Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models (CVPR 2018 Spotlight)

Presenter: Yongxin (Richard) Wang

Really quick self-introduction

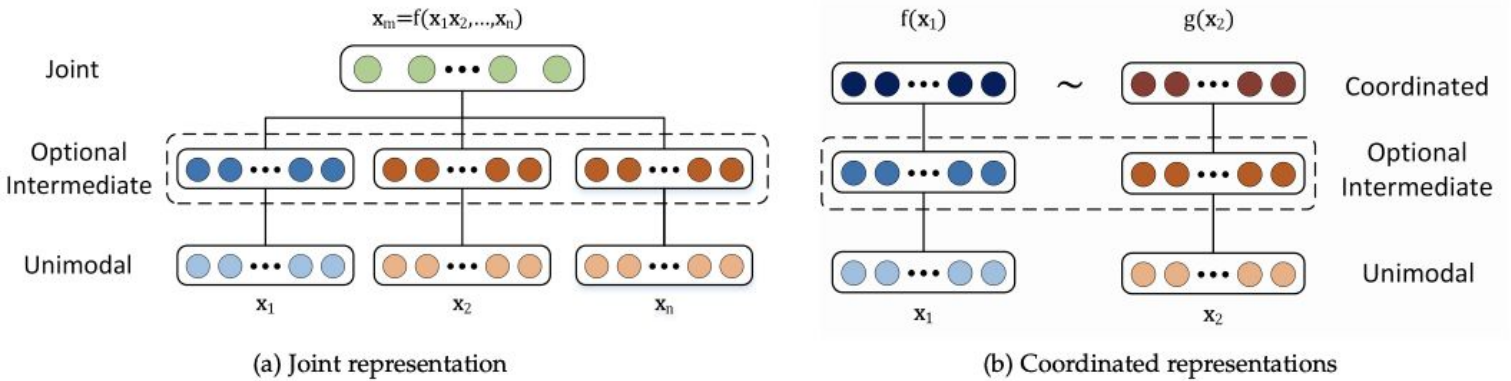
- First-year MS Computer Vision student @ CMU
- Background in Vision and ML (some NLP)
  - Vision: Human gaze analysis, visual object tracking,
  - ML: Domain adaptation
  - NLP: Image captioning

The problem - Cross-modal Retrieval

- Given an image, retrieve relevant texts
- Given some texts, retrieve relevant images

Motivation

- Current methods in cross-modal retrieval:
  - High level semantic features
  - Not sufficient for detailed local similarity (image) and word level similarity (language)



