UMR IRISA

Institut de Recherche en Informatique et Systèmes Aléatoires

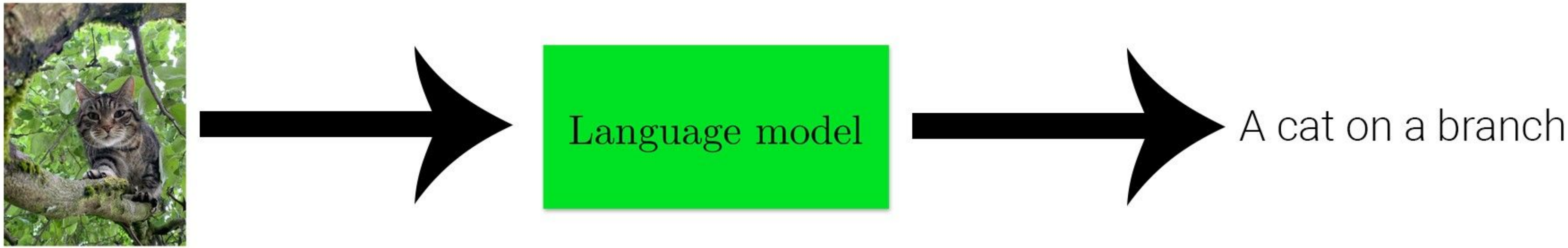
Distinctive Image Captioning: Leveraging Ground Truth Captions in CLIP Guided Reinforcement Learning

Antoine Chaffin,
Vincent Claveau,
Ewa Kijak

IMATAG

Image Captioning

- Language model conditioned on an image
- Create a **powerful cross-modal alignment**[1]



[1] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, Lucas Beyer. "Image Captioners Are Scalable Vision Learners Too". 2023

Distinctive Image Captioning

- Datasets captions only describe most salient objects, common to many images
- Higher word-matching metrics with words common across different images, not specific ones

A couple of dogs standing on a porch

Distinctive Image Captioning

- Datasets captions only describe most salient objects, common to many images
- Higher word-matching metrics with words common across different images, not specific ones

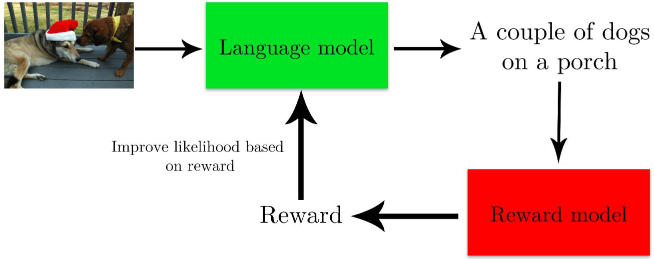
A couple of dogs standing on a porch



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

Cross-modal retriever rewards

- Reinforcement learning to optimize cross-modal similarity of the generated caption and the target image
  - A description that can let the retriever identify the image



Similarity reward

$$\nabla_\theta L_\theta \; (x) = - \, r_{sim}(x) \, \nabla_\theta \; \log p_\theta \; (x)$$

