Figure Captioning with Reasoning and Sequence-Level Training

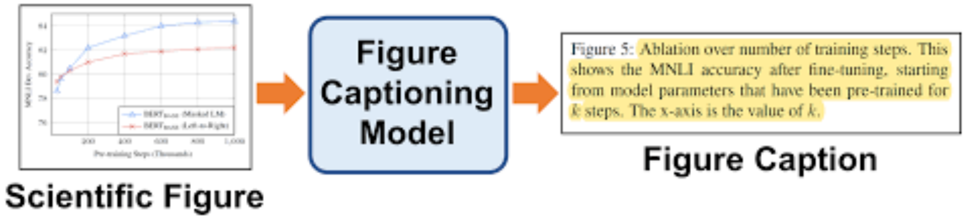
백인진 22.08.22

Image Captioning

Edit

377 papers with code • 27 benchmarks • 49 datasets

Image Captioning is the task of describing the content of an image in words. This task lies at the intersection of computer vision and natural language processing. Most image captioning systems use an encoder-decoder framework, where an input image is encoded into an intermediate representation of the information in the image, and then decoded into a descriptive text sequence. The most popular benchmarks are nocaps and COCO, and models are typically evaluated according to a BLEU or CIDER metric.



Introduction

- Main Contributions

1. We introduce a new dataset for figure captioning called *FigCAP*.

2. We propose two novel attention mechanisms to improve the decoder's performance.

   - The Label Maps Attention enables the decoder to focus on specific labels.

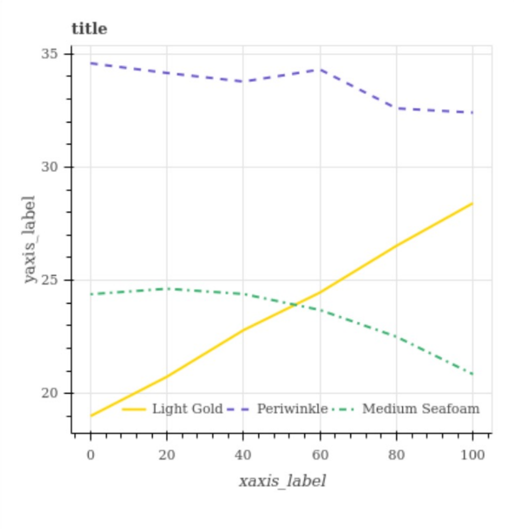
   - The Relation Maps Attention is proposed to discover the relations between figure labels.

3. We utilize sequence-level training with reinforcement learning to handle long sequence generation and alleviate the issue of exposure bias.

4. Empirical experiments show that the proposed models can effectively generate captions for figures under several metrics.

Background

- **Figure VQA**



Q: Does Medium Seafoam intersect Light Gold?
A: Yes

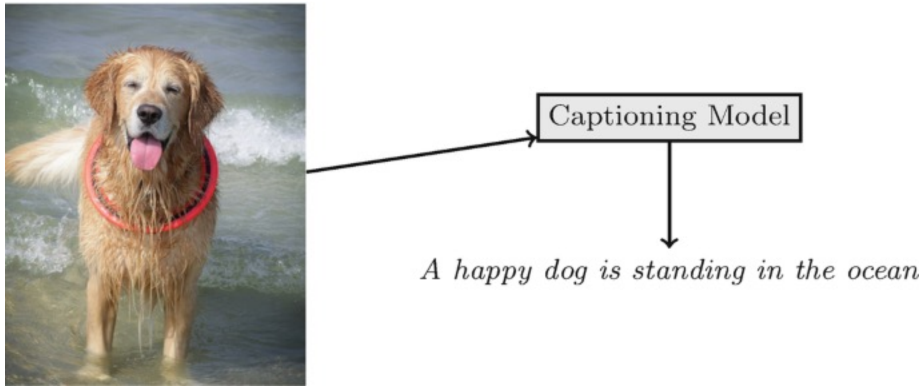
Q: Is Medium Seafoam the roughest?
A: No