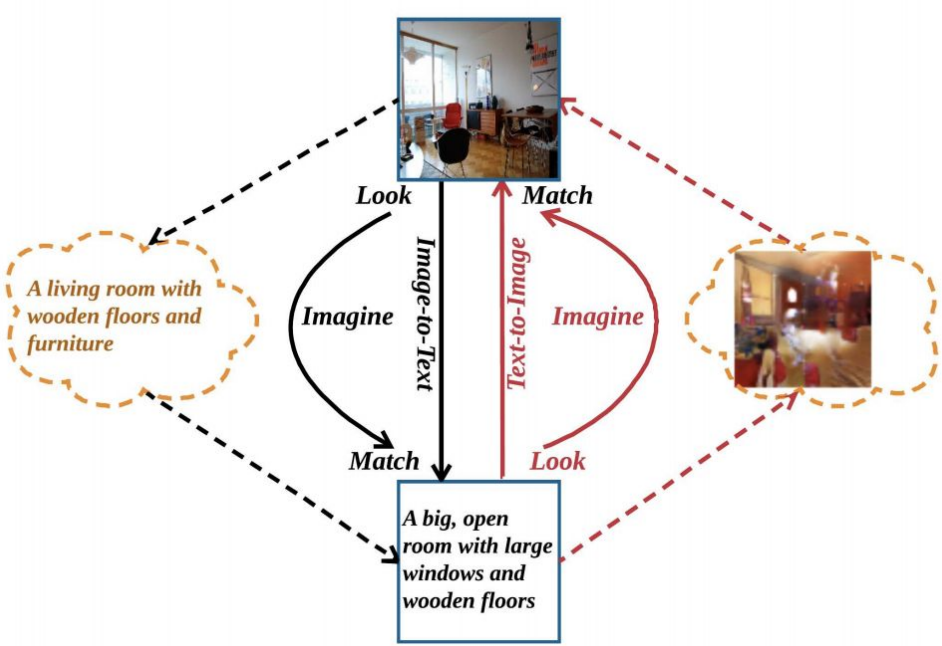
(a) Joint representation

(b) Coordinated representations

T. Baltrušaitis, C. Ahuja and L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019.

