Human Effort in Vision and Learning

- Annotator: most successful models are trained on fully supervised datasets and require expensive human annotations;
- Modeler: a large number of models are iterated but there is a lack of efficient model storage and management.

1- Visual n-grams: Learning from Web Data [1]

Learning from photo-comment pairs to predict n-grams from input images:

- Naive loss: an negative log-likelihood (NLL) loss that classifies n-grams as independent labels; does not handle out-of-vocabulary;

\[
\ell(I, w; \theta, E) = -\sum_{k=1}^{K} \log p(w | w_{i-k}^{i-1}; \phi(I; \theta); E)
\]

where the likelihood of a word conditioned on the \((n-1)\) words appearing before it is defined as:

\[
p(w | w_{i-k}^{i-1}) = \lambda \rho_{obs}(w | w_{i-k}^{i-1}) + (1 - \lambda) p(w | w_{i-k}^{i-1}).
\]

1- Quantitative Results: Perplexity

<table>
<thead>
<tr>
<th>Loss / Smoothing</th>
<th>“Stupid” back-off</th>
<th>Jelinek-Mercer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet + linear</td>
<td>349</td>
<td>233</td>
</tr>
<tr>
<td>Naive n-gram</td>
<td>297</td>
<td>212</td>
</tr>
<tr>
<td>Jelinek-Mercer</td>
<td>276</td>
<td>199</td>
</tr>
</tbody>
</table>

Table: Perplexity of visual n-gram models averaged over YFCC100M test set of 10,000 images (evaluated on in-dictionary words only). Results for two losses (rows) with and without smoothing at test time (columns). Lower is better. See [4] for Stupid backoff.

1- Qualitative Results: Retriving Images from n-gram queries


2- Dataset & Dictionary

- YFCC100M: user photos and comments (English only ≈ 30M);
- Preprocess comments and choose n-grams occurring over 1,000 times;
- Dictionary contains 142,806 n-grams (n = 1, 2, 3, 4, 5).

2- Loss functions

- Naive loss: an negative log-likelihood (NLL) loss that classifies n-grams as independent labels; does not handle out-of-vocabulary;

\[
\ell(I, w; \theta, E) = -\sum_{k=1}^{K} \log p(w | w_{i-k}^{i-1}; \phi(I; \theta); E)
\]

where the likelihood of a word conditioned on the \((n-1)\) words appearing before it is defined as:

\[
p(w | w_{i-k}^{i-1}) = \lambda \rho_{obs}(w | w_{i-k}^{i-1}) + (1 - \lambda) p(w | w_{i-k}^{i-1}).
\]

2- Quantitative Results: Perplexity

3- Model Search & Discovery [3]

- A ModelHub system hosts 54 neural networks fine-tuned from VGG-16;
- Result table cosine distance: cosine distance between final responses;
- Aligned parameter distance: \(d^2\) distance between aligned weight matrices;
- Highly correlated: Pearson correlation coefficient \(r = 0.8224\).

Reference


Acknowledgement. This work was supported in part by NSF under grants 1513972 and 1513443, and in part by the Office of Naval Research under grant N00014161271 entitled “Visual Common Sense Reasoning for Multi-agent Activity Prediction and Recognition.”