Q: Is Light Gold less than Periwinkle?
A: Yes

Q: Does Periwinkle have the maximum area under the curve?
A: Yes

Q: Does Medium Seafoam have the lowest value?
A: No

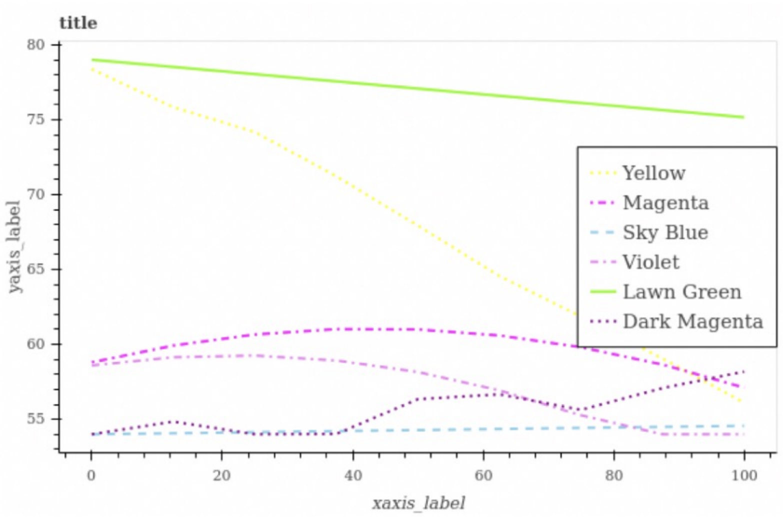
- **Image Captioning**



- Difference
  - Input : image & question
  - Output : the answer to the given question, commonly containing only a few words

- Difference
  - Input : normal image
  - Output : a relatively short length

Need to build a methodology only for figure captioning



High-level Caption

This figure is a line plot; it contains six categories: Yellow, Magenta, Sky Blue, Violet, Lawn Green and Dark Magenta.

Detailed Caption

Dark Magenta has the lowest value. Lawn Green has the highest value. Sky Blue is less than Lawn Green. Yellow is greater than Violet. Sky Blue has the minimum area under the curve. Lawn Green is the smoothest. Yellow intersects Magenta.

- FigCAP
  - Horizontal bar chart
  - Vertical bar chart
  - Pie chart
  - Line plot
  - Dotted line plot
- 2 different use cases
  - FigCAP-H : high-level descriptions
  - FigCAP-D : Detailed descriptions (relationship among the labels of categories)
- Challenging
  - Much longer than natural image captioning dataset
  - Logical information
  - How to capture key information and insights automatically

