Overview

- **Goal:** Generating a sentence to describe a given image.
- **Motivation:** The word-by-word caption generation process predicts attributes before referring to the subject they are referring to. Mixtures of attributes, subjects, and relations in a complete sentence affect training effectiveness.
- **Proposal:**
  - Step 1: global prediction of the objects and their relationship in the image.
  - Step 2: an object-wise attribute description.
  - Step 3: merge skeleton and attributes for complete caption prediction.

Our Coarse-to-fine Model

- **Skeleton Attribute decomposition** extracts skeleton text and attributes from a training caption, using the Stanford constituency parser.
  - **CNN:** extracts skeleton and attributes.
  - **ATT:** predicts the skeleton sentence given the image features.
  - **Attr-LSTM:** predicts the attribute sequence for each skeletal word, conditioned on the skeleton word and the context of skeleton sentence.
  - **Variable-length caption generation** uses a simple length factor.
  \[ \log(P) = \log(P_a) + y \cdot I \]
  - \( I \): the length of the sentence
  - \( y \): the length factor to encourage/disourage longer sentences

Tag enhancement takes a bag of detected tags as input, and add \( y \cdot w \) to each tag word log-probability. \( y \) is the encourage factor. \( w \) is the tag confidence.

Results on COCO and Stock3M

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Models</th>
<th>Baseline</th>
<th>COCO</th>
<th>Stock3M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcelona</td>
<td>0.778</td>
<td>0.793</td>
<td>0.785</td>
<td>0.787</td>
</tr>
<tr>
<td>METEO</td>
<td>0.709</td>
<td>0.720</td>
<td>0.717</td>
<td>0.719</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.769</td>
<td>0.780</td>
<td>0.778</td>
<td>0.780</td>
</tr>
<tr>
<td>COCO</td>
<td>0.742</td>
<td>0.777</td>
<td>0.774</td>
<td>0.776</td>
</tr>
<tr>
<td>Stock3M</td>
<td>0.723</td>
<td>0.756</td>
<td>0.753</td>
<td>0.755</td>
</tr>
</tbody>
</table>

Background

- **CNN:**
  - [Show and Tell, O. Vinyals, et al. 2015]
  - [Show, Attend and Tell, K. Xu, et al. 2016]

- **ATT:**
  - [Lin et al. 2014]

- **ResNet:**
  - [Krizhevsky et al. 2012]

- **LSTM:**
  - [Hochreiter and Schmidhuber, 1997]

- **Baseline**
  - [Vinyals et al. 2015]
  - [Xu et al. 2015]
  - [Zhang et al. 2015]

- **COCO**
  - [Lin et al. 2014]

- **Stock3M**
  - [Sharma et al. 2015]