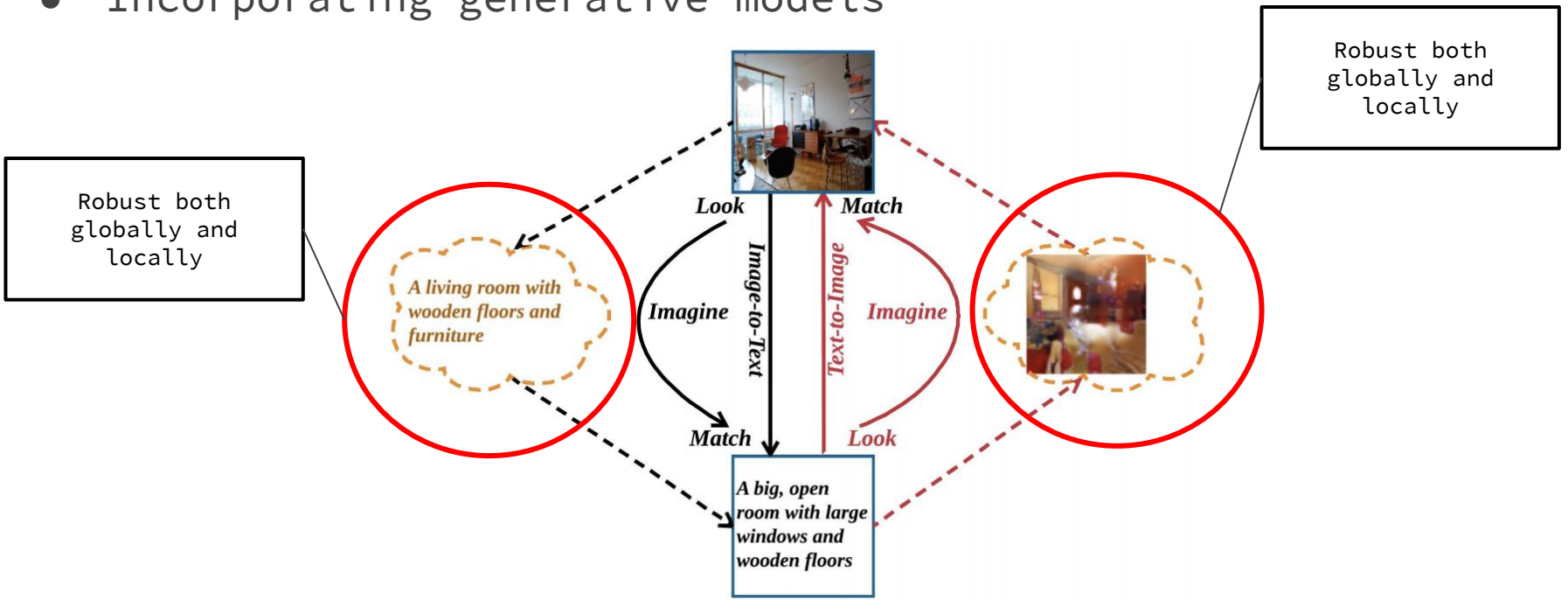
The idea - Look, Imagine, and Match

- Incorporating generative models



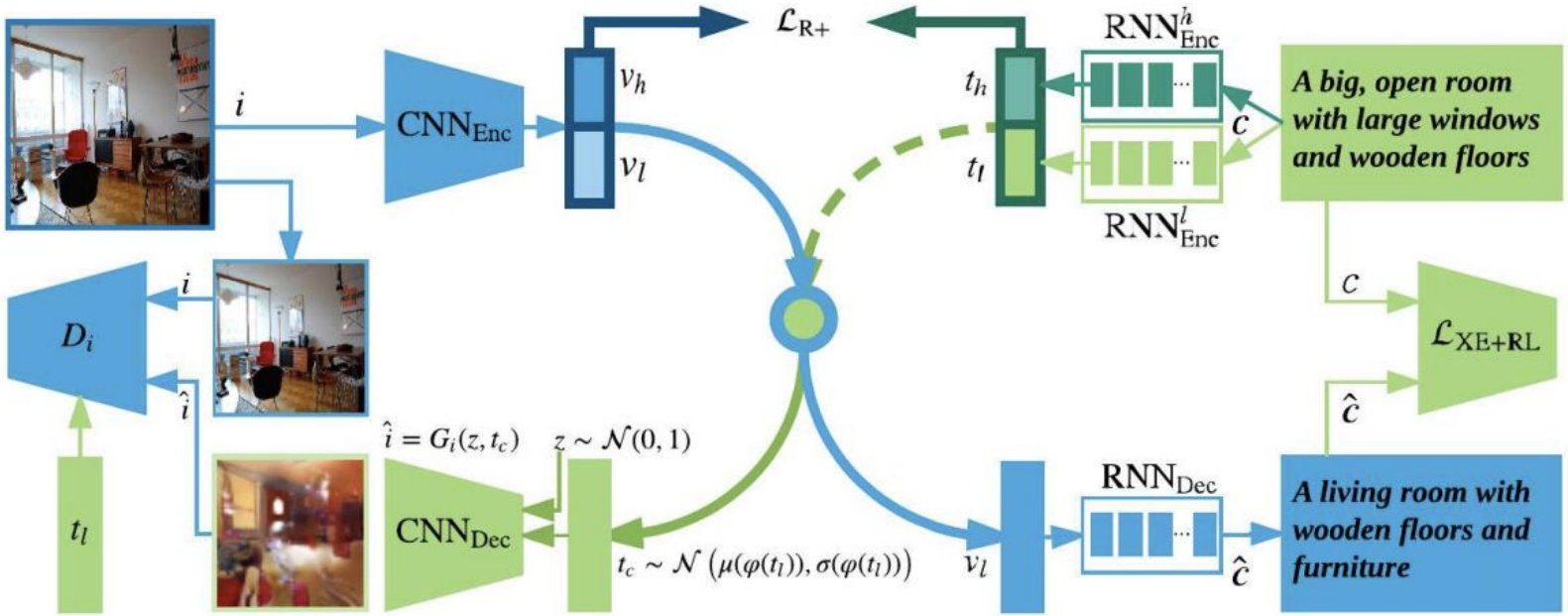
The idea - Look, Imagine, and Match

