Inference Networks, Graph Convolutional Networks

Greg Mori
School of Computing Science
Simon Fraser University
Outline

• Image annotation with label hierarchies
  • Hu et al. CVPR 2016

• Message passing with deep structured networks
  • Deng et al. BMVC 2015, CVPR 2016
Image Classification

• A natural image can be categorized with labels at different concept layers

Hu, Deng, Zhou, Liao, Learning Structured Inference Neural Networks with Label Relations, CVPR 2016
Label Correlation Helps

- Such categorization at different concept layers can be modeled with label graphs
- It is natural and straightforward to leverage label correlation
Goal: A generic label relation model

- Infer the entire label space from visual input
- Infer missing labels given a few fixed provided labels

Visual Architecture (CNN)

Inference Machine on Knowledge Graph

Activated or Partial Label

Visual Activation

Output Activation

Prediction

Refined Probability

Back-propagate Gradient from Loss Function

Hu, Deng, Zhou, Liao, Learning Structured Inference Neural Networks with Label Relations, CVPR 2016
Top-down Inference Neural Network

- Refine activations for each label
- Pass messages top-down and within each layer of label graph

Activation at current concept layer

Activation at last concept layer

Vertical weight propagates information across concept layers

Horizontal weight propagates information within concept layers

Production initial visual activation from CNN

$$x_t^i = W_t \cdot CNN(I^i) + b_t$$

Hu, Deng, Zhou, Liao, Learning Structured Inference Neural Networks with Label Relations, CVPR 2016
Bidirectional Inference Neural Network (BINN)

- Bidirectional inference to make information propagate across entire label structure
- Inference in each direction independently and blend results
Structured Inference Neural Network (SINN)

- BINN is hard to train
- Regularize connections with prior knowledge about label correlations
- Decompose connections into Positive correlation + Negative correlation

Hu, Deng, Zhou, Liao, Learning Structured Inference Neural Networks with Label Relations, CVPR 2016
Structured Inference Neural Network (SINN)

- Evolve BINN formulation with regularization in connections

ReLU neuron is essential to keep positive/negative contribution

\[ \gamma(x) = ReLU(x) \]
Prediction from Purely Visual Input

- Visual architecture (e.g. Convolutional Neural Network) produces visual activation
- SINN implements information propagation bidirectionally and produces refined output activation

Hu, Deng, Zhou, Liao, Learning Structured Inference Neural Networks with Label Relations, CVPR 2016
Prediction with Partially Observed Labels

- Reverse Sigmoid (logit) neuron produces activation from Partial labels
- SINN adapts both visual activation and activation from partial labels to infer the remaining labels
Reverse sigmoid (logit): produce activation from label

• Reverse the sigmoid function to produce sigmoid input

Inverse of sigmoid

\[ y = \sigma(x) = \frac{1}{1 + \exp^{-x}} \]

\[ a(y) = \log \frac{1}{1 - g(y)}, \]

\[ g(y) = \begin{cases} 
    y + \epsilon, & \text{if } y = 0, \\
    y - \epsilon, & \text{if } y = 1. 
\end{cases} \]

Use a small \texttt{epsilon} to keep numerical stability (0.005)
Image Datasets

- Evaluate with two types of experiments on three datasets

**Animals with Attributes** [Lampert et al. 2009]

- **Labels**
  - 28 taxonomy terms
  - 50 animal classes
  - 85 attributes

- **Task:** predict entire label set
  - Taxonomy terms are constructed from Word Net as [Hwang et al. 2012]
  - Knowledge graph constructed by combining class-attributes graph with taxonomy graph

**NUS-WIDE** [Chua et al. 2009]

- **Labels**
  - 698 image groups
  - 81 concepts
  - 1000 tags

- **Task:** predict 81 concepts with observing tags/image groups
  - Knowledge graph produced by Word Net using semantic similarity
  - 698 image groups constructed from image meta data

