For fine-grained feature learning from overhead imagery, we are investigating alternate techniques for training deep CNNs that yield results comparable to the famous Triplet Loss, but with a reduction in training time. The Triplet Loss is popular for the image retrieval task, but suffers from extensive sampling during training. To improve training time and possibly retrieval accuracy, we propose to use an alternate loss function, Weighted Approximate-Rank Pairwise Loss (WARP Loss), with some minor optimizations to improve runtime. We are also experimenting with cosine distance layers to better learn features, and improve image retrieval accuracy.

Image Retrieval

- For the image retrieval task, the goal is to provide a query image and return the most similar images.

- Two main types of models that learn features for the image retrieval task: Siamese Convolutional Neural Networks (CNNs) and CNNs trained with a Triplet Loss objective.

WARP Loss

- Focuses on learning using a pairwise ranking objective.
- Cross Entropy aggregates information across all classes.
- The WARP Loss Equation:

\[
L_{WARP}(x) = \frac{1}{N} \sum_{i=1}^{N} L_{triplet}(x_i, x_{i+1}, x_{i-1})
\]

where \( L_{triplet}(x_i, x_{i+1}, x_{i-1}) = \max \{d(a, p) - d(a, n) + \text{margin}, 0\} \)

Learning Features

- Input (n training classes)
- Classification Probability (Positive Sample)
- Linear Layer = Simple FC (Fully Connected) Layer.
- Cosine Layer = Normalization on input \( x \) and weights \( w \) before the FC layer.
- Triplet Loss Function:

\[
d(x_i, y_i) = \|x_i - y_i\|
\]

\[
L(a, p, n) = \max\{d(a, p) - d(a, n) + \text{margin}, 0\}
\]

Evaluation Metric: mean Average Precision (mAP)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training Time</th>
<th>Epochs</th>
<th>Classification Accuracy</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BYU-Cars</td>
<td>24 min</td>
<td>35</td>
<td>N/A</td>
<td>76.41</td>
</tr>
<tr>
<td>CrossEntropy No Cosine</td>
<td>22 min</td>
<td>50</td>
<td>83.8%</td>
<td>62.28</td>
</tr>
<tr>
<td>WARP Loss</td>
<td>10 min</td>
<td>10</td>
<td>84.10%</td>
<td>62.03</td>
</tr>
<tr>
<td>CrossEntropy with Cosine</td>
<td>22 min</td>
<td>50</td>
<td>90.46%</td>
<td>67.82</td>
</tr>
<tr>
<td>Satellite-Cars</td>
<td>1 hr 30 min</td>
<td>35</td>
<td>N/A</td>
<td>17.66</td>
</tr>
<tr>
<td>CrossEntropy No Cosine</td>
<td>1 hr 7 min</td>
<td>50</td>
<td>31.47%</td>
<td>16.76</td>
</tr>
<tr>
<td>WARP Loss</td>
<td>12 min</td>
<td>6</td>
<td>31.53%</td>
<td>16.10</td>
</tr>
<tr>
<td>CrossEntropy with Cosine</td>
<td>1 hr 7 min</td>
<td>50</td>
<td>34.17%</td>
<td>17.45</td>
</tr>
</tbody>
</table>

Conclusions

- Triplet Loss suffers in training time due to the online sampling process.
- Using image classification with CrossEntropy or WARP Loss is an alternate method for learning features for image retrieval, that can maintain relatively good mAP scores.
- Cosine distance-based classification leads to better classification accuracy and mAP (retrieval) scores.

References

- WSABIE: Scaling Up To Large Vocabulary Image Annotation, 2011.

WARP Loss + Cosine Distance Layer improves Fine-grained Image Retrieval Accuracy and Reduces Training Time.