- Incorporating generative models

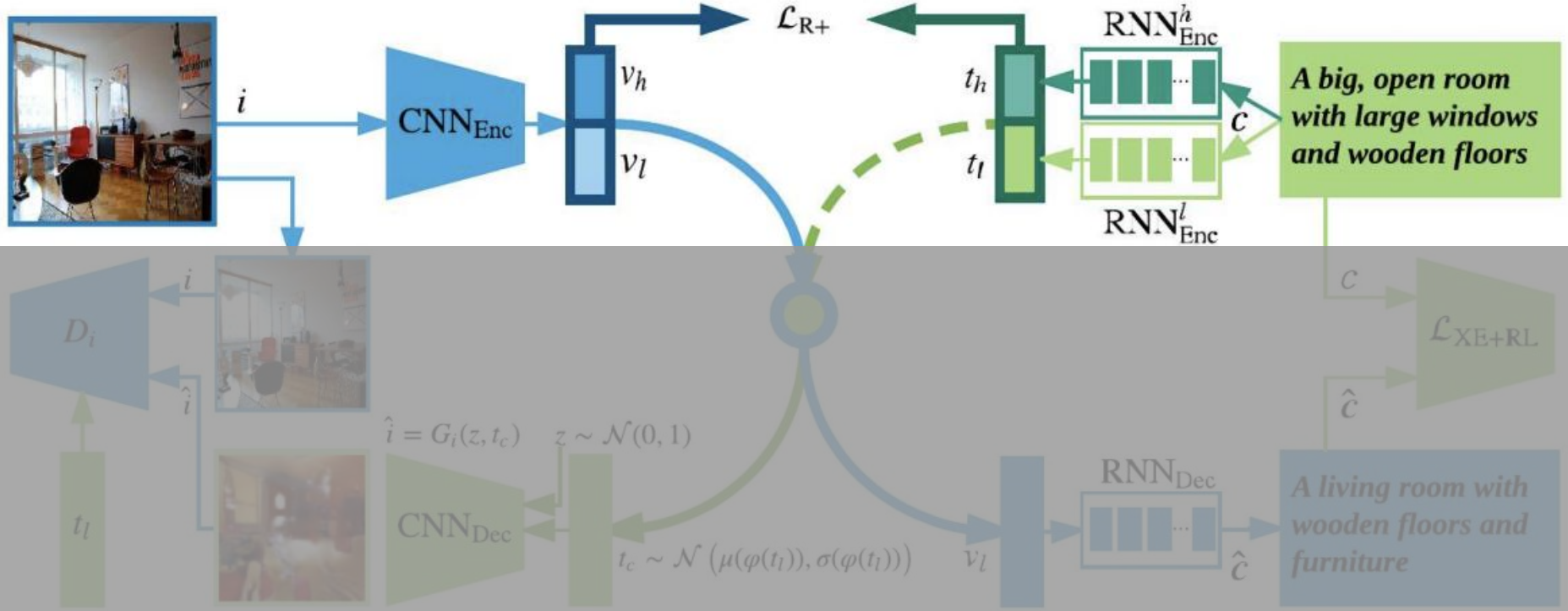


Architecture

- 3 Paths:

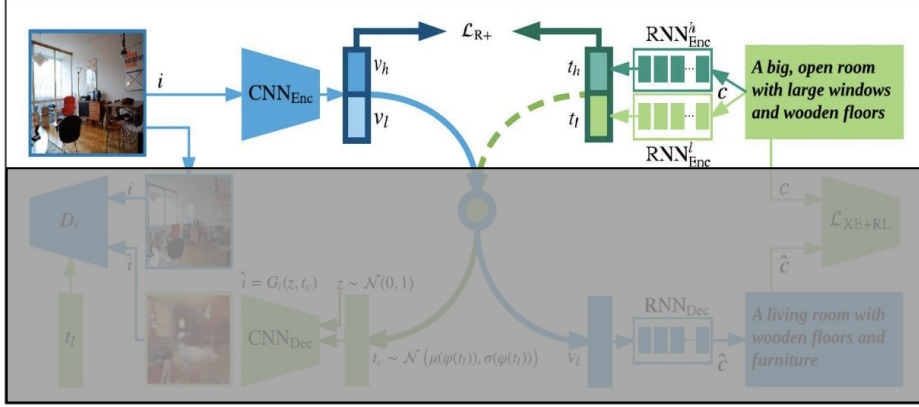


Architecture - Path 1 (Upper part of the figure)