Model

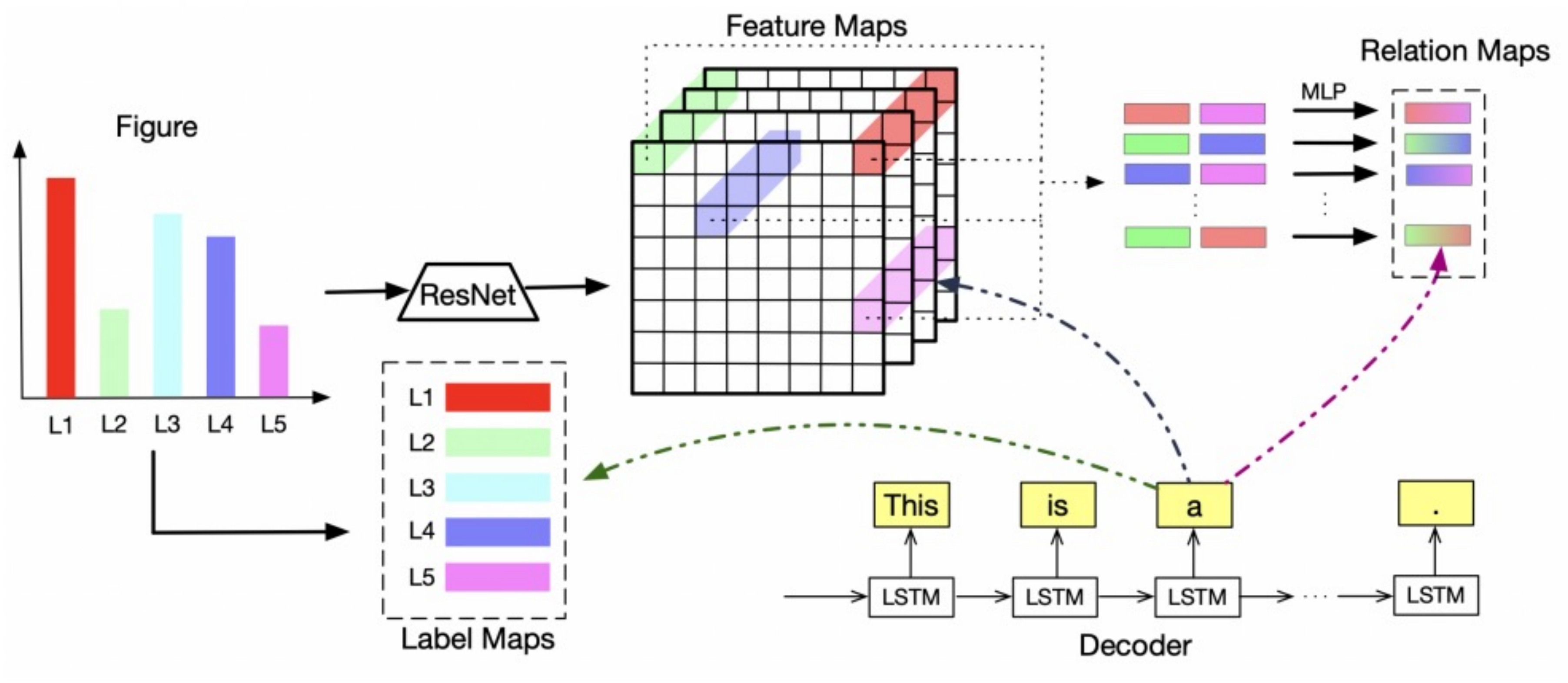
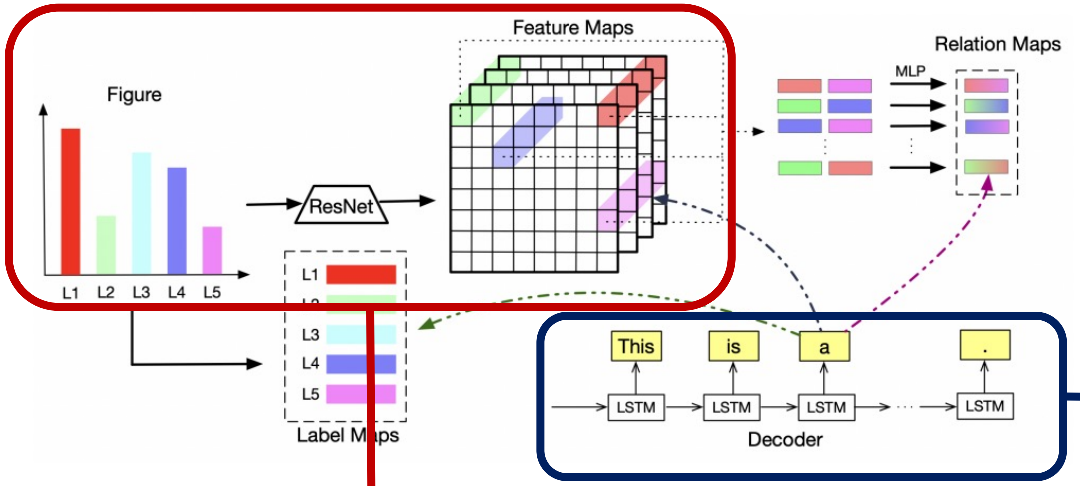


Figure
L1
L2
L3
L4
L5
ResNet
Feature Maps
Relation Maps
MLP
Label Maps
L1
L2
L3
L4
L5
This
is
a
.
LSTM
LSTM
LSTM
LSTM
Decoder



