Learning Deep Structure-Preserving Image-Text Embeddings

Liwei Wang Yin Li Svetlana Lazebnik

Presented by: Arjun Karpur

Outline

- Problem Statement
- Approach
- Evaluation
- Conclusion

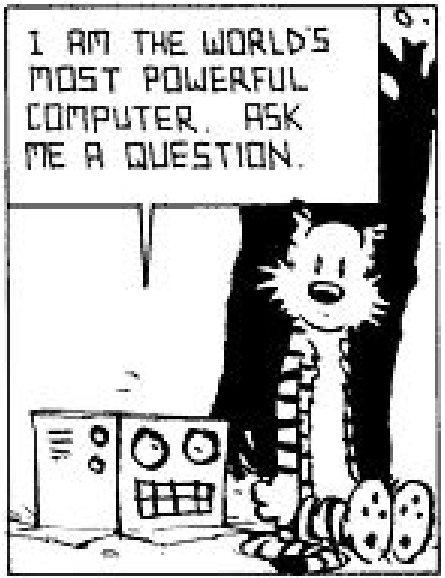


Image courtesy of: http://calvinandhobbes.wikia.com/

Problem Statement

Problem Statement

- Given collection of images, sentences
- Perform retrieval tasks...
  - Image-to-text
  - Text-to-image

“The quick brown fox jumped over the lazy dog”

Image courtesy of: http://nebraskaris.com/

Problem Statement

- Given collection of images, sentences
- Perform retrieval tasks...
  - Image-to-text
  - Text-to-image
- Useful for…
  - Image captioning
  - Visual question answering
  - etc...

“The quick brown fox jumped over the lazy dog”

Image courtesy of: http://nebraskaris.com/

Problem Statement

- Given collection of images, sentences
- Perform retrieval tasks...
  - Image-to-text
  - Text-to-image
- Useful for…
  - Image captioning
  - Visual question answering
  - etc...
- Utilize '**joint embedding**' to compare differing modalities

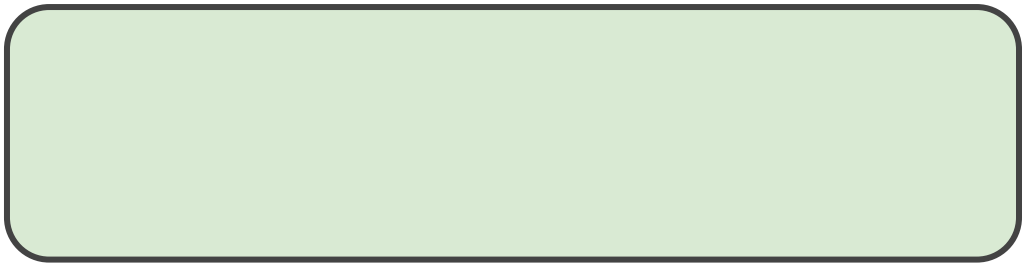
"The quick brown fox jumped over the lazy dog"

Image courtesy of: http://nebraskaris.com/

Joint Embedding

The dog plays in the park.

The student reads in the library



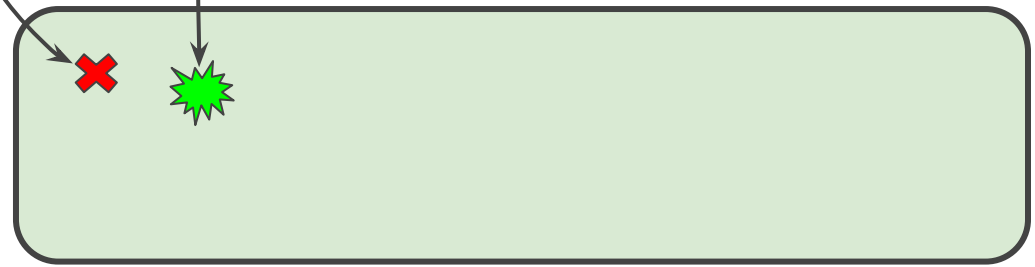
Embedding space

Images courtesy of: https://www.wikipedia.org

Joint Embedding

The dog plays in the park.