Loss - Path 1

- Global Semantic Similarity

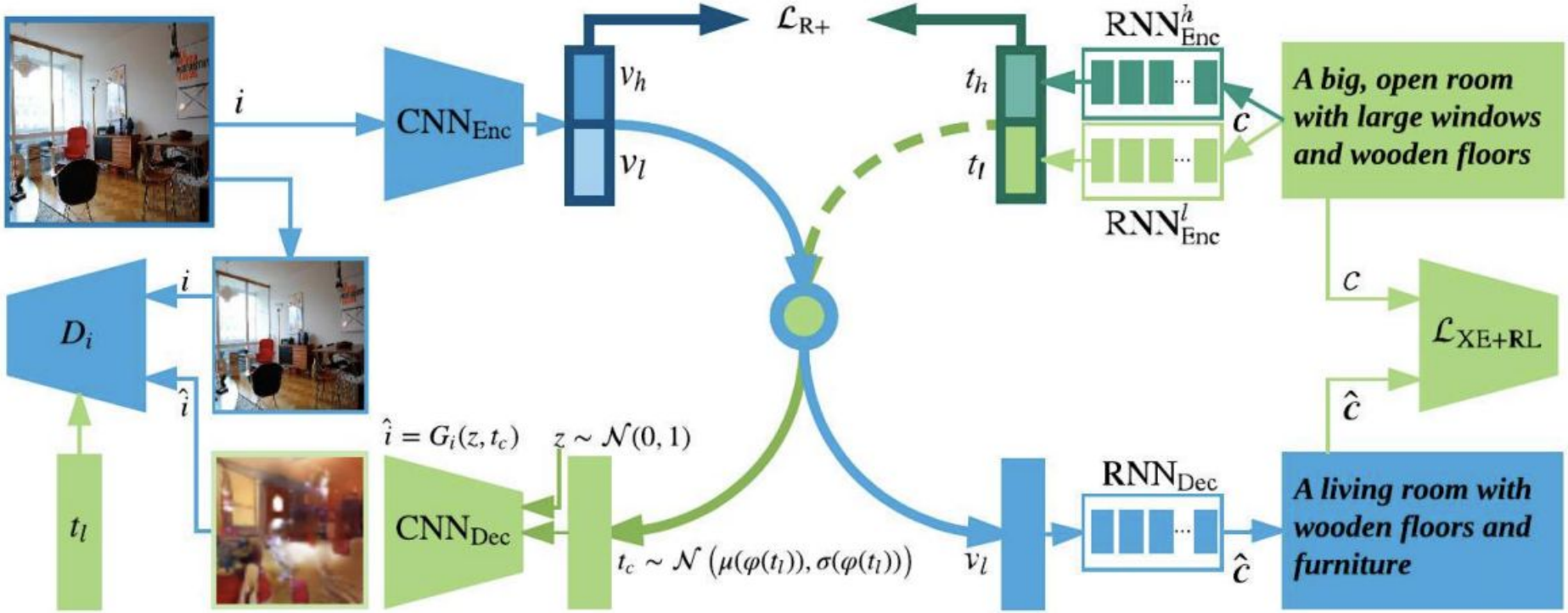


$$\mathcal{L}_{\text{R+}} = \sum_{t'} [\alpha - s^*(t_{h,l}, v_{h,l}) + s^*(t'_{h,l}, v_{h,l})]_+ + \sum_{v'} [\alpha - s^*(t_{h,l}, v_{h,l}) + s^*(t_{h,l}, v'_{h,l})]_+$$

$$\text{where } s^*(t_{h,l}, v_{h,j}) = \lambda s(t_h, v_h) + (1 - \lambda) s(t_l, v_l)$$

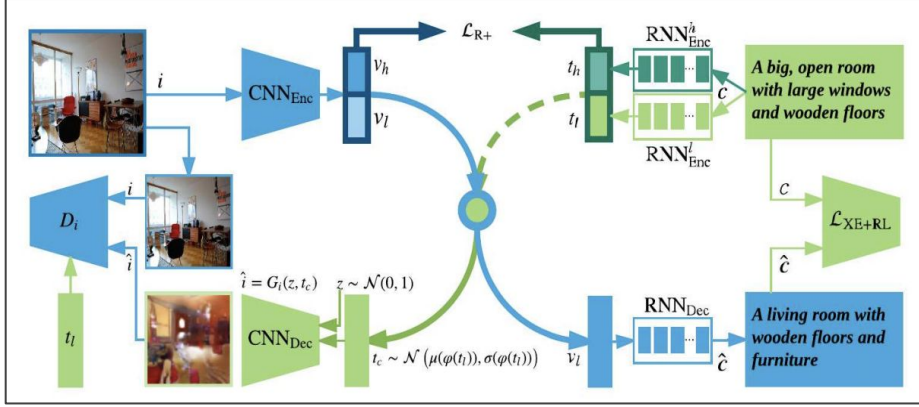
$$s(t, v) = -\| \max(0, v - t) \|^2$$

Architecture - Path 2: image to text (Blue Path)