Figure

$$F = ResNet(X)$$

X : the figure

F -> used to initialize a LSTM

$$c_0 = \sigma(W_{Ic}F)$$
$$h_0 = \sigma(W_{Ih}F)$$

Caption

words : one-hot encoding → e_t

gates

$$i_t = \sigma(W_{iy}e_t + W_{ih}h_t + W_{id}\boldsymbol{d_t} + b_i)$$
$$f_t = \sigma(W_{fy}e_t + W_{fh}h_t + W_{fd}\boldsymbol{d_t} + b_f)$$
$$o_t = \sigma(W_{oy}e_t + W_{oh}h_t + W_{od}\boldsymbol{d_t} + b_o)$$

cell state & hidden state

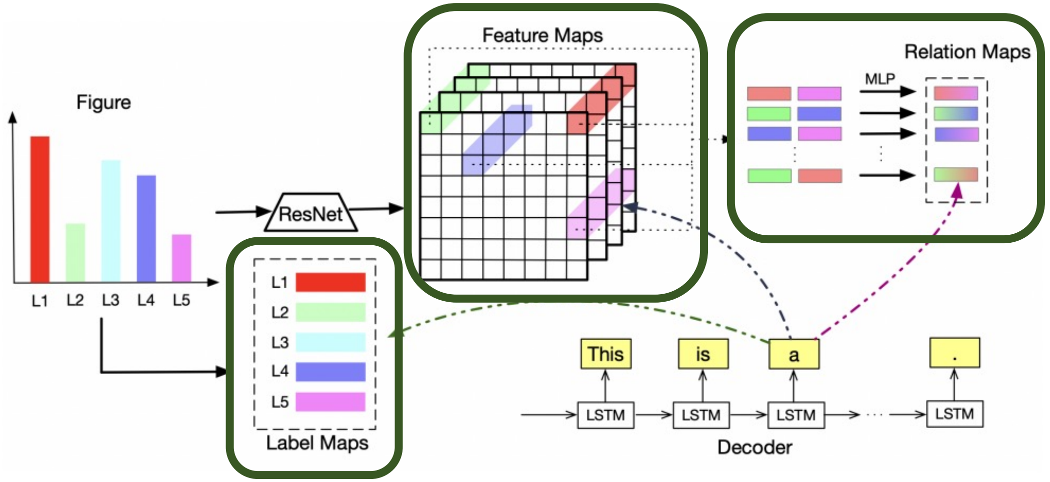
$$c_t = i_t \odot \phi(W_{cy}^{\otimes}\, e_t + W_{ch}^{\otimes}\, h_{t-1} + W_{cd}^{\otimes}\, d_t + b_c^{\otimes}) + f_t \odot c_{t-1}$$
$$h_t = o_t \odot \tanh(c_t)$$

next word prediction

$$\tilde{y}_t = \sigma(W_h h_t + W_d\, d_t)$$
$$y_t \sim softmax(\tilde{y}_t)$$

d_t : context vector

Model



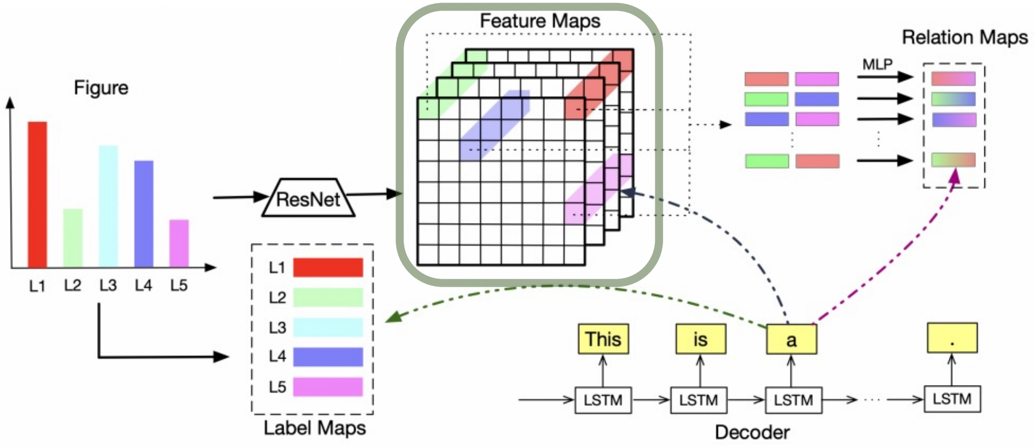
Attention Models

- Relation Maps attentions : Att_R
- Label Maps attentions : Att_L
- Feature Maps attentions : Att_F

d_t : context vector -> combination of (Att_R, Att_L, Att_F)