The student reads in the library



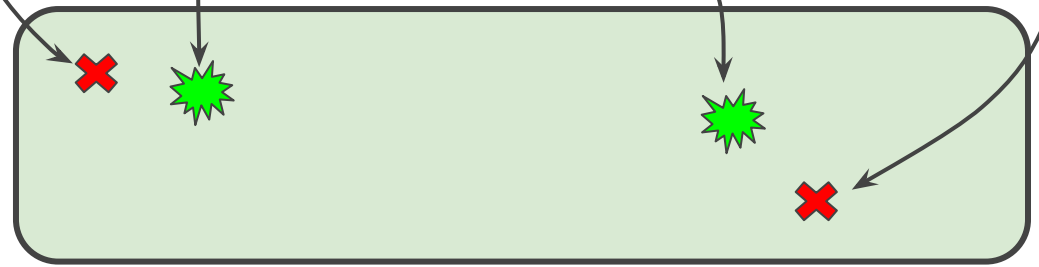
Embedding space

Images courtesy of: https://www.wikipedia.org

Joint Embedding

The dog plays in the park.

The student reads in the library



Embedding space

Images courtesy of: https://www.wikipedia.org

Approach

Approach

- **Multi-view shallow network** to project existing representations into embedding space
  - Any existing handcrafted or learned
  - One branch for each data mode

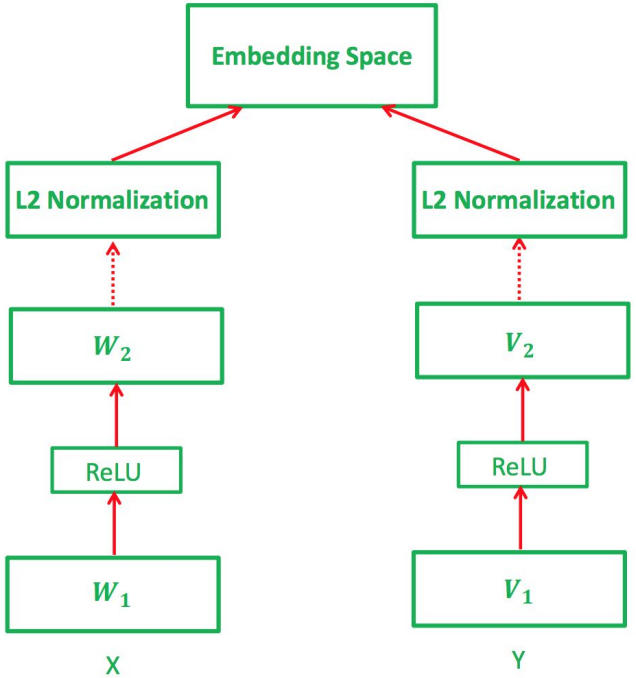


Image courtesy of: Wang et. al 2016

Approach

- **Multi-view shallow network** to project existing representations into embedding space
  - Any existing handcrafted or learned
  - One branch for each data mode
- **Nonlinearities** allow modeling of more complex functions

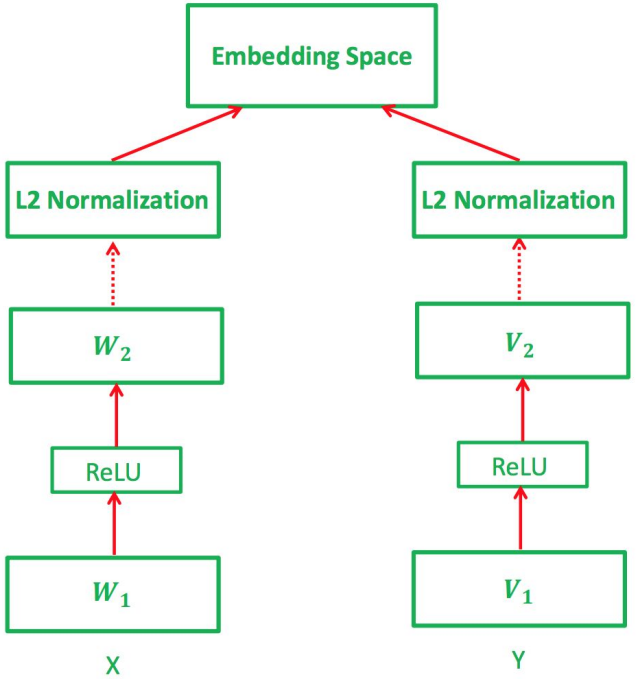


Image courtesy of: Wang et. al 2016

Approach

- **Multi-view shallow network** to project existing representations into embedding space
  - Any existing handcrafted or learned
  - One branch for each data mode
- **Nonlinearities** allow modeling of more complex functions
- Improve accuracy via **L2 normalization** before embedding loss