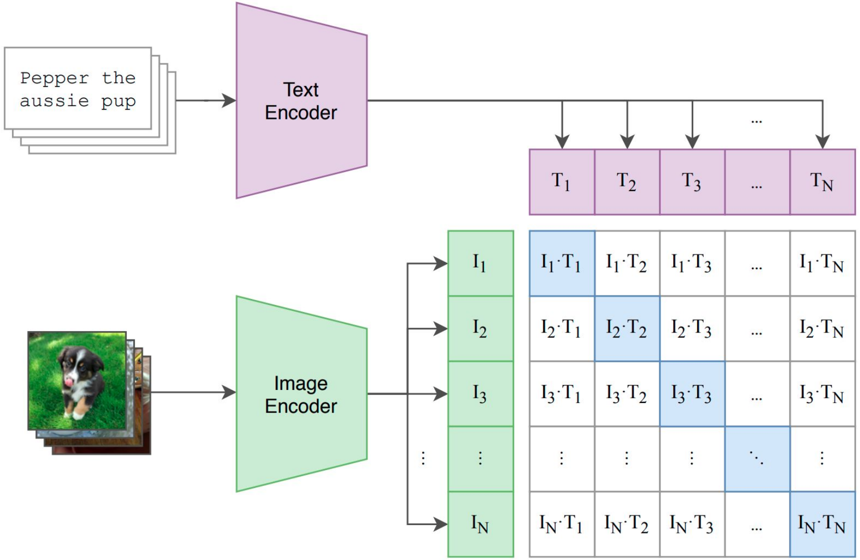
Sample from the generator

Likelihood

a couple of dogs wearing a santa hat on a porch

Contrastive learning

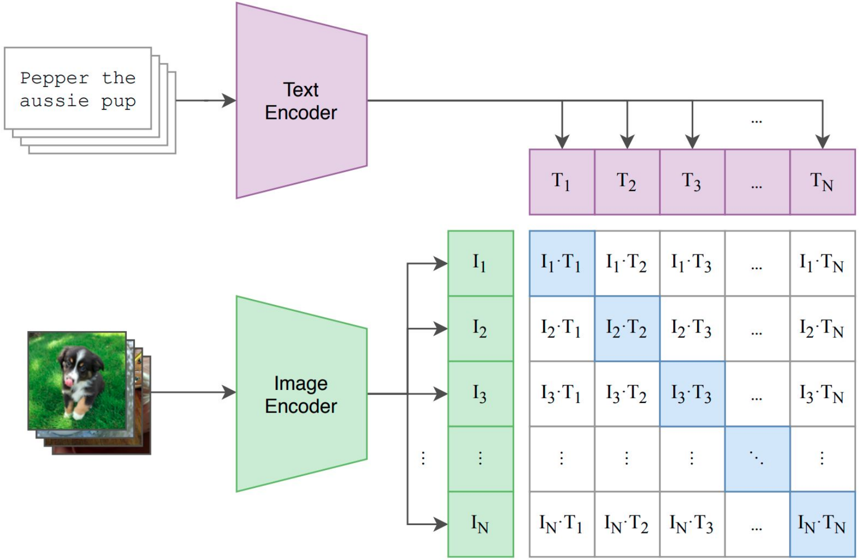
- Dual encoder, each projecting a modality separately
    - Similarity using dot product of both representations



Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". 2021

Contrastive learning

- Dual encoder, each projecting a modality separately
    - Similarity using dot product of both representations
- Couple closer than any element in the batch



$$\mathcal{L}_{\text{CLIP}} = \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{t \in \mathcal{T}} e^{\frac{t \cdot i_c}{\tau}}}}_{\text{image-to-text}} + \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{i \in \mathcal{I}} e^{\frac{t_c \cdot i}{\tau}}}}_{\text{text-to-image}}$$

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". 2021

Regularization

- 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*

Regularization

- Prevent the model from learning ill-formed solutions
- Regularization term in the reward
  - KL divergence, CIDEr value, **grammar network**…

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*

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

Experiments

- 3 different contributions to improve CLIP-based RL image captioning
  1. **Discriminator regularization**
  2. **RL objective on ground truth samples**
  3. **Bidirectional contrastive reward**

Experiments

- 3 different contributions to improve CLIP-based RL image captioning
  1. **Discriminator regularization**
  2. **RL objective on ground truth samples**
  3. **Bidirectional contrastive reward**
- MS COCO dataset
- Trade-off:
  - **Discriminativeness**: recall@k using generated caption (fixed CLIP model)
  - **Writing quality**: BLEU, ROUGE, CIDEr, METEOR and SPICE

Discriminator regularization

- Use generated text discriminator scores as regularization
- Simple MLP using CLIP representations as input

