Multimodal Learning

Veronika Krejčířová

Introduction

- Who am I?
- What am I going to talk about?
- Whyyy?

Overview

- Intro - who, what, why
- Multimodal learning
- Text and image learning
  - The 3 main approaches
    - Principle
    - Example
    - More examples
- Conclusion

Definition

“Multimodal data mining refers to analyzing more than one form of data for discovering hidden patterns.” [1]

- can integrate text, images, video, audio, sensor data or structured information
- to derive and validate insights none of which may be possible to obtain from any single source
- has been a popular research area since 2006

Examples

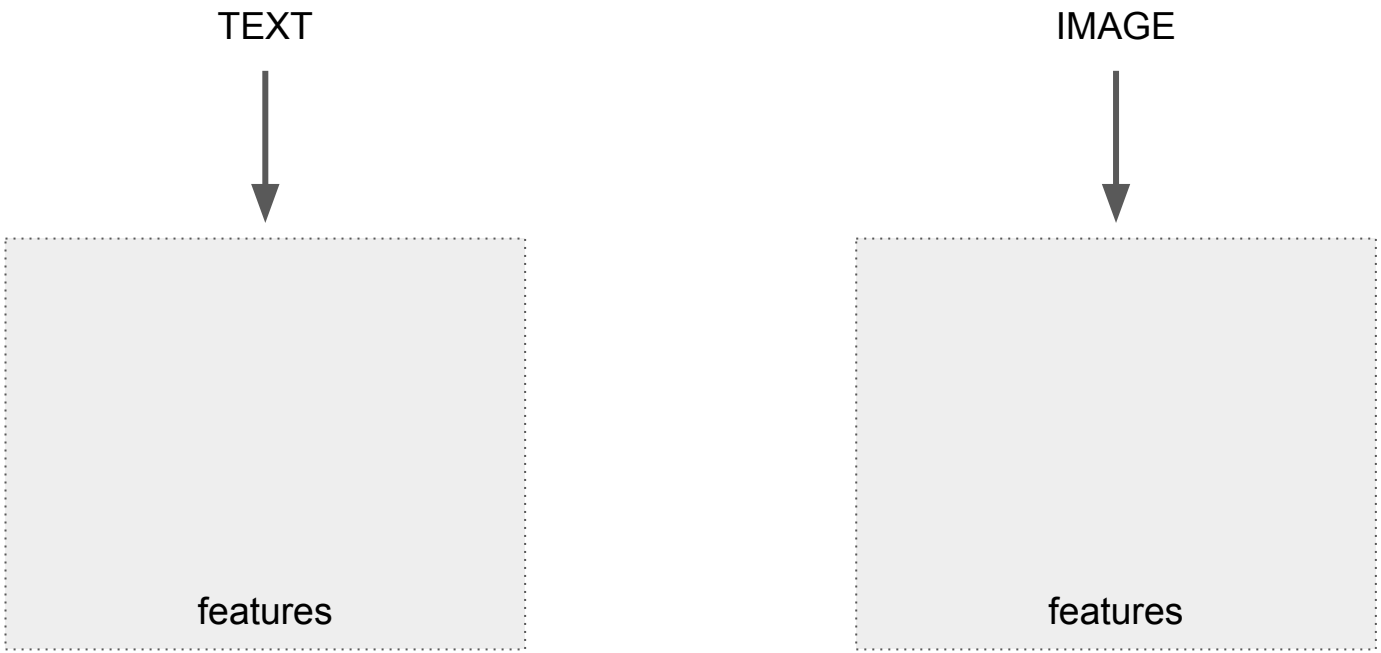
- News - texts and images
- Speach recognition - voice recording and video
- Business - sales data/market share reports and text data about organization or its products.
- Medicine - patient data, demographic information and imaging modalities or genomic-related tests.

