Decoupled Box Proposal and Featurization with Ultrafine-Grained Semantic Labels Improve Image Captioning and Visual Question Answering

Soravit (Beer) Changpinyo

Joint work with Bo Pang, Piyush Sharma, and Radu Soricut
Faster R-CNN
Trained to predict object + attributes of Visual Genome

Figure adapted from Ren et al. 2016 (TPAMI version)
Detected Objects by Faster R-CNN (Visual Genome)

Each box comes with <rank>: <optional attribute class> <object class> <score x 100>
Ultrafine-grained Semantic Labels

Spectrum of Semantic Similarity

<table>
<thead>
<tr>
<th>Category-level (Coarse-grained)</th>
<th>Fine-grained level</th>
<th>Instance level (Ultrafine-grained)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bridge</td>
<td>Steel red bridge</td>
<td>Golden Gate Bridge</td>
</tr>
</tbody>
</table>

Figure adapted from Juan et al. 2019
Graph-RISE on Internal Data

Juan et al. 2019

Labeled Images

Graphs of Unlabeled Images

ResNet-101

Graph Regularizer

Image Embedding

40M classes

Subsampling 100K classes

Figure adapted from Juan et al. 2019
Our Work (To be presented at EMNLP 2019)

Decoupled Boxes + Ultrafine-Grained Featurization

= Better Image Captioning & Visual Question Answering
Image Captioning on Conceptual Captions

Region Faster-RCNN
“a woman walks through the streets .”

Region Ultra
“monks walking in front of a temple”

“monks clean a garden at a temple .”
Informative Words induced by Ultra

“mughal structure on the way”

“lollipop on a yellow background”

“green algae in the sea”

“breakfast with coffee and croissants on a white wooden background.”
Possible to predict from Top 5 FRCNN objects?

“mughal structure on the way”

“green algae in the sea”

“lollipop on a yellow background”

“breakfast with coffee and croissants on a white wooden background”
VQA on VizWiz

“Is it sunny outside?”
yes

“How much money is this?”
1 dollar
20

“What is this?”
beer
bbq sauce

“What does it say on this card?”
unanswerable
unanswerable

Data: Gurari et al. 2018
Model: Anderson et al. 2018, Jiang et al. 2018