Learning Deep Structure-Preserving Image-Text Embeddings

같은 Embedding Space에 Image와 Text 나타내기

Abstract

- Paper: https://arxiv.org/abs/1511.06078
- Github: https://github.com/lwwang/Two_branch_network

- Linear projection + nonlinear activation function으로 구성된 two-branch neural network를 이용해 image와 text의 join embedding을 학습하는 방법 제안
- “structure preservation”을 위한 Loss Function 사용
- Image-to-text, text-to-image retrieval의 accuracy 향상

Instruction

컴퓨터 비전 연구의 동향(2016년 기준)

- 과거: 이미지를 category별로 분류하기
- 현재: 이미지에 대한 설명 생성하기

문제

- 시각 데이터와 텍스트 데이터 간의 similarity를 어떻게 개산할 것인가?
- 텍스트와 이미지를 한 공간에 임베딩하는 방법이 필요하다

목표

- 텍스트와 이미지를 한 공간에 임베딩한다.

Related Work

Canonical Correlation Analysis (CCA, 정준상관분석) 기반

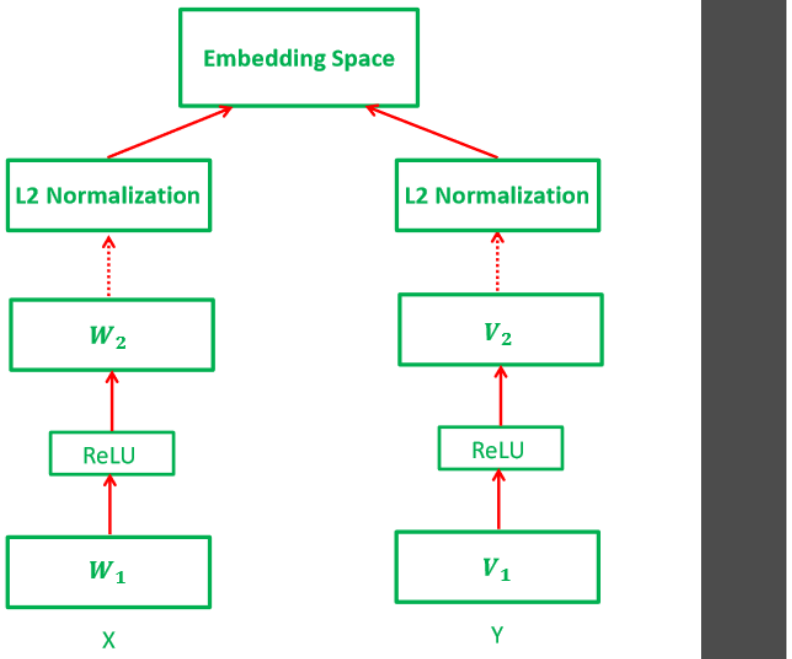
- CCA: 두 변수 집단의 연관성(association)을 각 변수집단에 속한 변수들의 선형결합의 상관계수를 이용하여 설명
- CCA를 이용하여 correlation이 최대가 되는 linear projection을 찾는다

Ranking loss 이용

- Ranking loss: retrieval 등에서 ranking을 잘못 맞출수록 증가하는 loss
- Ranking loss가 감소하는 방식으로 linear transformation을 학습한다

딥러닝 이용

- Boltzmann machine, Autoencoder, LSTM, RNN을 이용하여 non-linear mapping을 하려는 시도가 있었다.



Method (모델 구조)

- X: 이미지 feature vector
- Y: 텍스트 Feature vector
- Feature vector는 pre-trained된 값 사용
- W_1, W_2, V_1, V_2: Matrix. Dense layer 하나로 봐도 무방.
- L2 normalization: 임베딩에 nonlinearity를 주기 위해

Method

Bi-directional ranking constraints

- 같은 category의 이미지 벡터와 텍스트 벡터 간의 거리가 가까워야 한다.
- (x: 이미지, y: 텍스트)

$$d(x_i, y_j) + m < d(x_i, y_k) \quad \forall y_j \in Y_i^+, \forall y_k \in Y_i^- \ . \quad (1)$$

$$d(x_{j'}, y_{i'}) + m < d(x_{k'}, y_{i'}) \quad \forall x_{j'} \in X_{i'}^+, \forall x_{k'} \in X_{i'}^-, \quad (2)$$

