Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning

1. Distinctive Image Captioning

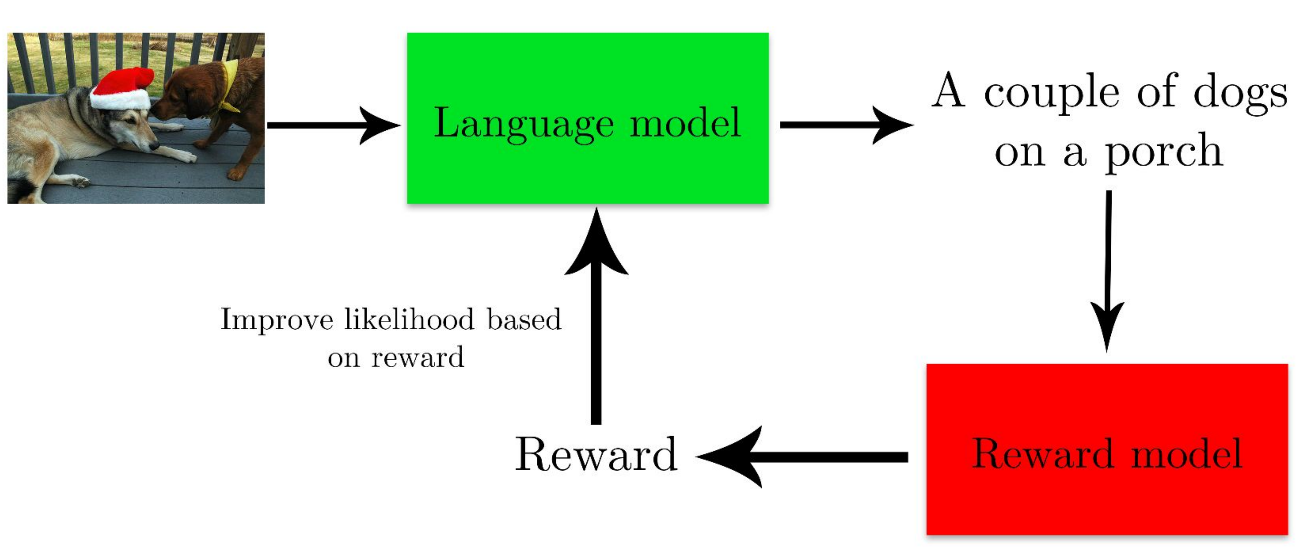
- Image captioning training datasets only describe most salient objects, common to many images
- Metrics push the focus on words common across different images, not specific ones
  - Image captioning models produce very generic texts **describing the image but could describe a lot of others**