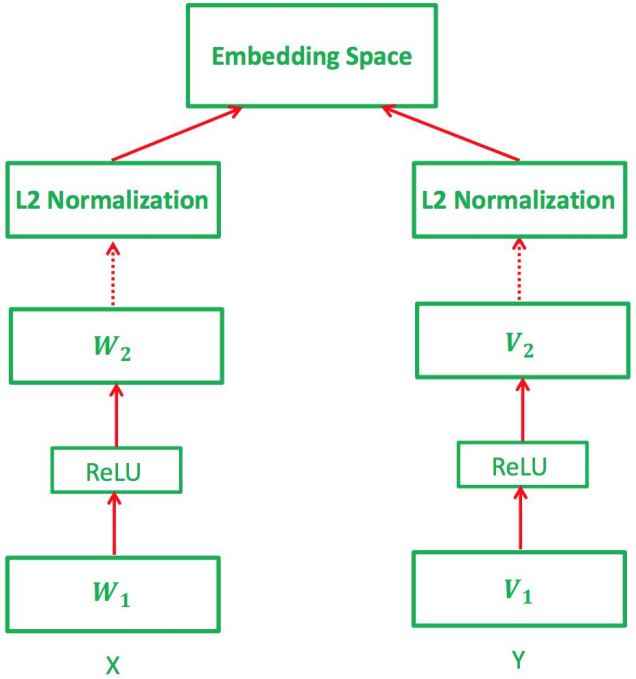
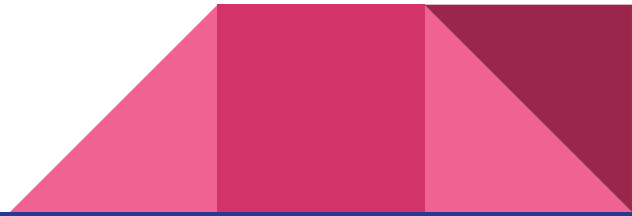


Image courtesy of: Wang et. al 2016

Training Objective

- Loss function comprising of…

    a. **Bi-directional ranking constraints** - encourage short distances between an image/sentence and its positive matches and large distances between image/sentence and negatives
        - Cross-view matching



Training Objective

- Loss function comprising of…

    a. **Bi-directional ranking constraints** - encourage short distances between an image/sentence and its positive matches and large distances between image/sentence and negatives
        - Cross-view matching

    b. **Structure-preserving constraints** - images (and sentences) with identical semantic meanings are separated from others by some margin
        - Within-view matching

Bi-directional Ranking Constraints

- Given a training image x_i, let Y_i^- and Y_i^+ represent its matching and non-matching sentences

Bi-directional Ranking Constraints

- Given a training image x_i, let Y_i^- and Y_i^+ represent its matching and non-matching sentences
- Want distance between x_i and $\forall y_j \in Y_i^+$ to be less than distance between x_i and $\forall y_k \in Y_i^-$ by some margin m...

Bi-directional Ranking Constraints

- Given a training image x_i, let Y_i^- and Y_i^+ represent its matching and non-matching sentences
- Want distance between x_i and $\forall y_j \in Y_i^+$ to be less than distance between x_i and $\forall y_k \in Y_i^-$ by some margin m...

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

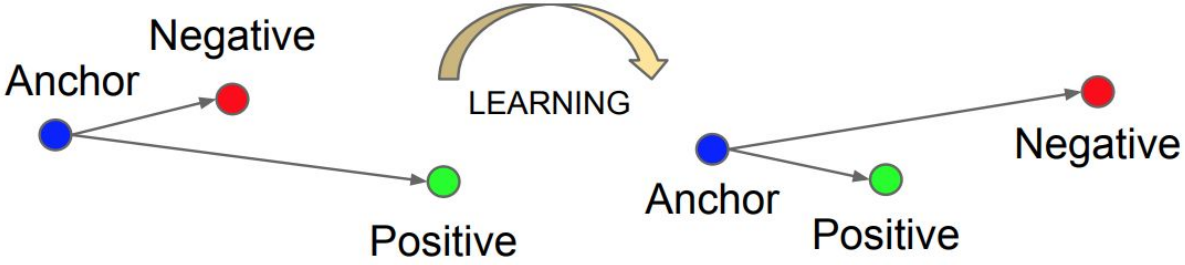


Image courtesy of: FaceNet [Schroff et. al]

Structure-preserving Constraints

- Neighborhood $N(x_i)$ of images (or sentences - same modality) with shared meaning