**SUN 397** [Xiao et al. 2012]

- **Labels**
  - 3 coarse
  - 16 general
  - 397 fine-grained

- **Task 1:** predict entire label set
- **Task 2:** predict fine-grained scene given coarse scene category
  - Knowledge graph provided by dataset
Ex1: Inference from visual input

- Produce predictions on entire label space
- Evaluate on each concept layer (measured by mAP per class)
- Consistent improvement over baselines on different concept layers

Animal With Attributes

SUN 397
Ex2: Inference from partial labels (NUS-WIDE)

- Produce predictions given partial 1k tags and 698 image groups

Correct predictions are marked in blue while incorrect are marked in red
Ex2: Inference from partial labels (NUS-WIDE)

- Evaluate on standard 81 ground truth classes of NUSWIDE
- Outperform all baselines by large margin

Hu, Deng, Zhou, Liao, Learning Structured Inference Neural Networks with Label Relations, CVPR 2016
Ex2: Inference with partial labels (SUN397)

- Produce predictions given coarse-level labels (3 coarse categories)

Correct predictions are marked in **blue** while incorrect are marked in **red**
Ex2: Inference with partial labels (SUN397)

- Evaluate on 397 fine-grained scene categories
- Significantly improved performance

Hu, Deng, Zhou, Liao, Learning Structured Inference Neural Networks with Label Relations, CVPR 2016
Video Dataset: YouTube-8M

- Youtube-8M V1 / V2
  - 8 million / 7 million videos
  - ~500K hours of video
  - 4800 possible labels
  - 1.8 / 3.4 labels per video average

- Inception V3 frame features
- Neural network audio features
# Results

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>gAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>YouTube-8M v1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM [Abu El Haija et al.]</td>
<td>26.6</td>
<td>N/A</td>
</tr>
<tr>
<td>Logistic regression [Abu El Haija et al.]</td>
<td>28.1</td>
<td>N/A</td>
</tr>
<tr>
<td>CNN features</td>
<td>27.98</td>
<td>60.34</td>
</tr>
<tr>
<td><strong>YouTube-8M v2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BINN</td>
<td>31.18</td>
<td>64.74</td>
</tr>
</tbody>
</table>

### Figure 1: Diagram of proposed model for performing video label inference

- **Input Frames**
- **Feature Extraction**
- **Feature Aggregation**
- **Hierarchical Label Inference**
- **Coarse-Grained Classification**
- **Fine-Grained Classification**
- **Animal**
- **Cat**
- **Adorable**
- **Bread**

Each frame of a video is fed through a pre-trained CNN, followed by a mean pooling for temporal aggregation. Inference is performed in label space, and predictions are made at multiple levels of granularity.
Summary

• Inference in structured label space

• Relations within and across levels of a label space

• Model positive and negative correlations between labels in end-to-end trainable model
Outline

• Image annotation with label hierarchies
  • Hu et al. CVPR 2016

• Message passing with deep structured networks
  • Deng et al. BMVC 2015, CVPR 2016
Overview

• Group Activity Recognition: structures
Overview

• Flexible Structure to adaptively capture discriminative dependencies [Lan et al. 10, Amer et al. 14]
Overview: Probabilistic Graphical Models

\[ P(\text{fall}|\text{image}) \]
Belief Propagation

- Messages' contents depend on graphical model parameters.
- Messages passed depend on graphical model structure.

\[ P(\text{fall}|\text{image}) \]
Overview

• Combine Graphical Models and Deep Learning, enable message passing
Overview

- Combine Graphical Models and Deep Learning, enable **message passing** and **structure learning**
Method

1. Message passing in a Recurrent Neural Network (RNN)

1. Gating function between nodes

1. Structure learning of graphical model
Recurrent Networks
Method: Message Passing in RNN

- Another view of belief propagation: Inference Machine, classify messages [Ross et al. CVPR’11]
Method: Message Passing in RNN

