NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
School of Computer Science and Engineering

Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models

Jiuxiang Gu[1], Jianfei Cai[1], Shafiq Joty[1], Li Niu[2], Gang Wang[3]
[1] Nanyang Technological University, [2] Rice University, [3] Alibaba AI Labs

Alibaba.com Ali A.I. Labs 阿里人工智能实验室
2018 CVPR SALT LAKE CITY • JUNE 18-22

Overview

- This work focuses on the problem of multi-modal information retrieval, which is to retrieve the images (*resp.* texts) that are relevant to a given textual (*resp.* image) query.
- We incorporate two generative models into the conventional textual-visual feature embedding, which is able to learn concrete *grounded* representations that capture the detailed similarity between the two modalities.
- Our results demonstrate that the combination of the *grounded* and the *abstract* representations can significantly improve the state-of-the-art performance on cross-modal image-caption retrieval.

Challenges

1. The multi-modal data exhibits heterogeneous properties, making it difficult for users to search information of interest effectively and efficiently.
2. The conventional cross-modal feature embedding at the global semantic level are not sufficient to retrieve images with detailed local similarity (e.g., *spatial layout*) or sentences with *word-level* similarity.

Motivation

1. As humans, we can relate a textual (*resp.* image) query to relevant images (*resp.* texts) more accurately, if we pay more attention to the finer details of the images (*resp.* texts).

2. The retrieved results of conventional learning approach (*global semantic level*) in blue color, and our approach (*global semantic level + local similarity*) in yellow color.

Generative Cross-modal Learning Network

The proposed generative cross-modal learning framework (GXN) consists of three training paths: **cross-modal feature embedding** (*the entire upper part*), **image-to-text generative feature learning** (*the blue path*), and **text-to-image generative adversarial feature learning** (*the green path*). It includes *six* networks: two *sentence encoders* $\mathrm{RNN}^h_{\mathrm{Enc}}$ and $\mathrm{RNN}^l_{\mathrm{Enc}}$, one *image encoder* $\mathrm{CNN}_{\mathrm{Enc}}$, one *sentence decoder* $\mathrm{RNN}_{\mathrm{Dec}}$, one *image decoder* $\mathrm{CNN}_{\mathrm{Dec}}$ and one *discriminator* D_i.

1. Considering we have two branches of cross-modal feature embedding, which result in two pairs of cross-modal features: the abstract features (t_h, v_h) and the grounded features (t_l, v_l), where the embedding and mapping of each modality is formulated as:

$$v_k = P_v^k(\mathrm{CNN}_{\mathrm{Enc}}(i;\theta_i)), \quad t_k = P_t^k(\mathrm{RNN}^k_{\mathrm{Enc}}(c;\theta_c^k)), \quad k \in \{h, l\} \tag{1}$$

2. We optimize the improved ranking loss as follows:

$$\mathcal{L}_{\mathrm{R^+}} = \sum_{t'}[\alpha - s^*(t_{h,l}, v_{h,l}) + s^*(t'_{h,l}, v_{h,l})]_+ + \sum_{v'}[\alpha - s^*(t_{h,l}, v_{h,l}) + s^*(t_{h,l}, v'_{h,l})]_+ \tag{2}$$

- where $s^*(t_{h,l}, v_{h,j}) = \lambda s(t_h, v_h) + (1-\lambda)s(t_l, v_l)$ is a combined score with λ being the tradeoff weight, $s(t, v) = -\|\max(0, v-t)\|^2$ is the order-violation penalty used as a similarity, t' and v' denote the representations of the negative samples, $[x]_+$ represents $\max(x, 0)$.

Image-to-text Generative Feature Learning

1. The goal is to encourage the *grounded* visual feature v_l to be able to generate sentences that are similar to the ground-truth captions.
   - We encode the image with $\mathrm{CNN}_{\mathrm{Enc}}$, and then decode the *grounded* visual feature into a sentence with $\mathrm{RNN}_{\mathrm{Dec}}$.
   - We first train our model on a cross-entropy (XE) loss:

$$\mathcal{L}_{\mathrm{xe}} = -\sum_{t=0}^{T-1} \log p_{\theta_t}(w_t | w_{0:t-1}, v_l; \theta_t) \tag{3}$$

   - Then, we directly optimize our model by minimizing the negative expected reward given by:

$$\mathcal{L}_{\mathrm{rl}} = -\mathbb{E}_{\tilde{c} \sim p_{\theta_l}}[r(\tilde{c})] \tag{4}$$

   - The expected gradients can be approximated as:

$$\nabla_{\theta_l}\mathcal{L}_{\mathrm{rl}} \approx -(r(\tilde{c}) - r_b)\nabla_{\theta_l}\log p_{\theta_l}(\tilde{c}) \tag{5}$$

2. To ensure the readability and fluency of the generated caption during the early stage of training, we use a mixture of XE and RL losses:

$$\mathcal{L}_{\mathrm{xe+rl}} = (1-\gamma)\mathcal{L}_{\mathrm{xe}} + \gamma\mathcal{L}_{\mathrm{rl}} \tag{6}$$

   - where γ is a tuning parameter used to balance the two losses.
   - For annealing and faster convergence, we start with optimizing XE loss in Equation (3), and then move to optimizing the joint loss in Equation (6).