Model

- Feature Maps Attention ; Att_F
  - Caption Text – Figure



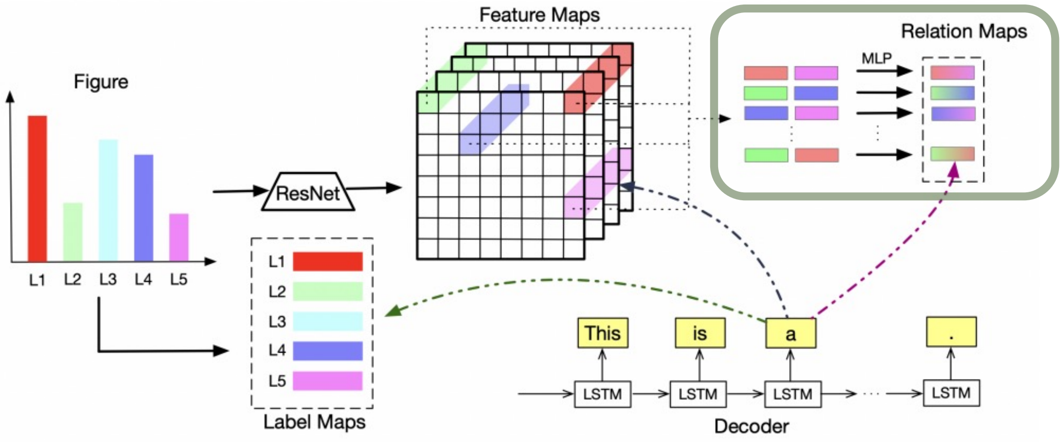
- $$e_{tj} = Att\_F(\boldsymbol{h}_{t-1}, \boldsymbol{f}_j) \qquad (1)$$
$$= \boldsymbol{v}_a^T \tanh(\boldsymbol{W}_a \boldsymbol{f}_j + \boldsymbol{U}_a \boldsymbol{h}_{t-1})$$
$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{m} \exp(e_{tk})}, \quad \boldsymbol{c}_t = \sum_{j=1}^{m} \alpha_{tj} \cdot \boldsymbol{f}_j$$

where $\boldsymbol{f}_j$ is the j-th feature in the feature maps $\boldsymbol{F}$, $\boldsymbol{c}_t$ is the context vector and α_{tj} is an attention weight.

  - h_{t-1}: LSTM output
  - c_t: the weighted sum of all features in the feature maps