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

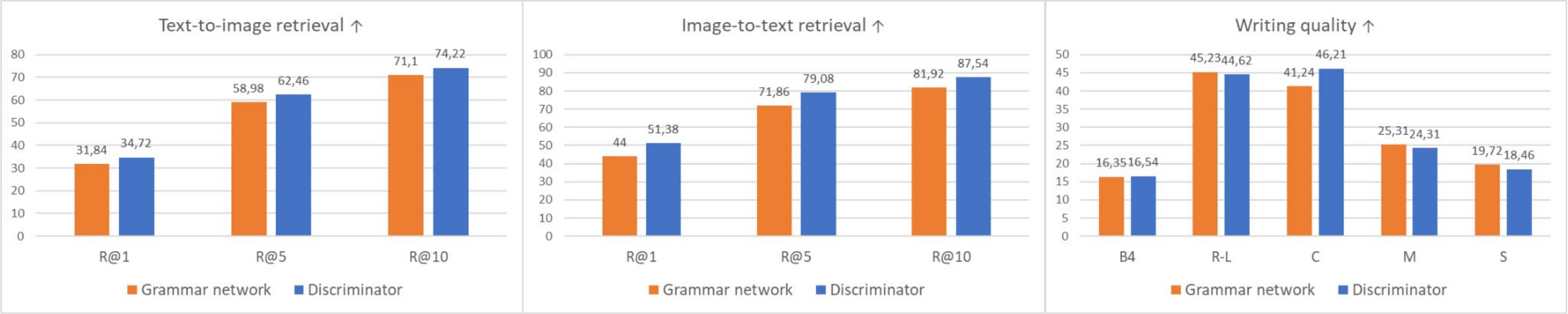
IRISA

Discriminator regularization

- Use generated text discriminator scores as regularization
- Simple MLP using CLIP representations as input

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

- Higher retrieval rate without degrading written quality



Weighted Teacher Forcing

- RL learns from high-scoring sequences
- Ground truths are (relatively) good solutions

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Ground truth sample}} = -\Big(\overbrace{\alpha \, r_{sim}(x)}^{\text{Similarity reward}} + \overbrace{(1-\alpha)\, r_{regu}(x)}^{\text{Regularization reward}} \Big) \nabla_\theta \underbrace{\log p_\theta (x)}_{\text{Likelihood}}$$

Weighted Teacher Forcing

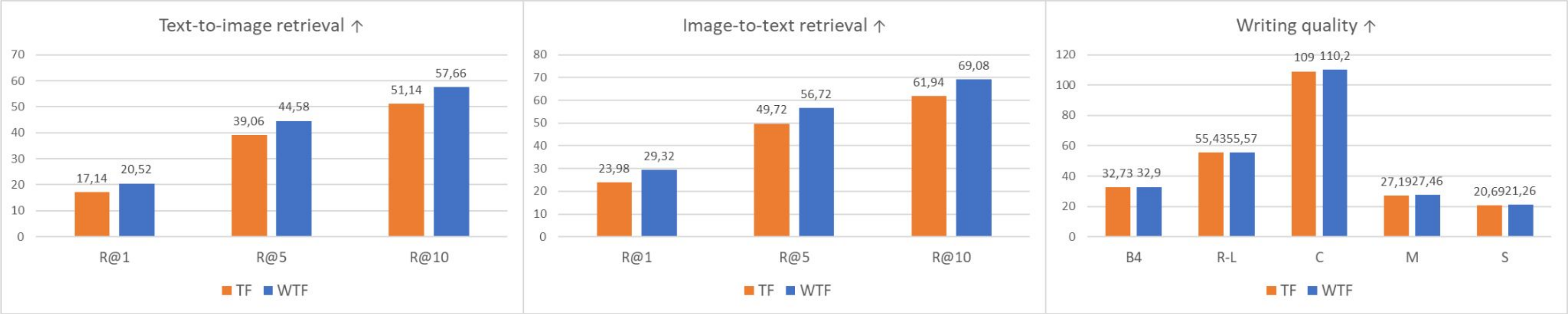
- RL learns from high-scoring sequences
- Ground truths are (relatively) good solutions
- Learn to reproduce human-written sequence (TF) but focuses on highly descriptive ones

$$\nabla_\theta L_\theta \underbrace{(x)}_{\text{Ground truth sample}} = -\left(\overbrace{\alpha \, r_{sim}(x)}^{\text{Similarity reward}} + \overbrace{(1-\alpha)\, r_{regu}(x)}^{\text{Regularization reward}} \right) \nabla_\theta \underbrace{\log p_\theta (x)}_{\text{Likelihood}}$$

(1) there is an adult bear that is walking in the forest
(2) picture of an exterior place that looks wonderful.

Weighted Teacher Forcing

- Improve retrieval metrics using only ground truth, without degrading writing quality
- Better regularization objective to couple with traditional RL



Baseline rewards

- Subtract a baseline to the reward to reduce variance

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

Baseline rewards

- Subtract a baseline to the reward to reduce variance

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

1. Use the model itself as a baseline[1]

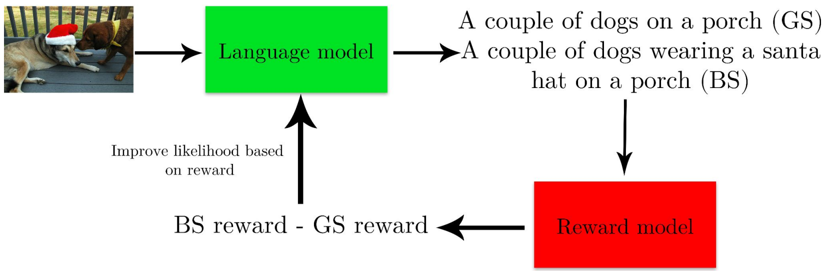


Image-to-text baseline

[1] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, Mohit Bansal. “Fine-grained Image Captioning with CLIP Reward”. 2022

