Localization and Detection with Natural Language Queries

- Tradition object detection: person, cat, dog, car, motorbike, airplane, bed, television, ...
- Natural language query: a group of bikers, a building with lots of windows, ...

**Discriminative Bimodal Networks (DBNet)**

- Typical previous work
  - Caption generation model
  - Output in a huge language space
  - Using only positive training samples (matched text and regions), or a limited amount of negative samples
- Our DBNet
  - A fully discriminative model
  - Learning a binary classifier is easier
  - Able to use all possible negative samples across the training set

**Model Architectures**

- Text features: Character-level CNN
  - Replicating a phrase until reaching the input length of the CNN (256 characters)
  - Non-saturating neurons: LeakyReLU
  - Replaceable by other text embedding (e.g., skip-thoughts)

- The detection score of the image region \(i\) is given by a linear classifier dynamically generated according to the text feature \(\Phi(t_i)\): \(\text{score}(x_i) = \text{linear}(\Phi(t_i), \Theta(t))\).

- An extra regularization term on the classifier parameter is beneficial for the SGD stability: \(\text{linear}(\Phi(t_i), \Theta(t)) + \lambda \text{L2}()\).

**Benchmarking Protocol**

- Based on the Visual Genome database
  - Bounding boxes with text phrase descriptions
  - Additional spell checking and auto-correction

- Localization: find a known-to-exist entity

- Detection: localize all matched entities if exists any
  - Should have negative images (no matched entity)
  - Impractical to test all queries on each image

- We proposed the first benchmarking protocol for visual entity detection with language queries.

- 3 difficulty levels with increasing number of random negative images per query phrase:
  - Level 0: no negative image
  - Level 1: the same number as the positive image
  - Level 2: 5 times as the number of positive images or 20 ( whichever is greater) for each test phrase

- Detection metric: average precision (AP)
  - mean AP (mAP): averaging APs over all phrases; each phrase has its own decision threshold.
  - global AP (gAP): a single AP for any phrases; all phrases share the same decision threshold.

**Model Training: Labels, Objectives, and Optimization**

- Any region-text pair \((i, t)\) can be possibly used during training.
- \(r\) can be GT or proposed (EdgeBox), \(r\) can be an annotation on the same image as \(i\), or it can be from the rest of the training set.
- Spatial overlapping based training labels:
  - \(\{1\} + \{r\} \) has large overlap with a ground truth region of \(i\)

- Text similarity based uncertainty augmentation:
  - If \(f\) is similar to \(t\), \(r\) should also have uncertain label with \(<\)

- Given a region, the non-(uncertain) phrases are categorized into: positive phrases (pos), negative phrases from the annotations on the same image (neg), and negative phrases from the rest images (rest).

- Training loss is normalized separately for the three categories:

\[
L = \lambda_{\text{pos}} L_{\text{pos}} + \lambda_{\text{neg}} L_{\text{neg}} + \lambda_{\text{rest}} L_{\text{rest}}
\]

where \(\lambda_{\text{pos}}, \lambda_{\text{neg}}, \lambda_{\text{rest}}\) and the sampling strategy in SGD determines \(\lambda_{pos}, \lambda_{neg}\) (smaller).

- Optimization: (initialization) Initializing the visual pathway with the pretrained Faster R-CNN;
  - (phase 1) training the text pathway from scratch with the visual pathway unchanged;
  - (phase 2) jointly finetuning the two pathways; (phase 3) decreasing the learning rate.

**References**

[Huang et al., 2016; Liu et al., 2017; Xie et al., 2016; He et al., 2017; Sun et al., 2017; Xiang et al., 2017; Xiong et al., 2017; Zhang et al., 2017; Zhao et al., 2017; Zellers et al., 2017]