Structure-preserving constraints

- 같은 category의 이미지 벡터 간의 거리, 텍스트 벡터 간의 거리가 가까워야 한다.
- 논문에서 새롭게 제안한 방식

$$d(x_i, x_j) + m < d(x_i, x_k) \quad \forall x_j \in N(x_i), \forall x_k \notin N(x_i), \quad (3)$$

$$d(y_{i'}, y_{j'}) + m < d(y_{i'}, y_{k'}) \quad \forall y_{j'} \in N(y_{i'}), \forall y_{k'} \notin N(y_{i'}), \quad (4)$$

Method (Loss Function)

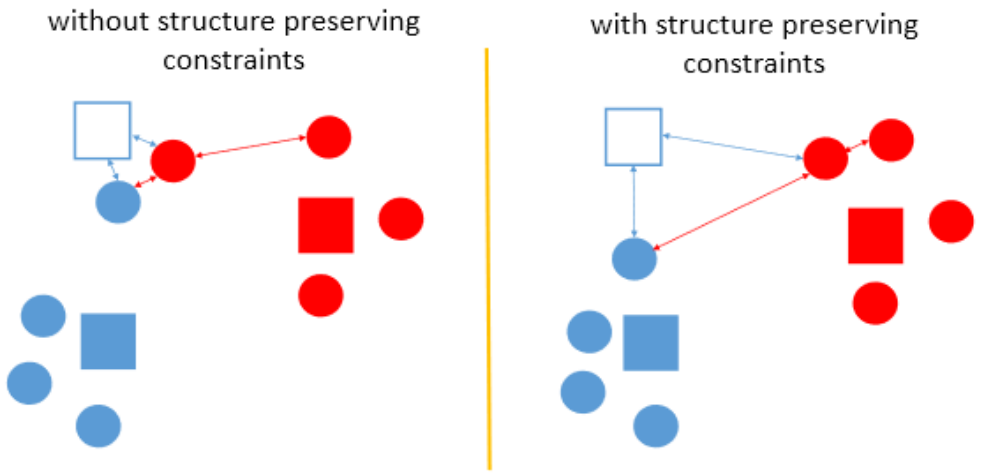


Figure 2. Illustration of the proposed structure-preserving constraints for joint embedding learning (see text). Rectangles represent images and circles represent sentences. Same color indicates matching images and sentences.

$$
\begin{aligned}
L(X,Y) = & \sum_{i,j,k} \max[0, m + d(x_i, y_j) - d(x_i, y_k)] \\
& + \lambda_1 \sum_{i',j',k'} \max[0, m + d(x_{j'}, y_{i'}) - d(x_{k'}, y_{i'})] \\
& + \lambda_2 \sum_{i,j,k} \max[0, m + d(x_i, x_j) - d(x_i, x_k)] \\
& + \lambda_3 \sum_{i',j',k'} \max[0, m + d(y_{i'}, y_{j'}) - d(y_{i'}, y_{k'})] \,,
\end{aligned}
\tag{5}
$$

Method (Sampling)

- Loss function 특성상 Mini-batch를 돌릴 때 data를 잘 골라야 함
  - Target instance
  - Positive match
  - Negative match
- Positive Pair 선택(텍스트-이미지 쌍이 맞는 쌍)
- 각각의 Positive Pair마다 상위 K개의 Negative pair 선택
  - 논문에선 K = 50

Experiments

- Image feature: VGG19의 feature layer
- Sentence Feature: 300차원 Word2Vec을 기반으로 한 Fisher Vector Representation
- Flickr30k, MSCOCO 데이터셋을 이용해 image-sentence retrieval 테스트 진행
  - 기존 모델(CCA 등)에 비해 성능 향상이 있었음
- Flickr30k 데이터셋을 이용해 Phrase Localization 테스트 진행
  - 100개의 EdgeBox region 추출 후 이 모델을 이용
  - 기존 모델(CCA 기반)에 비해 성능 향상이 있었음

