A Fast and Accurate One-Stage Approach to Visual Grounding

Presenter: Tianlang Chen
Zhengyuan Yang  Boqing Gong  Liwei Wang  Wenbing Huang  Dong Yu  Jiebo Luo
Visual grounding

- Grounding a language query onto a region of the image

Query: bottom right grass
Existing framework

- Two-stage framework

Query: center building
Existing framework

- Performance is capped by the region candidates
- Slow in speed

Query: center building

Query: bottom right grass
One-stage visual grounding

- One-stage approach
- Generally applicable for sub-tasks in grounding
Why one-stage visual grounding

- No region candidates -> 7~20% higher in accuracy
- One-stage -> 10x faster
Architecture overview

- Encoder
- Fusion module
- Grounding module
Architecture

- Encoder
- Fusion module
- Grounding module

- Visual encoder: DarkNet53+FPN
- Language encoder: Bert, LSTM, FV
- Spatial encoder: location related queries
Architecture

- Encoder
- Fusion module
- Grounding module

- Image-level fusion
Architecture

- Encoder
- Fusion module
- Grounding module

- Output format: box + confidence
Datasets

• Phrase localization: Flickr 30K Entities
• Referring expression comprehension: ReferItGame

Flickr 30K Entities

ReferItGame
Comparison to other methods

ReferItGame

Flickr30K Entities

Inference Speed
Qualitative results

- Union of multiple objects
- Stuff as opposed to things
- Challenging regions
A Fast and Accurate One-Stage Approach to Visual Grounding

Code & models: https://github.com/zyang-ur/onestage_grounding

Poster: #26

Contact: zyang39@cs.rochester.edu