LinkNet: Relational Embedding for Scene Graph

- Conference: NeurIPS 2018
- Code: Unofficial
- Authors:
  - Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon - KAIST, South Korea

This paper proposes LinkNet, a new model for scene graph generation. LinkNet model consists of three modules.

1. Object relational embedding
2. Global context encoding (GCE)
3. Geometric layout encoding

Input : Object proposals and features from a region proposal network (RPN).
Each object proposal is represented as a vector $o_i = (f_i^{RoI}, K_0 l_i, c)$. K_0 is a parameter matrix which maps the distribution of predicted labels l_i of each of the object proposal $i = 1, ..., N$.

1. Object relational embedding

Object features are learnt using a graph-based approach.

$$\mathbf{R}_1 = \mathrm{softmax}\left((\mathbf{O}_0\mathbf{W}_1)(\mathbf{O}_0\mathbf{U}_1)^{\mathbf{T}}\right) \in \mathbb{R}^{\mathbf{N}\times\mathbf{N}} - \mathrm{Relational\ embedding}$$

$$\mathbf{O}_1 = \mathbf{O}_0 \oplus fc_0\left((\mathbf{R}_{\mathbf{1}}(\mathbf{O}_0\mathbf{H}_1))\right) \in \mathbb{R}^{\mathbf{N}\times 4808}$$

$$\mathbf{O}_2 = fc_1(\mathbf{O}_1) \in \mathbb{R}^{\mathbf{N}\times 256} - \mathrm{Relation\text{-}aware\ embedding}$$

$\oplus$ denotes elementwise summation. O_1 can be considered as applying a graph convolutional (GCN) layer with a residual connection. The resultant features O_2 is once again fed into a similar set of layers to get $O_4 \in \mathbb{R}^{N\times C_{obj}}$.

2. Global context encoding (GCE)

$$c \in \mathbb{R}^{512} - \mathrm{Average\ pooling\ of\ RPN\ image\ featurs}$$

Features c is concatenated with other RPN featurs to get o_i.

3. Geometric layout encoding

This encodes relative location and scale information of an object.

$$\mathbf{b_{o|s}} = \left(\frac{\mathbf{x_o} - \mathbf{x_s}}{\mathbf{w_s}}, \frac{\mathbf{y_o} - \mathbf{y_s}}{\mathbf{h_s}}, \log\left(\frac{\mathbf{w_o}}{\mathbf{w_s}}\right), \log\left(\frac{\mathbf{h_o}}{\mathbf{h_s}}\right)\right)$$

x_o, y_o, h_o, w_o: coordinates, height, and width of the object proposal of object o
o and s stand for object and subject respectively. These features are used for learning *edge-relational embeddings*.

Loss function

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{obj\_}cls} + \lambda_1 \mathcal{L}_{\text{rel\_}cls} + \lambda_2 \mathcal{L}_{\text{gce}}$$

By default λ_1 and λ_2 are set to 1.