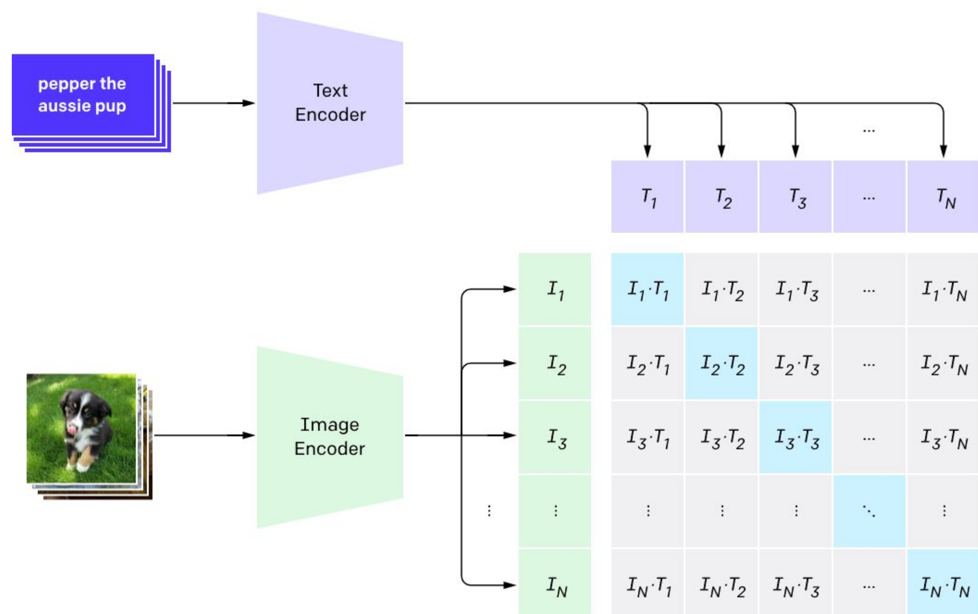
A couple of dogs standing on a porch

- Fine-grained alignment to describe **the input image and only this one**

2. Reinforcement Learning

- Optimize cross-modal similarity of the generated caption and the target image[1,2]
  - Learn to **generate a description that lets the retriever identify the image**
- Dual encoder (CLIP) projects both modalities separately and compute all the similarities in a batch using **simple dot products**





3. Discriminator Regularization

- CLIP is not trained to evaluate written quality
  - Regularization to prevent the model from learning **ill-formed solutions**

*a close up of two **brown** and **black dogs** wearing a **santa hat** on a **black** and **brown dog** with a **red hat** on a backyard with a fence in the background*

- **Simple MLP** using CLIP representations as input

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = -\Big[\Big(\overbrace{\alpha\, r_{sim}(x)}^{\text{Similarity reward}} + \overbrace{(1-\alpha)\, r_{regu}(x)}^{\text{Regularization reward}}\Big)\nabla_\theta \underbrace{\log p_\theta(x)}_{\text{Likelihood}}\Big]$$

4. Bidirectional Contrastive Rewards

- A baseline is subtracted to the reward to reduce variance

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Sample from the generator}} = -\Big(\overbrace{r(x)}^{\text{Reward}} - \overbrace{b}^{\text{Baseline}}\Big)\nabla_\theta \underbrace{\log p_\theta(x)}_{\text{Likelihood}}$$

- Similarity of another caption from the model (image-to-text)[1] or a similar mined image (text-to-image)[2]
- **Decoupled contrastive loss uses the closest element in the batch for both cross-modal directions**

$$r_{bicont}(t_c) = \tau\left(\underbrace{\log\frac{e^{\frac{t_c\cdot i_c}{\tau}}}{\sum_{t\in\mathcal{T}\setminus t_c} e^{\frac{t\cdot i_c}{\tau}}}}_{\text{Image-to-text reward } r_{i2t}(t_c)} + \underbrace{\log\frac{e^{\frac{t_c\cdot i}{\tau}}}{\sum_{i\in\mathcal{I}\setminus i_c} e^{\frac{t_c\cdot i}{\tau}}}}_{\text{Text-to-image reward } r_{t2i}(t_c)}\right)$$

5. Weighted Teacher Forcing

- RL learns from high-scoring sequences and ground truth are good solutions
- RL using GT: **learn to reproduce human-written sequence (TF) but focuses on highly descriptive ones**

✔ there is an adult bear that is walking in the forest
✘ picture of an exterior place that looks wonderful.

6. Experiments & Results

- Trade-off **discriminativeness** (recall@k) using generated caption (fixed CLIP model) and **writing quality** (BLEU, ROUGE, CIDEr, METEOR and SPICE) on MS COCO
  - **MLP on top of CLIP can be used as regularization** (higher retrieval rate without degrading written quality)
  - Weighted Teacher Forcing **improves retrieval metrics using only ground truths, without degrading writing quality**
  - **Both cross-modal directions are needed** for a caption highly descriptive of this image and this image only

Antoine Chaffin
IMATAG, LightOn, IRISA, France
Vincent Claveau
CNRS, DGA, France
Ewa Kijak
IRISA

Contact
antoine.chaffin@lighton.ai

References
[1] Fine-grained Image Captioning with CLIP Reward.
Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, Mohit Bansal
[2] Distinctive Image Captioning via CLIP Guided Group Optimization
Youyuan Zhang, Jiuniu Wang, Hao Wu, Wenjia Xu