Baseline rewards

- Subtract a baseline to the reward to reduce variance

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

1. Use the model itself as a baseline[1]
2. Similarity with other (similar) images[2]

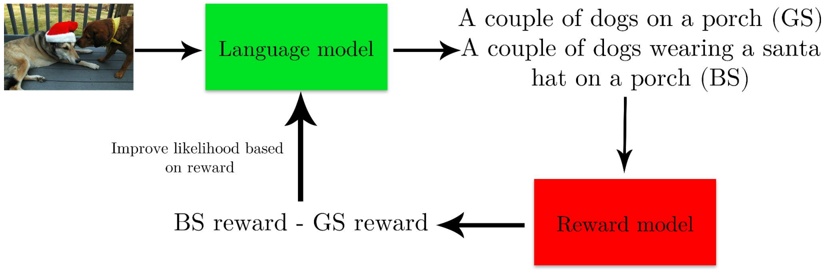
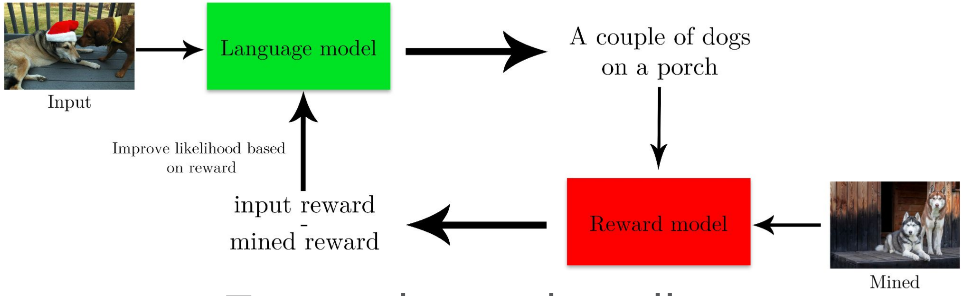


Image-to-text baseline

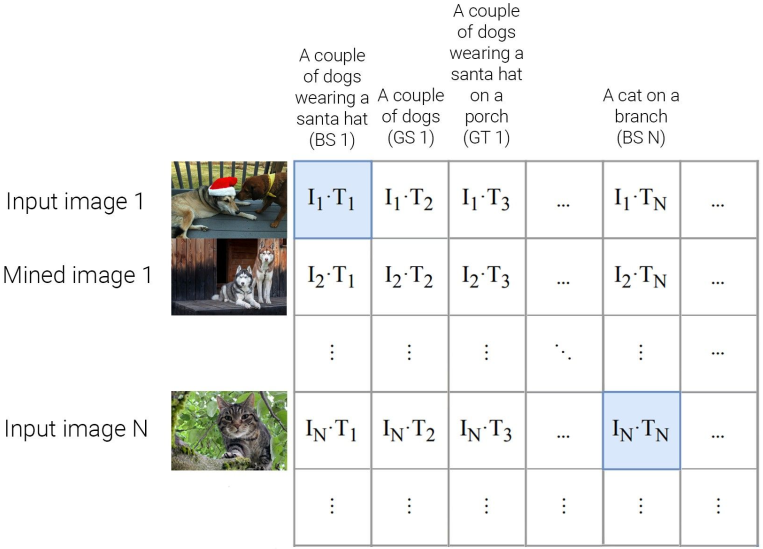


Text-to-image baseline

[1] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, Mohit Bansal. “Fine-grained Image Captioning with CLIP Reward”. 2022
[2] Youyuan Zhang, Jiuniu Wang, Hao Wu, Wenjia Xu. “Distinctive Image Captioning via CLIP Guided Group Optimization”. 2022

