Evaluation of the Use of Attention in Image Classification

COMP3740 Jingyi Zhang u7029630
Supervised by Prof Stephen Gould
MOTIVATION AND RELATED WORK
CONTRIBUTIONS
NETWORK STRUCTURE
EXPERIMENT RESULTS ANALYSIS
FUTURE WORK
PART 1

MOTIVATION AND RELATED WORK
Attention Mechanism for Image Classification

- Attention mechanism works by selecting a focused location while enhancing different representations of objects at that location[1].
- Its importance is shown by evidences from human perception process[2] and Mnih et al. [2] also suggest that an attention-based model may be better than a convolutional neural network at both dealing with clutter and scaling up to large input images.
Attention Mechanism for Image Classification

Figure 1: Left: an example shows the interaction between features and attention masks. Right: example images illustrating that different features have different corresponding attention masks in our network. The sky mask diminishes low-level background blue color features. The balloon instance mask highlights high-level balloon bottom part features. (F. Wang et al, 2017)
Residual Attention Network (RAN)

- Fei Wang et.al proposed a CNN model named “Residual Attention Network” [1].
- The Residual Attention Network is constructed by stacking multiple Attention Modules. Each Attention Module has a trunk branch $T(x)$, which performs feature processing task and a mask branch, which learns a mask $M(x)$ of the same size that serves as control gates of the neurons of the trunk branch.
Residual Attention Learning

where $i$ ranges over all spatial positions of the feature map, $c \in \{1, \ldots, C\}$ is the index of the channel and $M(x)$ ranges from $[0, 1]$. 

Figure 2. Inner structure of attention module

$$H_{i,c}(x) = (1 + M_{i,c}(x)) \times T_{i,c}(x)$$
Gaps

• Not too much work about some detailed settings of attention mechanism has been done.
  1. Not many kinds of constraints to the attention features have been tried, while some methods can possibly result in better classification accuracy.
     • In the RAN proposed by Wang et al. [1], they only tried with Sigmoid as the constraining operation.
     • Many of the constraining operations do not have closed-form solutions.
  2. Exploration of the effects of adding different normalization to the attention output are needed.
Deep Declarative Network (DDN)

- Gould et al. [3] proposed Deep Declarative Network (DDN), which is a new class of end-to-end learnable models wherein data processing nodes (or network layers) are defined in terms of the solution to an optimization problem rather than an explicit forward function.
- Any explicitly defined forward processing function can also be defined as a declarative node in a DDN.
Declarative Nodes

1. Definition of the Node of a projection operation[3]

\[ y \in \arg \min_{u \in \mathbb{R}^n} \frac{1}{2} \|x - u\|^2 \]

subject to \( h(u) = 0 \)
## 2. Gradient Calculation for Back-propagation[3]

### Table 1. Gradient Calculation

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>1</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h(u)$</td>
<td>$|u|_2 - 1$</td>
<td>$|u|_1 - 1$</td>
<td>$|u|_\infty - 1$</td>
</tr>
<tr>
<td>$D_Y h(y)$</td>
<td>$y$</td>
<td>$\text{vec}{\text{sign}(y_i)}^T$</td>
<td>$\text{vec}{[i \in I^<em>]\text{sign}(y_i)}^T$, $I^</em> = {i</td>
</tr>
<tr>
<td>$D^2_{YY} h(y)$</td>
<td>$I - yy^T$</td>
<td>$0_{n \times n}$</td>
<td>$0_{n \times n}$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>$1 - |x|_2$</td>
<td>$\text{sign}(y_i)(y_i - x_i)\forall i$</td>
<td>$\left[i \in I^<em>\right]\text{sign}(y_i)(y_i - x_i)\forall i \in I^</em>$</td>
</tr>
<tr>
<td>$D_y(x)$</td>
<td>$\frac{1}{|x|_2}(I - yy^T)$</td>
<td>$I - \frac{D_Y h(y)^T D_Y h(y)}{D_Y h(y) D_Y h(y)^T}$</td>
<td>$I - \frac{D_Y h(y)^T D_Y h(y)}{D_Y h(y) D_Y h(y)^T}$</td>
</tr>
</tbody>
</table>
PART 2

