Interpretable and Globally Optimal Prediction for Textual Grounding using Image Concepts

Raymond A. Yeh[1] Jinjun Xiong[2] Wen-mei W. Hwu[1] Minh N. Do[1] Alexander G. Schwing[1]

1. Dept. of ECE, University of Illinois at Urbana-Champaign 2. IBM Research

Motivation

Task: Phrase localization (*i.e.*, given a phrase and image find the corresponding bounding box described in the phrase)

- Challenging multimodal task
- Joint understanding of language and image
- Variety of objects and instance level descriptions

Contributions:

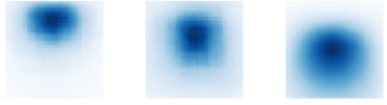
- A unified framework
- Does not rely on object proposals
- Globally optimal prediction
- Learned model parameters are interpretable

From Plummer *et al.* 2017

Introduction

Problem Formulation:

- Input: $x = (Q, I)$
- Output: $y = (y_1, ..., y_4)$
- Bounding box prediction: $\hat{y} = \arg\min_{y \in \mathcal{Y}} E(x, y, w)$



Intuition of Our Approach:

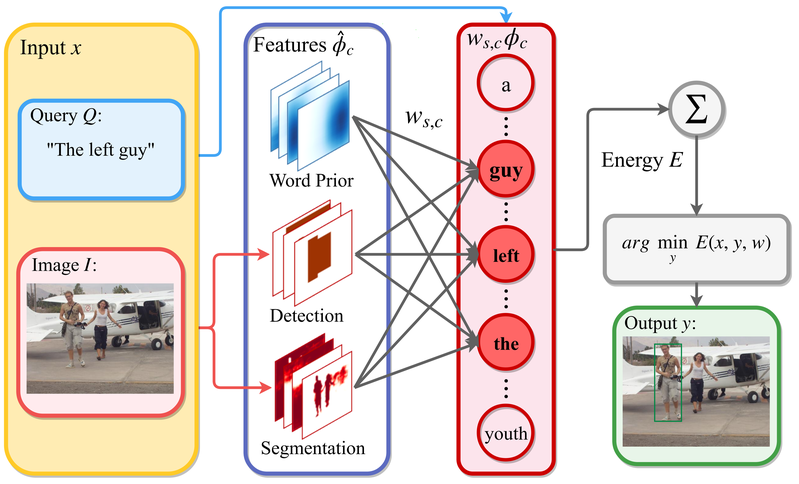
- Explicit word-concept modeling
- Combine "image concepts"

Energy Function Details:

- Image concept $c \in \mathcal{C}$ is attached to a parametric score map $\hat{\phi}_c(x, w_r) \in \mathbb{R}^{W \times H}$
- $\phi_c(x, y, w_r) \in \mathbb{R}$ refers to the score accumulated within the bounding box y of score map $\hat{\phi}_c(x, w_r)$
- Designed energy function:

$$E(x, y, w) = \sum_{s \in \mathcal{S}: \delta(s \in Q) = 1} \sum_{c \in \mathcal{C}} w_{s,c} \phi_c(x, y, w_r)$$

Our Approach



Learning the Parameters:

- Structured support vector machine based surrogate loss minimization:

$$\min_w \quad \frac{C}{2}\|w\|_2^2 + \sum_{(x,y) \in \mathcal{D}} \max_{\hat{y} \in \mathcal{Y}} \left(-E(x, \hat{y}, w) + L(\hat{y}, y)\right) + E(x, y, w)$$

- Task loss $L(\hat{y}, y)$ is Intersection over Union (IoU)
- Parameters w includes score map parameters w_r and top layer parameters $w_{s,c}$
- Parameter $w_{s,c}$ connects word $s \in \mathcal{S}$ and concept $c \in \mathcal{C}$
- With fixed w_r, training is equivalent to training a structured SVM
- The cutting-plane algorithm of Joachims *et al.* 2009 works well
- To solve the inference problem effectively, we utilize efficient subwindow search by Lampert *et al.* 2009

Results

Method	Acc. (%)
SCRC (Hu 2016)	27.80
DSPE (Wang 2016)	43.89
GroundeR(Rohrbach 2016)	47.81
CCA (Plummer 2017)	50.89
Ours	**53.97**

Results on Flickr30k dataset

Method	Acc. (%)
SCRC (Hu 2016)	17.93
GroundeR (Rohrbach 2016)	26.93
Ours [Prior + Geo]	25.56
Ours [Prior + Geo + Seg]	**33.36**
Ours [Prior + Geo + Seg + Det]	**34.70**

Results on ReferItGame dataset

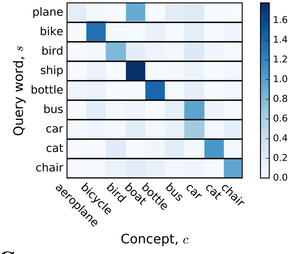
Success Cases:

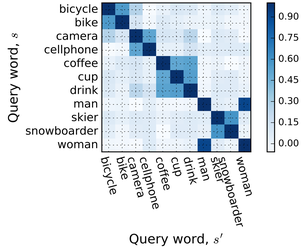
second bike from right in front

painting next to the two on the left

person all the way to the right

Weights Interpretability:





Failure Cases:

her shoes

a red shirt

a dirt bike