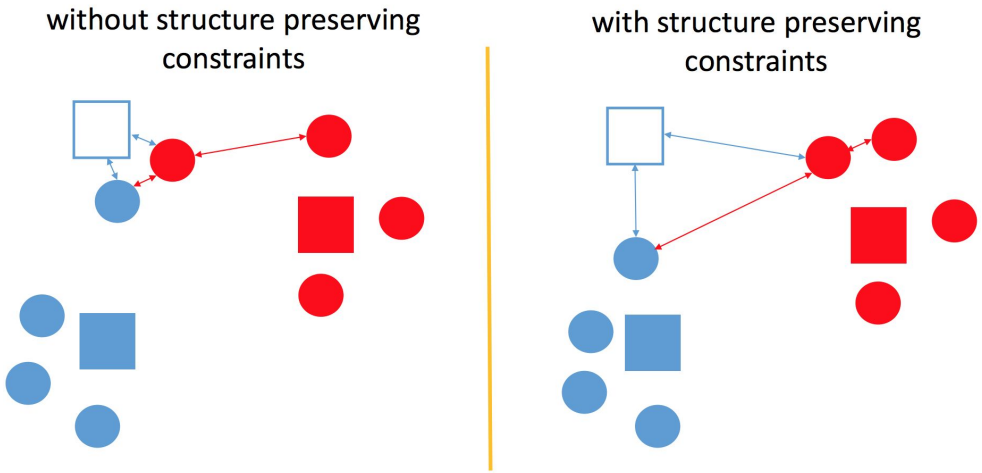


Image courtesy of: Wang et al 2016

Structure-preserving Constraints

- Neighborhood $N(x_i)$ of images (or sentences - same modality) with shared meaning
- Enforce margin between $N(x_i)$ and points outside

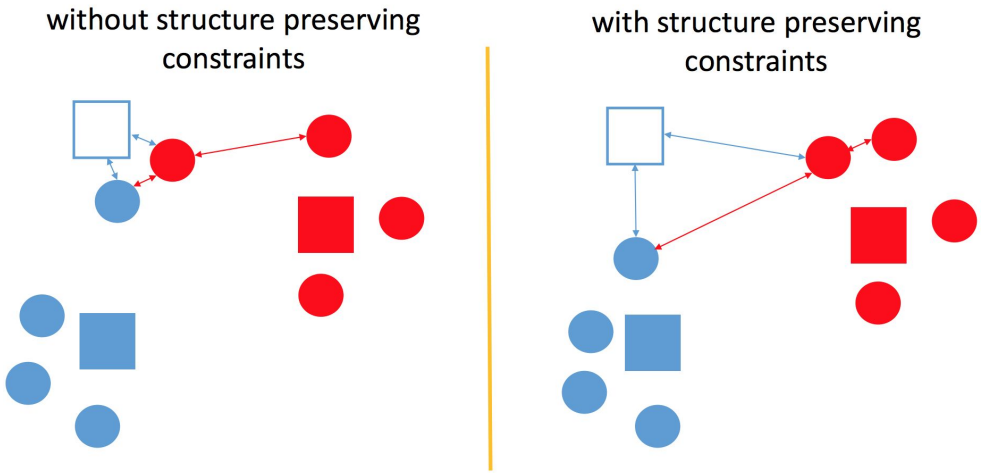


Image courtesy of: Wang et al 2016

Structure-preserving Constraints

- Neighborhood $N(x_i)$ of images (or sentences - same modality) with shared meaning
- Enforce margin between $N(x_i)$ and points outside
- Remove ambiguity for a query image/sentence