Loss - Path 2

- Cross Entropy + Similarity Reward



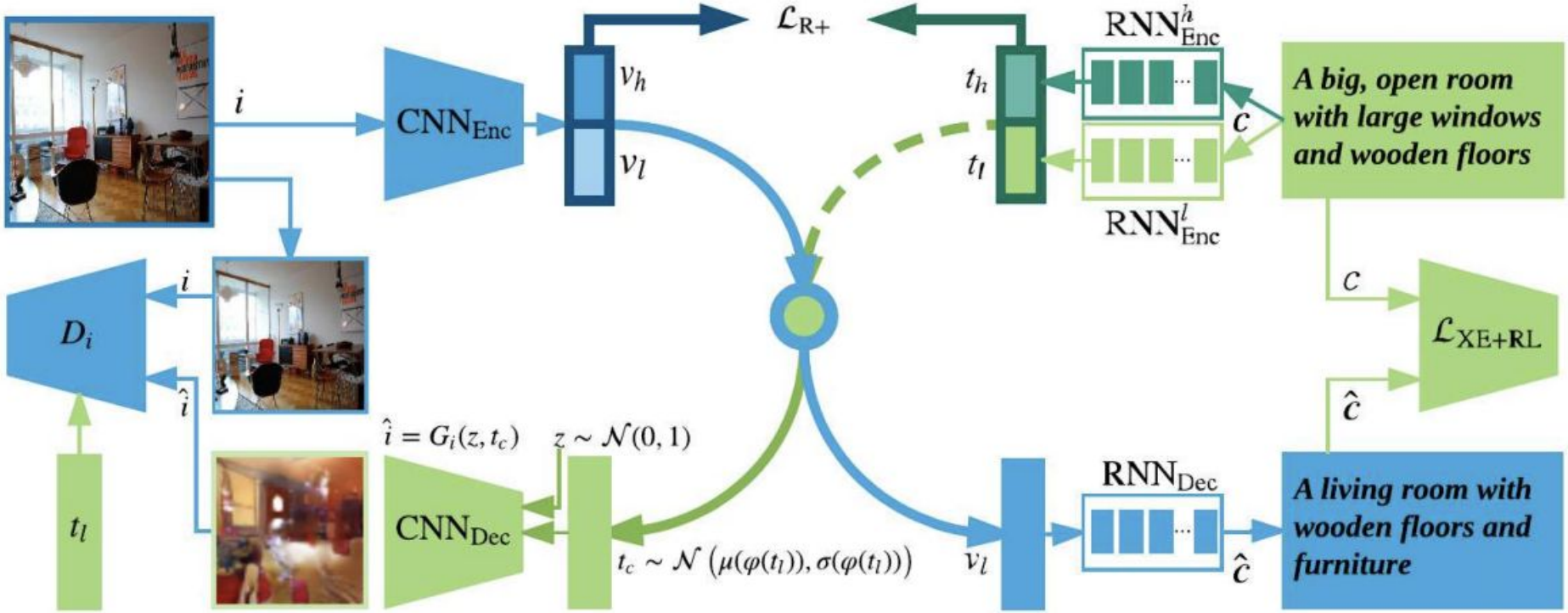
$$\mathcal{L}_{\text{xe}} = -\sum_{t=0}^{T-1} \log p_{\theta_t}(w_t | w_{0:t-1}, v_l; \theta_t)$$

$$\mathcal{L}_{\text{rl}} = -\mathbb{E}_{\tilde{c} \sim p_{\theta_t}}[r(\tilde{c})]$$

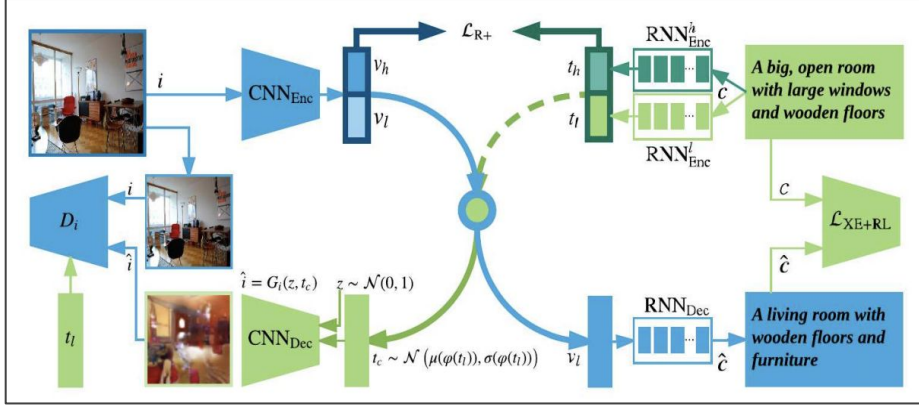
$$\begin{aligned} \nabla_{\theta_t} \mathcal{L}_{\text{rl}} &= -\mathbb{E}_{\tilde{c} \sim p_{\theta_t}}[r(\tilde{c}) \cdot \nabla_{\theta_t} \log p_{\theta_t}(\tilde{c})] \\ &\approx -r(\tilde{c}) \nabla_{\theta_t} \log p_{\theta_t}(\tilde{c}) \\ &\approx -(r(\tilde{c}) - r_b) \nabla_{\theta_t} \log p_{\theta_t}(\tilde{c}) \end{aligned}$$

$$\mathcal{L}_{\text{xe+rl}} = (1 - \gamma)\mathcal{L}_{\text{xe}} + \gamma \mathcal{L}_{\text{rl}}$$

Architecture - Path 3: text to image (Green Path)