Nodes for each person's action

Make output prediction

CNN

Prediction Layer

Prediction Layer

$t$

$t+1$
Method: Message Passing in RNN

• Represent potential functions in message passing process

• The same potential function corresponds to the same message classifier

• Weights are shared across all instances with same semantic meaning

• The same with prediction layer
Method: Gating Function between Instances

• We introduce instance level gates:

waiting

walking
Method: Gating Function

- C.f. Long Short-Term Memory (LSTM), Gated Recurrent Unit

```
Content(t-1)  Gate  Content(t)
```
Method: Graphical Model Structure Learning

- Use instance gates in inference machines:
Method: General Purpose Inference Machine

- Untying weights between recurrent units
Method: Review of Pipeline
Method: Review of Pipeline

Structure Inference Machine

Walking?
Waiting
Waiting
Walking
Waiting
Waiting
Walking
Waiting
Waiting
Waiting
Waiting
Walking
Experiment: Collective Activity Dataset

- Contains 44 video clips
- Action classes: others, crossing, waiting, queuing, walking, talking
- Scene classes: crossing, waiting, queuing, walking, talking
Experiment: Collective Activity Dataset Extended

- Contains 72 video clips
- Action classes: others, crossing, waiting, queuing, talking, dancing, jogging
- Scene classes: crossing, waiting, queuing, talking, dancing, jogging

Choi et al., CVPR 2011
Experiment: Nursing Home Dataset

- Contains 80 videos captured from a nursing home
- Action classes: walking, standing, sitting, bending, squatting and falling
- Scene classes: fall and non-fall

Deng et al., BMVC 2015
## Quantitative Results

### Collective Activity Dataset

<table>
<thead>
<tr>
<th>Iterations</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tied</td>
<td>73.86%</td>
<td>74.02%</td>
<td>74.02%</td>
</tr>
<tr>
<td>Gated Tied</td>
<td>80.12%</td>
<td>80.9%</td>
<td>81.22%</td>
</tr>
<tr>
<td>Gated Untied</td>
<td>80.12%</td>
<td>81.06%</td>
<td>81.22%</td>
</tr>
</tbody>
</table>

### Collective Activity Extended Dataset

<table>
<thead>
<tr>
<th>Iterations</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tied</td>
<td>84.45%</td>
<td>87.97%</td>
<td>87.97%</td>
</tr>
<tr>
<td>Gated Tied</td>
<td>89.51%</td>
<td>90.14%</td>
<td>90.14%</td>
</tr>
<tr>
<td>Gated Untied</td>
<td>89.51%</td>
<td>90.14%</td>
<td>90.23%</td>
</tr>
</tbody>
</table>

### Nursing Home Dataset

<table>
<thead>
<tr>
<th>Iterations</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tied</td>
<td>83.68%</td>
<td>84.91%</td>
<td>84.91%</td>
</tr>
<tr>
<td>Gated Tied</td>
<td>84.46%</td>
<td>85.32%</td>
<td>85.32%</td>
</tr>
<tr>
<td>Gated Untied</td>
<td>84.46%</td>
<td>85.50%</td>
<td>85.50%</td>
</tr>
</tbody>
</table>
Qualitative Results

• Gates between person-level node and group activity
Summary

- Recurrent network for inference in a graphical model
- Nodes in graphical model represented by units in recurrent network
- Inference by repeated message passing
- Structure learning by gating functions
Graph Convolutional Networks

• These are both examples of a broad class of methods known as Graph Convolutional Networks

  Neural network architectures designed to run over graphs

  “Convolutions” defined over adjacent nodes in the graph, filters shared over all nodes in the graph

  Variations in terms of form of function, normalization, layers, adjacency
Reading list

• Duvenaud et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints, NIPS 2015

• Kipf and Welling. Semi-Supervised Classification with Graph Convolutional Networks, ICLR 2017

• Jain et al. Structural-RNN: Deep Learning on Spatio-Temporal Graphs, CVPR 2016

• Santoro et al. A simple neural network module for relational reasoning, NIPS 2017

• Ibrahim and Mori. Hierarchical Relational Networks for Group Activity Recognition and Retrieval, ECCV 2018