CONTRIBUTIONS
Main Contribution

- Validate whether declaratively defined nodes can be easily integrated into the RAN and use declaratively defined nodes to implement different constraining operations.
- Conduct experiments on whether using projections as attention constraining operations is better than using conventional methods like Sigmoid and Softmax.
- Conduct experiments on whether adding normalization to the attention map helps get better accuracy.
PART 3

NETWORK STRUCTURE
Network Structure with Attention

Figure 3. Structure of the Network
Structure of the attention module

Figure 2. Inner structure of attention module
PART 4

EXPERIMENT RESULTS
Experiment Results

1. Normalization Mechanisms Comparison

Figure 4. Accuracy Comparisons
Experiment Results

1. Normalization Mechanisms Comparison

1. The distributions of accuracy are slightly different for using different normalization methods.
2. Not adding normalization to the soft mask branch outperform others methods overall.

Figure 5. Frequency Comparisons
Experiment Results

2. Constraints Mechanisms Comparison
2.1 Combinations Resulting in the Highest Accuracy

Table 2. Top-5 Results

<table>
<thead>
<tr>
<th>Stage1</th>
<th>Stage2</th>
<th>Stage3</th>
<th>Normalization</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Softmax</td>
<td>L₂-Sphere</td>
<td>Softmax</td>
<td>N/A</td>
<td>94.92</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>L∞-Sphere</td>
<td>Sigmoid</td>
<td>N/A</td>
<td>94.8</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>L₂-Sphere</td>
<td>Sigmoid</td>
<td>N/A</td>
<td>94.72</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>L₂-Sphere</td>
<td>L₂-Sphere</td>
<td>N/A</td>
<td>94.72</td>
</tr>
<tr>
<td>L₂-Sphere</td>
<td>L∞-Sphere</td>
<td>L₂-Sphere</td>
<td>N/A</td>
<td>94.68</td>
</tr>
<tr>
<td>ResNet-152</td>
<td></td>
<td></td>
<td>N/A</td>
<td>75.92</td>
</tr>
</tbody>
</table>
Experiment Results

2. Constraints Mechanisms Comparison
2.2 Best Choice for Each Stage

Figure 6. Comparison of Distribution of Accuracy
Experiment Results

2. Constraints Mechanisms Comparison

2.2 Best Choice for Each Stage

Observation:
1. L2-sphere projection outperforms other constraints in all stages, especially in stage 2.
Experiment Results

2. Constraints Mechanisms Comparison

2.2 Best Choice for Each Stage

T-test:
\( \alpha = 0.05 \)

HA: Choosing a particular constraint operation for a particular stage is resulting in high accuracy.

H0: Choosing a particular constraint operation for a particular stage is not related to high accuracy.

The result:
- \( P \)-value=0.008 << \( \alpha/2 \)
- Choosing L2-Sphere projection for stage 2 is resulting in high accuracy.
Experiment Results

2. Constraints Mechanisms Comparison
2.3 Combinations that Perform Well

T-test:
$\alpha = 0.05$

HA: Choosing a particular combination for two particular stages is resulting in high accuracy.
H0: Choosing a particular combination for two particular stages is not resulting in high accuracy.

Table 3. p-value Results

<table>
<thead>
<tr>
<th>first stage</th>
<th>first method</th>
<th>second stage</th>
<th>second method</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage1</td>
<td>Softmax</td>
<td>Stage2</td>
<td>L2-Sphere</td>
<td>0.0127</td>
</tr>
<tr>
<td>Stage1</td>
<td>Softmax</td>
<td>Stage3</td>
<td>L1-Sphere</td>
<td>0.0173</td>
</tr>
<tr>
<td>Stage2</td>
<td>L2-Sphere</td>
<td>Stage3</td>
<td>Softmax</td>
<td>0.0192</td>
</tr>
<tr>
<td>Stage1</td>
<td>Sigmoid</td>
<td>Stage2</td>
<td>L2-Sphere</td>
<td>0.0216</td>
</tr>
</tbody>
</table>
Future Work

- More experiments are needed to exclude randomness in the result.
- Interpret the results in theory.
- Try on different datasets.
- Try with models consisting of different numbers of attention stages to find other factors involved resulting in the accuracy difference.
Citation


Thanks!