Model

- Relation Maps Attention : Att_R
  - Caption Text – Feature Relation

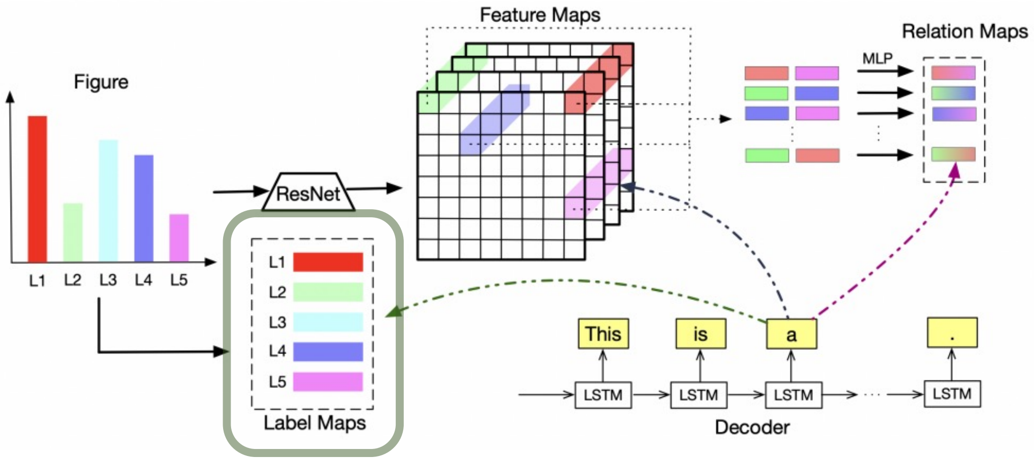


- $$\hat{e}_{tk} = Att\_R(\boldsymbol{h}_{t-1}, \boldsymbol{r}_k) \qquad (3)$$
  $$= \boldsymbol{v}_b^T \tanh(\boldsymbol{W}_b \boldsymbol{r}_k + \boldsymbol{U}_b \boldsymbol{h}_{t-1})$$
  $$\beta_{tk} = \frac{\exp(\hat{e}_{tk})}{\sum_{l=1}^{m^2} \exp(\hat{e}_{tl})}, \qquad \hat{\boldsymbol{c}}_t = \sum_{k=1}^{m^2} \beta_{tk} \cdot \boldsymbol{r}_k$$
  - $r_k = MLP\left(concat(f_i, f_j)\right)$, relation vector

- Representing abstract objects that implicitly represent objects in the figure, not explicitly representing one specific object like a bar or a line.

Model

- Label Maps Attention ; Att_L
  - Caption Text – Figure Label



- $$\tilde{e}_{tj} = Att\_L(\boldsymbol{h}_{t-1}, \boldsymbol{l}_j) \tag{4}$$
  $$= \boldsymbol{v}_c^T \tanh(\boldsymbol{W}_c \boldsymbol{l}_j + \boldsymbol{U}_c \boldsymbol{h}_{t-1}),$$
  $$\gamma_{tj} = \frac{\exp(\tilde{e}_{tj})}{\sum_{j=1}^{n} \exp(\tilde{e}_{tj})}, \quad \tilde{\boldsymbol{c}}_t = \sum_{j=1}^{n} \gamma_{tj} \cdot \boldsymbol{l}_j$$
  - l_j: figure label , extracted from figure using OCR techniques
    -> subset of word embeddings

Context Vector & Objective

- Context Vector ; $d_t = concat(c_t, \widehat{c_t}, \widetilde{c_t})$
  - used as input to the decoder

- Hybrid training objective
  - Traditional "teacher forcing" : exposure bias & indirectly optimizing the evaluation metric
  - Reinforcement learning : alleviating the mentioned problems by directly optimizing the sequence-level evaluation metric
    - Loss for RL : $L_{rl} = \underbrace{-\left(r\left(\hat{Y}^s\right) - r\left(\hat{Y}^b\right)\right)}_{\text{Reward (CIDEr)}} \underbrace{\Sigma_{t=1}^{T} logp(\hat{y}_t^s | \hat{Y}_{t-1}^s, x)}_{\text{Sampled sequence}}$

      $\hat{Y}^s$: sampling / $\hat{Y}^b$:greedy (baseline)

    - Hybrid loss
      - RL loss : purely optimizing sequence-level evaluation metric may lead overfitting
      - To tackle this issue, use hybrid training objective
      - Word-level loss (L_{sl} : provided by MLE) & Sequence-level loss (L_{rl} : provided by RL)
      - $L_{hybrid} = \lambda L_{rl} + (1 - \lambda L_{sl})$

Result

Models	Evaluation Metrics CIDEr	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE
CNN-LSTM	0.232	0.332	0.255	0.201	0.157	0.188	0.270
CNN-LSTM+*Att_F*	0.559	0.333	0.262	0.210	0.168	0.209	0.334
CNN-LSTM+*Att_F*+*Att_L*	**1.018**	**0.337**	**0.269**	**0.215**	**0.170**	**0.227**	**0.368**

Table 3: Results for FigCAP-H: High-level Caption Generation.

- $FigCAP\_H$
  - No relation, much shorter
  - Label maps attention improve model performances
    - Features specific to figures, such as labels, can be utilized to boost the model's performance