	Methods on Flickr30K	Image-to-sentence			Sentence-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
ıe art	Deep CCA [33]	27.9	56.9	68.2	26.8	52.9	66.9
	mCNN(ensemble) [29]	33.6	64.1	74.9	26.2	56.3	69.6
	m-RNN-vgg [31]	35.4	63.8	73.7	22.8	50.7	63.1
	Mean vector [26]	24.8	52.5	64.3	20.5	46.3	59.3
	CCA (FV HGLMM) [26]	34.4	61.0	72.3	24.4	52.1	65.6
	CCA (FV GMM+HGLMM) [26]	35.0	62.0	73.8	25.0	52.7	66.0
	CCA (FV HGLMM) [37]	36.5	62.2	73.3	24.7	53.4	66.8
ctor	Linear + one-directional	33.5	61.7	73.6	21.0	47.4	60.5
	Linear + bi-directional	34.6	64.3	74.9	24.2	52.0	64.2
	Linear + bi-directional + structure	35.2	66.8	76.2	25.6	54.8	66.5
	Nonlinear + one-directional	37.5	65.6	76.9	22.4	50.9	63.3
	Nonlinear + bi-directional	39.3	68.0	78.3	28.1	59.2	71.2
	Nonlinear + bi-directional + structure	**40.3**	**68.9**	**79.9**	**29.7**	**60.1**	**72.1**
tor	Nonlinear + bi-directional	33.5	60.2	71.9	22.8	52.5	65.0
	Nonlinear + bi-directional + structure	35.7	62.9	74.4	25.1	53.9	66.5
	Nonlinear + bi-directional	38.7	66.6	76.9	27.6	57.0	69.0
	Nonlinear + bi-directional + structure	40.1	67.6	78.2	28.1	58.5	69.8

ieval results. The numbers in (a) come from published papers, and the numbers in (b-d) are results of o
tures. Note that the Deep CCA results in [33] were obtained with AlexNet [27]. The results of our
% higher than those of [33] for image-to-sentence retrieval and 1% higher for sentence-to-image retri

	Methods on MSCOCO 1000 testing set	Image-to-sentence			Sentence-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
(a) State of the art	Mean vector [26]	33.2	61.8	75.1	24.2	56.4	72.4
	CCA (FV HGLMM) [26]	37.7	66.6	79.1	24.9	58.8	76.5
	CCA (FV GMM+HGLMM) [26]	39.4	67.9	80.9	25.1	59.8	76.6
	DVSA [22]	38.4	69.9	80.5	27.4	60.2	74.8
	m-RNN-vgg [31]	41.0	73.0	83.5	29.0	42.2	77.0
	mCNN(ensemble) [29]	42.8	73.1	84.1	32.6	68.6	82.8
(b) Fisher Vector	Nonlinear+bi-directional	47.5	77.6	88.3	36.8	72.2	85.6
	Nonlinear+bi-directional+structure	**50.1**	**79.7**	**89.2**	**39.6**	**75.2**	**86.9**
(c) Mean Vector	Nonlinear+bi-directional	39.6	74.0	84.8	32.0	67.3	81.6
	Nonlinear+bi-directional+structure	40.7	74.2	85.3	33.5	68.7	83.2
(d) tf-idf	Nonlinear+bi-directional	45.3	77.6	86.8	35.4	70.2	83.4
	Nonlinear+bi-directional+structure	46.7	77.9	87.7	36.2	72.3	84.7

Table 2. Bidirectional retrieval results on MSCOCO 1000-image test set.

Methods	R@1	R@5	R@10	mAP(all)
CCA baseline	40.11	61.52	67.17	41.96
Our method without negative mining				
(a) $\lambda_1 = 2,\ \lambda_2 = 0,\ \lambda_3 = 0$	35.83	60.51	66.70	40.50
(b) $\lambda_1 = 2,\ \lambda_2 = 0,\ \lambda_3 = 0.1$	36.59	60.44	66.92	40.85
(c) $\lambda_1 = 2,\ \lambda_2 = 0.1,\ \lambda_3 = 0$	**36.74**	60.35	66.73	**41.22**
(d) $\lambda_1 = 2,\ \lambda_2 = 0.1,\ \lambda_3 = 0.1$	36.72	**61.14**	**67.21**	41.13
Fine-tuned with negative mining				
Fine-tuning (a) for 5 epochs	41.77	63.01	68.27	46.55
Fine-tuning (b) for 5 epochs	43.77	64.22	**68.84**	47.38
Fine-tuning (c) for 5 epochs	42.88	63.41	68.47	46.78
Fine-tuning (d) for 5 epochs	**43.89**	**64.46**	68.66	**47.72**

results on Flickr30K Entities using Fast-RCNN features. We use 100 EdgeBox
6.91.

Experiments

Conclusion

- Two-branch network를 이용하여 Text와 Image 임베딩을 같은 space에 나타낼 수 있다.
- Bi-directional ranking에다 structure-preserving을 추가한 loss 함수를 추가하여, 같은 category에 속하는 Text/Image끼리 더 잘 뭉치게 되었다.
- Flick30K, MSCOCO 기반으로 한 retrieval 결과는 상당한 향상이 있었다.

References

- https://arxiv.org/abs/1511.06078
- http://stat.snu.ac.kr/time/download/7.%EC%A0%95%EC%A4%80%EB%B6%84%EC%84%9D.pdf