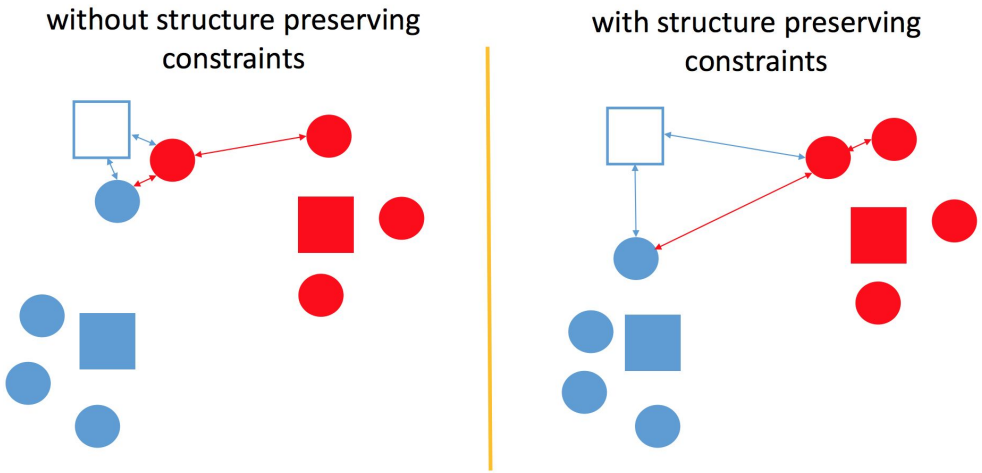


Image courtesy of: Wang et al 2016

Loss Function

$$L(X,Y) = \sum_{i,j,k} \max[0, m + d(x_i, y_j) - d(x_i, y_k)]$$

$$+\lambda_1 \sum_{i',j',k'} \max[0, m + d(x_{j'}, y_{i'}) - d(x_{k'}, y_{i'})]$$

$$+\lambda_2 \sum_{i,j,k} \max[0, m + d(x_i, x_j) - d(x_i, x_k)]$$

$$+\lambda_3 \sum_{i',j',k'} \max[0, m + d(y_{i'}, y_{j'}) - d(y_{i'}, y_{k'})]\,,$$

Cross-view

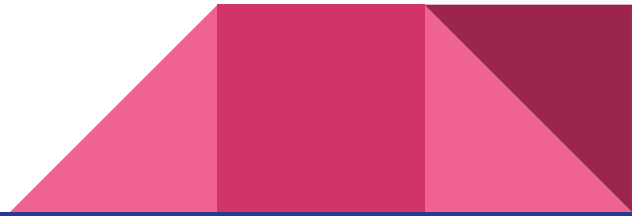
Within-view

Use ‘triplet sampling’ to efficiently train, given nearly infinite triplets

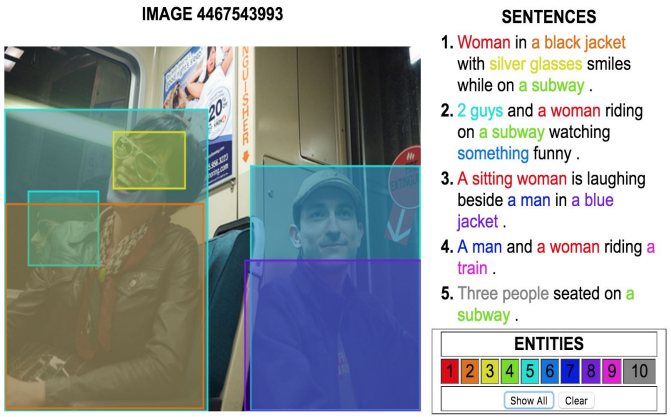
Evaluation

Evaluation

- Evaluate image-to-sentence and sentence-to-image retrieval
- Datasets
  - **Flickr30K** - 31783 images, each described by 5 sentences
  - **MSCOCO** - 123000 images, each described by 5 sentences
- Perform Recall@K (K = 1,5,10) for 1000 test images and corresponding sentences