Result

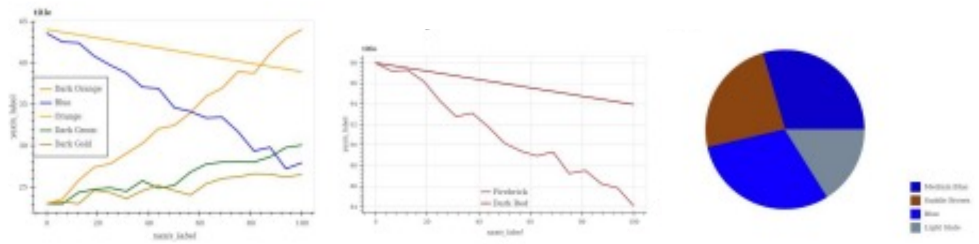
Models	Evaluation Metrics CIDEr	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE
CNN-LSTM	0.158	0.055	0.050	0.044	0.038	0.115	0.244
CNN-LSTM+*Att_F*	0.868	0.215	0.200	0.181	0.159	0.200	0.401
CNN-LSTM+*Att_F*+*Att_L*	0.917	0.232	0.214	0.194	0.170	0.207	0.413
CNN-LSTM+*Att_All*	1.036	0.312	0.290	0.264	0.233	0.231	0.468
CNN-LSTM+*Att_All*+RL	**1.179**	**0.404**	**0.367**	**0.324**	**0.270**	**0.263**	**0.489**

Table 4: Results for FigCAP-D: Detailed Caption Generation. *Att_All*=*Att_F*+*Att_L*+*Att_R*.

- $FigCAP\_D$
  - Relation, much longer
  - Points
    - To discover the relations between the labels in the figures
    - To generate the long sequences of captions
  - Label maps attention improve model performances
  - Using $Att\_F\&Att\_L$ is better in FigCAP_H than $\mathrm{FigCAP\_D}$
  - $Att\_R\&RL$ case is the best. It means that relation and RL can effectively model the relations between the labels of figures and the long sequence generation

Discussions

- The effects of Att_F is more higher in FigCAP-D than FigCAP-H
  - High-level descriptions dose not actually need complex attention models since it is **more likely a classification task** which can be accomplished based on general information of the figure

- Error in label names
  - 

(d) It is a line plot, with five lines; their names are orange, **mediumblue, olive, orangered, greenyellow**.

(f) Firebrick has the minimum area under the curve; darkred has the maximum area under the curve; darkred is the smoothest; firebrick is the roughest; firebrick has the **lowest** value; firebrick intersects darkred.

(h) **Mediumblue** is the maximum; mediumblue is greater than saddlebrown; saddlebrown is less than **saddlebrown**; mediumblue is the high median; saddlebrown is the low median.

  - *Future plan to incorporate a ranking model, which allows **current models select the label with the highest score as the candidate from a set of similar labels.***

- How to make RELATION maps
  - NOW : Caption Text – Feature Relation

$$\hat{e}_{tk} = Att\_R(\boldsymbol{h}_{t-1}, \boldsymbol{r}_k) = \boldsymbol{v}_b^T \tanh(\boldsymbol{W}_b \boldsymbol{r}_k + \boldsymbol{U}_b \boldsymbol{h}_{t-1})$$

$$\beta_{tk} = \frac{\exp(\hat{e}_{tk})}{\sum_{l=1}^{m^2} \exp(\hat{e}_{tl})}, \quad \hat{\boldsymbol{c}}_t = \sum_{k=1}^{m^2} \beta_{tk} \cdot \boldsymbol{r}_k$$

  - Captio on
  - Relatio
    - $r_k = MLP\left(concat(f_i, f_j)\right)$, relation vector

- No code in github…

- Relation ~ 위치, 숫자, 그림의 형태 등이 제대로 학습만 된다면 관계를 찾아주는 것도 충분히 학습이 될 것
  - 보편적으로 많은 종류의 그래프에 대해서 모델이 적용될 수 있는가..
  - Figure에 나타난 object의 type, number를 잘 매핑할 수 있도록 임베딩…
  - 만약에 한다고 하면 type을 줄여서 잘 작동이 되는지 보고 점차적으로 늘여가는 것으로

References

https://openaccess.thecvf.com/content_WACV_2020/html/Chen_Figure_Captioning_with_Relation_Maps_for_Reasoning_WACV_2020_paper.html

@InProceedings{Chen_2020_WACV,
author = {Chen, Charles and Zhang, Ruiyi and Koh, Eunyee and Kim, Sungchul and Cohen, Scott and Rossi, Ryan},
title = {Figure Captioning with Relation Maps for Reasoning},
booktitle = {Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)},
month = {March},
year = {2020}
}