Text-to-image Generative Adversarial Feature Learning

1. The goal is to encourage the *grounded* text feature t_l to be able to generate an image that is similar to the ground-truth one.
   - Unlike the image-to-text path, the reverse path suffers from the highly multi-modal distribution of images conditioned on a text representation.
   - We use a conditional GAN to model such a conditional distribution.

$$V(D_i, G_i) = \mathcal{L}_{D_i} + \mathcal{L}_{G_i} \tag{7}$$

   - the discriminator D_i is trained to distinguish the real samples ⟨real image, true caption⟩ from the generated samples of ⟨fake image, true caption⟩ as well as samples of ⟨real image, wrong caption⟩.
2. Directly combining t_l with z cannot produce satisfactory results. This is because of the limited amount of data and the unsmoothness between t_l and z.
   - We improved the discriminator loss $\mathcal{L}_{D_i}$ and the generator loss $\mathcal{L}_{G_i}$ as follows:

$$\mathcal{L}_{D_i} = \mathbb{E}_{i \sim p_{\mathrm{data}}}[\log D_i(i, t_l)] + \beta_f \mathbb{E}_{\hat{i} \sim p_G}[\log(1 - D_i(\hat{i}, t_l))] + \beta_w \mathbb{E}_{i \sim p_{\mathrm{data}}}[\log(1 - D_i(i, t'_l))] \tag{8}$$

$$\mathcal{L}_{G_i} = \mathbb{E}_{\hat{i} \sim p_G}[\log(1 - D_i(\hat{i}, t_l))] + \beta_s \mathcal{D}_{\mathrm{KL}}(\mathcal{N}(\mu(\varphi(t_l)), \sigma(\varphi(t_l))) \,\|\, \mathcal{N}(0,1)) \tag{9}$$

   - t_c is sampled from a Gaussian distribution of $\mathcal{N}(\mu(\varphi(t_l)), \sigma(\varphi(t_l)))$, $\mu(\varphi(t_l))$.
   - t_l and t'_l denote the encoded grounded feature vectors for a matched and a mismatched captions, respectively, i is the matched real image from the true data distribution p_{data}, and $\hat{i} = G_i(z, t_l)$ is the generated image by the generator G_i conditioned on t_l and a noise z.
   - β_f and β_w are the tuning parameters.

Qualitative Results

Visualization of word embedding (GXN (fine-tune) vs. GXN (i2t+t2i))

Quantitative Results

R@K is defined as the percentage of queries in which the ground-truth matchings are contained in the first K retrieved results.

The contribution of each component.

Model	Image-to-Text R@1	Image-to-Text R@10	Image-to-Text Med r	Text-to-Image R@1	Text-to-Image R@10	Text-to-Image Med r
GRU(VGG19)	51.4	91.4	**1.0**	39.1	86.7	2.0
$\mathrm{GRU_{Bi}}$(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**

Comparison with State-of-the-art Methods.

Model	Image-to-Text Retrieval R@1	Image-to-Text Retrieval R@10	Image-to-Text Retrieval Med r	Text-to-Image Retrieval R@1	Text-to-Image Retrieval R@10	Text-to-Image Retrieval Med r	Sum
	MSCOCO 1K Test Images						
Order-embeddings	46.7	88.9	2.0	38.9	85.9	2.0	260.4
DSPE+Fisher Vector	50.1	89.2	-	39.6	86.9	-	265.8
sm-LSTM	53.2	91.5	**1.0**	40.7	87.4	2.0	272.8
∗VSE++ (ResNet152)	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**
	MSCOCO 5K Test Images						
Order-embeddings	23.3	65.0	5.0	18.0	57.6	7.0	163.9
∗VSE++ (ResNet152)	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**
	Flirck30K 1K Test Images						
*VSE++ (ResNet152)	52.9	87.2	**1.0**	39.6	79.5	**2.0**	262.2
GXN (i2t+t2i)	**56.8**	**89.6**	**1.0**	**41.5**	**80.1**	**2.0**	**271.1**

http://www.facebook.com/scse.ntu