Bidirectional Contrastive Reward

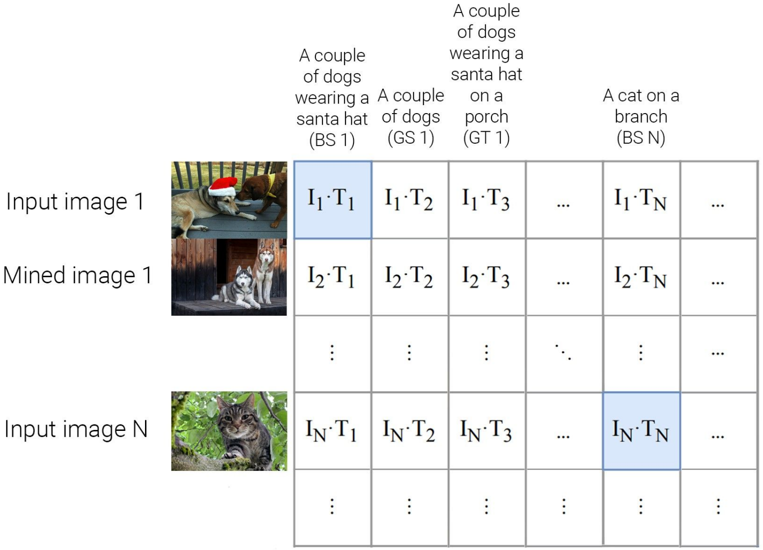
- Decoupled contrastive loss



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

Bidirectional Contrastive Reward

- Decoupled contrastive loss
- Closest element in the batch as baseline
- Natively handle both cross-modal directions



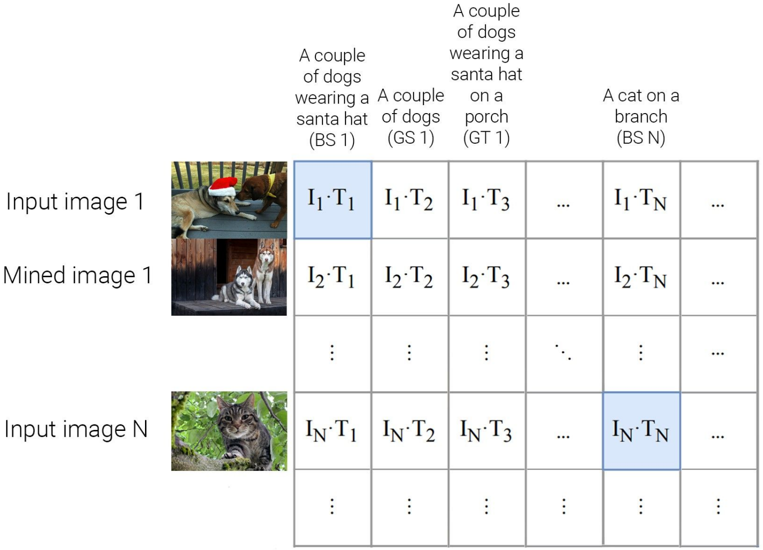
$$r_{bicont}(t_c) = \tau \left(\underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{t \in \mathcal{T} \setminus t_c} e^{\frac{t \cdot i_c}{\tau}}}}_{\text{Image-to-text reward } r_{i2t}(t_c)} + \underbrace{\log \frac{e^{\frac{t_c \cdot i_c}{\tau}}}{\sum_{i \in \mathcal{I} \setminus i_c} e^{\frac{t_c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

$$r_{i2t}(t_c) = \tau \left(\log \left(e^{\frac{t_c \cdot i_c}{\tau}} \right) - \log \left(\sum_{t \in \mathcal{T} \setminus t_c} e^{\frac{t \cdot i_c}{\tau}} \right) \right)$$

$$\approx t_c \cdot i_c - \max_{t \in \mathcal{T} \setminus t_c} \{ t \cdot i_c \}$$

Bidirectional Contrastive Reward

- Decoupled contrastive loss
- Closest element in the batch as baseline
- Natively handle both cross-modal directions
- The caption is **very descriptive of the image and this image only**



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

$$r_{i2t}(t_c) = \tau \left(\log \left(e^{\frac{t_c \cdot i_c}{\tau}} \right) - \log \left(\sum_{t\in\mathcal{T}\setminus t_c} e^{\frac{t\cdot i_c}{\tau}} \right) \right)$$

$$\approx t_c \cdot i_c - \max_{t\in\mathcal{T}\setminus t_c} \{t \cdot i_c\}$$

Bidirectional reward

- Unidirectional image-to-text reward only yield significantly lower text-to-image retrieval results
- Both cross-modal directions are needed for a caption highly descriptive of **this image and this image only**