Datasets - Flickr30k



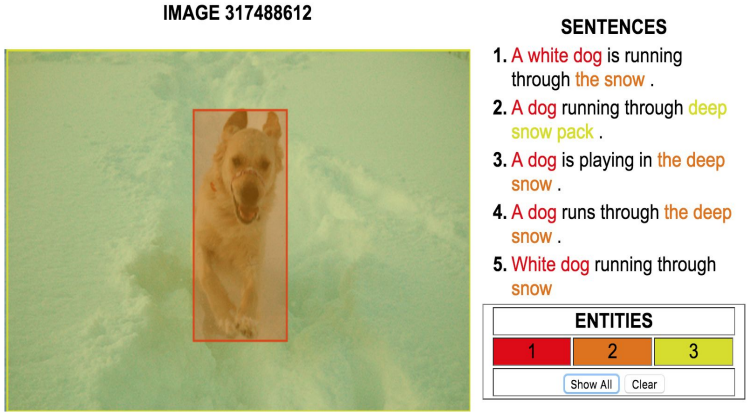


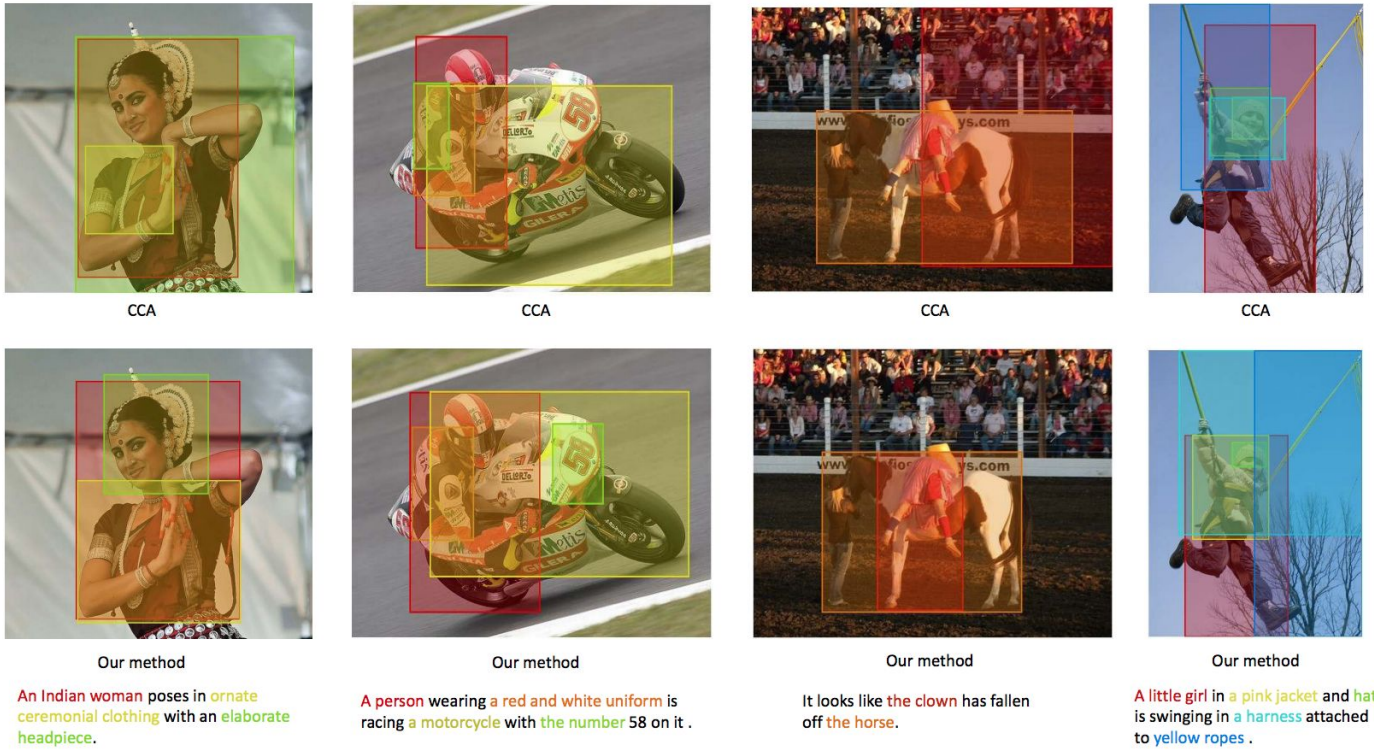
Image courtesy of: http://web.engr.illinois.edu/~bplumme2/Flickr30kEntities/

	Methods on Flickr30K	Image-to-sentence			Sentence-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
(a) State of the 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
(b) Fisher vector	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**
(c) Mean vector	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
(d) tf-idf	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

	Methods on Flickr30K	Image-to-sentence			Sentence-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10
(a) State of the 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
(b) Fisher vector	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**
(c) Mean vector	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
(d) tf-idf	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

Quantitative Results - Recap

- Using joint-loss, fine-tuning method on top of handcrafted feature outperforms deep methods
- All components of loss function contribute to good results



Compared to baselines, achieve high results even without focusing on object detection

Image courtesy of: Wang et al 2016

Conclusion

Strengths & Weaknesses

+	-
• Works with **any pre-existing embedding** (finetune or train from scratch) • Robust **2-way** embedding method • **L2 normalization** allows for easy Euclidean distance comparisons	• Hard to find a **single sentence** that describes **multiple images** (or vice versa) • Only allows for **retrieval, not synthesis** (image captioning) • Requires **large collection** of labeled pairs

Extensions

- Use framework for other data pairs in **different modalities** (audio + video)
- Leverage data pairs that arise naturally in the world for **unsupervised learning**

References

- Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. APA
- Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
- Various image sources...

Comments + Questions