Loss - Path 3



- Objective:
  - Distinguish <Real Im, True Cap> from <Real Im, Wrong Cap> and <Fake Im, True Cap>

$$\min_{G_i} \max_{D_i} V(D_i, G_i) = \mathcal{L}_{D_i} + \mathcal{L}_{G_i}.$$

- GAN loss

$$\begin{aligned}\mathcal{L}_{D_i} =& \mathbb{E}_{i\sim p_{\text{data}}}[\log D_i(i, t_l)] + \beta_f \mathbb{E}_{\hat{i}\sim p_G}[\log(1 - D_i(\hat{i}, t_l))] + \\ & \beta_w \mathbb{E}_{i\sim p_{\text{data}}}[\log(1 - D_i(i, t_l'))] \qquad (11)\\ \mathcal{L}_{G_i} =& \mathbb{E}_{\hat{i}\sim p_G}[\log(1 - D_i(\hat{i}, t_l))] + \\ & \beta_s \mathcal{D}_{\text{KL}}(\mathcal{N}(\mu(\varphi(t_l)), \sigma(\varphi(t_l))) \,\|\, \mathcal{N}(0,1)) \qquad (12)\end{aligned}$$

Training

---

while True:

- { Train path2 / Train path1 } Gen_{i2t}-GXN
- { Train path3 / Train path1 } Gen_{t2i}-GXN

Evaluation

- Metrics
  - R@K: percentage of queries where GT matchings are contained in the first K retrievals
  - Med r: median rank of the first retrieved GT matching
- Dataset: MS COCO

Quantitative Results

Table 3: Comparisons of the cross-modal retrieval results on MSCOCO dataset with the state-of-the-art methods. We mark the unpublished work with $*$ symbol. Note that 'Sum' is the summation of the two R@1 scores and the two R@10 scores.

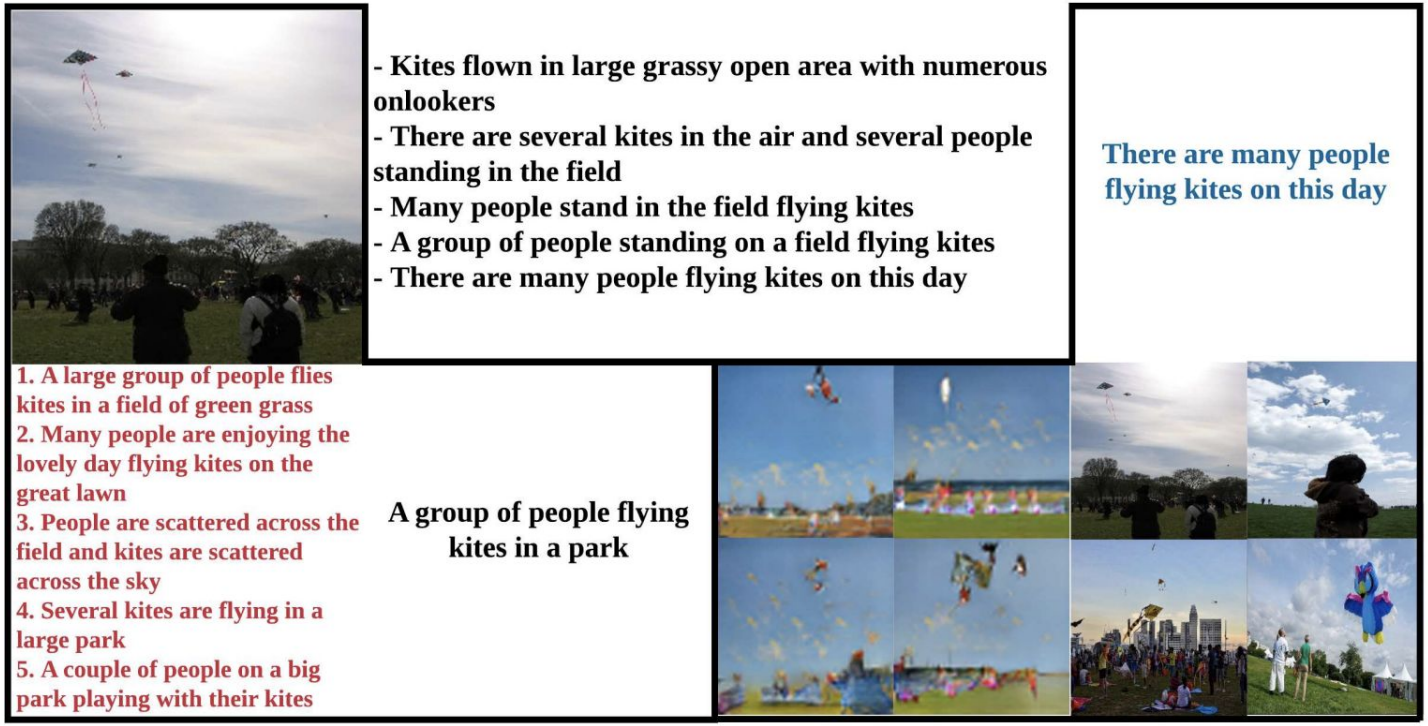
Model	Image-to-Text Retrieval			Text-to-Image Retrieval			Sum
	R@1	R@10	Med r	R@1	R@10	Med r	
	1K Test Images						
m-CNN [19]	42.8	84.1	2.0	32.6	82.8	3.0	242.3
HM-LSTM [24]	43.9	87.8	2.0	36.1	86.7	3.0	254.5
Order-embeddings [36]	46.7	88.9	2.0	38.9	85.9	2.0	260.4
DSPE+Fisher Vector [37]	50.1	89.2	-	39.6	86.9	-	265.8
sm-LSTM [9]	53.2	91.5	**1.0**	40.7	87.4	2.0	272.8
$*$VSE++ (ResNet152, fine-tune) [3]	64.7	95.9	**1.0**	52.0	92.0	**1.0**	304.6
GXN (i2t+t2i)	**68.5**	**97.9**	**1.0**	**56.6**	**94.5**	**1.0**	**317.5**
	5K Test Images						
Order-embeddings [36]	23.3	65.0	5.0	18.0	57.6	7.0	163.9
$*$VSE++ (ResNet152, fine-tune) [3]	41.3	81.2	**2.0**	30.3	72.4	4.0	225.2
GXN(t2i+t2i)	**42.0**	**84.7**	**2.0**	**31.7**	**74.6**	**3.0**	**233.0**

