}}(\hat{p}, p) = \lambda \frac{1}{\sum_i^n p_i} \sum_i^n p_i \text{BCE}(\hat{p}_i, p_i) + (1 - \lambda) \frac{1}{\sum_i^n (1 - p_i)} \sum_i^n (1 - p_i) \text{BCE}(\hat{p}_i, p_i),
\]

\[
\text{BCE}(\hat{p}_i, p_i) = -[p_i \log(\hat{p}_i) + (1 - p_i) \log(1 - \hat{p}_i)].
\]
Unlikelihood Learning (UL)

- Goal: suppress the probabilities of contradicting/exclusive concepts.
- Prior knowledge: WordNet antonym[1].
- Obstacles: Context, globally/locally exclusive.
- Propose new UL loss function inspired by [2,3].

\[
\text{Loss}_{UL}(\hat{p}, p) = \frac{1}{\sum_i p_i} \sum_i -p_i \sum_{t \in T} \log(1 - \hat{p}_i) \times (1 - p_t)
\]

[3] Roller et al., Don’t say that! making inconsistent dialogue unlikely with unlikelihood training, ACL, 2020
New architecture

• Embedding search, concept search and fusion search
Advantages of the new model

- Make query embedding less sensitive to query formulation
- Likelihood training can address the **missing labels** problem
- Unlikelihood training avoids **frequent** and **contradicting** concepts

Top-15 concepts in the query embedding interpretation:
[interpret, communicate, sign, signs, language, deaf, interpreter, asl, use, convey, translate, message, words, show, person]
<table>
<thead>
<tr>
<th>Submitted run</th>
<th>Model</th>
<th>Concept search</th>
<th>Embedding search</th>
<th>Fusion search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline #1</td>
<td>Original Dual-task model</td>
<td>0.167</td>
<td>0.167</td>
<td>0.193</td>
</tr>
<tr>
<td>Baseline #2</td>
<td>Feature enhancement dual-task model*</td>
<td>0.269</td>
<td>0.278</td>
<td>0.305</td>
</tr>
<tr>
<td>Baseline #3</td>
<td>Feature enhancement dual encoding model* [1]</td>
<td>/</td>
<td>0.287</td>
<td>/</td>
</tr>
<tr>
<td>RUN1</td>
<td>Phrase model*</td>
<td>0.216</td>
<td>0.301</td>
<td>0.317</td>
</tr>
<tr>
<td>RUN2</td>
<td>(Un)likelihood model*</td>
<td>0.270</td>
<td>0.290</td>
<td>0.330</td>
</tr>
<tr>
<td>RUN3</td>
<td>RUN1+RUN2</td>
<td>/</td>
<td>/</td>
<td>0.336</td>
</tr>
<tr>
<td>RUN4</td>
<td>RUN1+RUN2+Feature enhancement</td>
<td>/</td>
<td>/</td>
<td>0.355</td>
</tr>
<tr>
<td>Novelty run</td>
<td>Concept searches of RUN1 and RUN2+manual queries</td>
<td>0.297</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

*(+ video features[2,3] + VATEX dataset [4])

Benefit from the phrase vocabulary

- 676 Find shots of a **white dog**

(a) Dual-task\textsubscript{concept} (xinfAP=0.167)

(b) Phrase model\textsubscript{concept} (xinfAP=0.4252)
Benefit from the unlikelihood training

- 662 Find shots of a **woman** wearing sleeveless top

(a) Dual-task \textsubscript{embedding} (xinfAP=0.355)  

(b) UL model \textsubscript{embedding} (xinfAP=0.580)
Suffer from small number of training cases

- 678 Find shots of a man sitting on a barber chair in a shop

(a) UL model (xinfAP=0.409)  
(b) Phrase model_{concept} (xinfAP=0.133)
Automatic Versus Manual (Novelty) runs

- Automatic runs outperform manual runs.
- Manual (Novelty) runs are sensitive to query formulation.
Summary

• Enhanced features and additional dataset significantly improve the performance.

• (Un)likelihood model effectively pull down contradicted videos.

• With phrases, interpretable embeddings are more robust, but concept phrase retrieval rate could be limited by having a small number of training samples.

• Manual runs are sensitive to query formulation and the results are depend on the training data and the video dataset.
Thank you
Q&A