Ablation Study

Table 1: Cross-modal retrieval results on MSCOCO 1K-image test set (bold numbers are the best results).

Model	Image-to-Text			Text-to-Image		
	R@1	R@10	Med	R@1	R@10	Med
GRU(VGG19)	51.4	91.4	**1.0**	39.1	86.7	2.0
GRU_{Bi}(VGG19)	53.6	90.2	**1.0**	40.0	87.8	2.0
GXN(ResNet152)	59.4	94.7	**1.0**	47.0	92.6	2.0
GXN(fine-tune)	64.0	97.1	**1.0**	53.6	94.4	**1.0**
GXN(i2t,xe)	68.2	98.0	**1.0**	54.5	**94.8**	**1.0**
GXN(i2t,mix)	68.4	98.1	**1.0**	55.6	94.6	**1.0**
GXN(t2i)	67.1	**98.3**	**1.0**	56.5	**94.8**	**1.0**
GXN (i2t+t2i)	**68.5**	97.9	**1.0**	**56.6**	94.5	**1.0**

Qualitative Results



Qualitative Results

Query Image

Ground-Truth Captions

- A man wearing a clown wig while riding on skis.
- Two people posing on a mountain wearing skis.
- Group of skiers in colorful outfits on top of a mountain.
- Two people that are standing beside one another while wearing snow skis.
- Two people riding skis at a ski slope

Query Caption

Two people posing on a mountain wearing skis

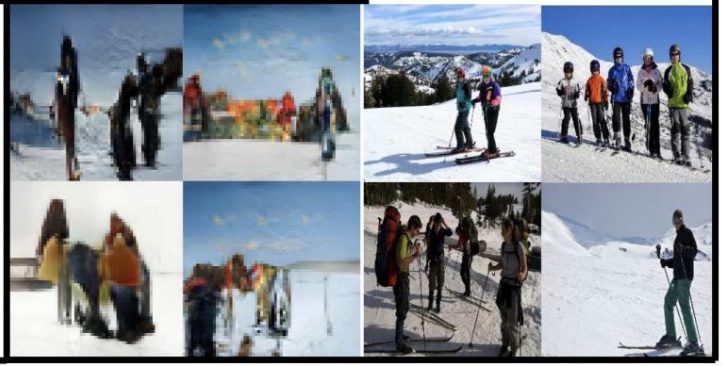
Retrived Captions

1. A couple of people on skis stand on a snowy hill top
2. Two people posing on a mountain wearing skis
3. A couple of cross country skiers on a bright sunny day on mountain
4. Two people standing on a ski slope looking down the hill
5. a pair of skiers on a snowy hillside dressed for cold weather

Generated Caption

Two people standing on skis in the snow

Generated Images



Retrieved Images

Thannks!