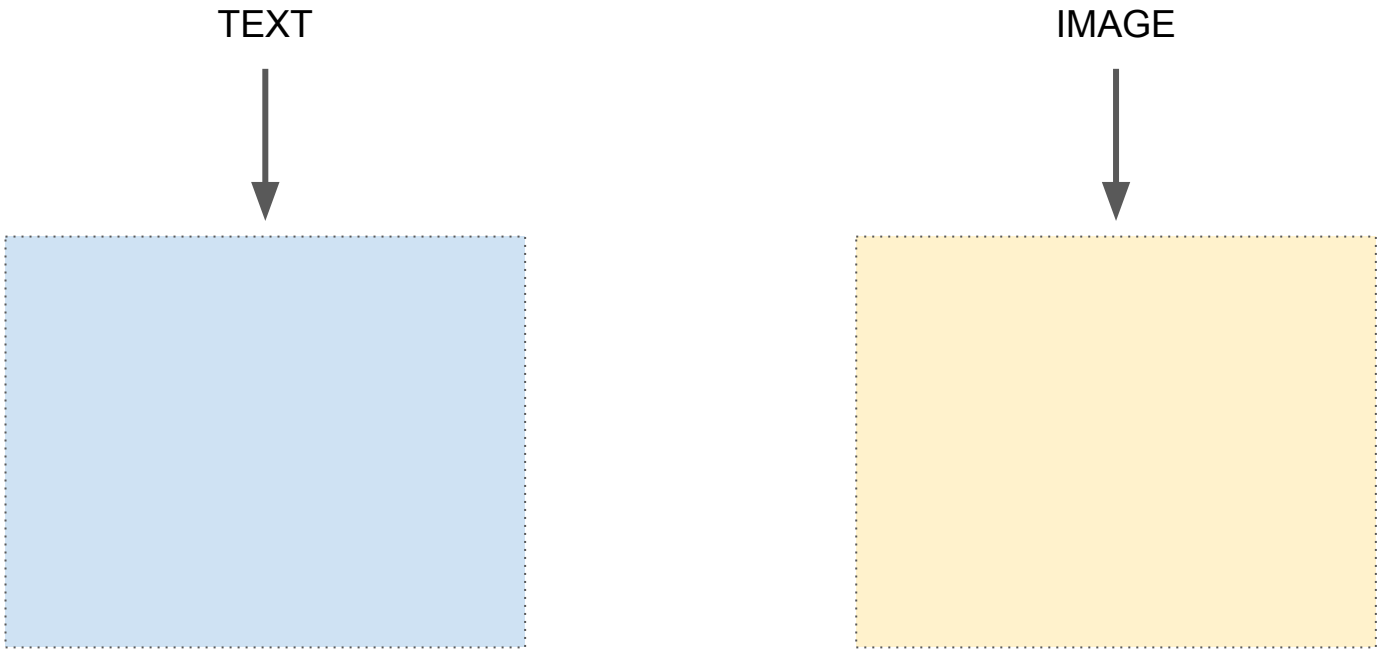
“The major advantage of using multiple data sources is to be able to integrate multiple perspectives about the same event.”

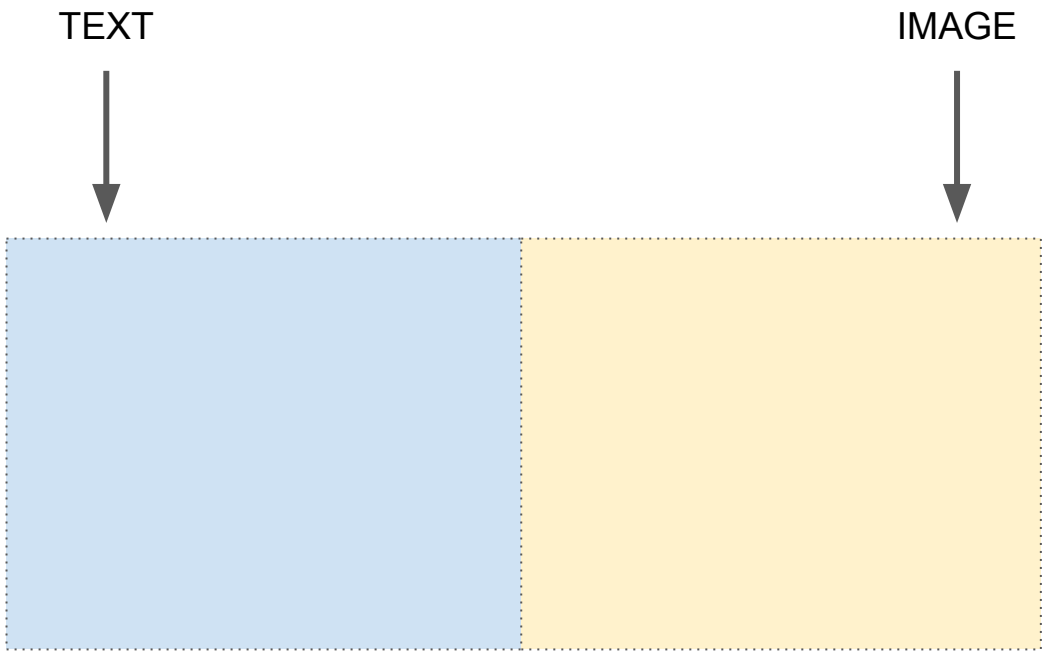
Joint Text and Image Mining

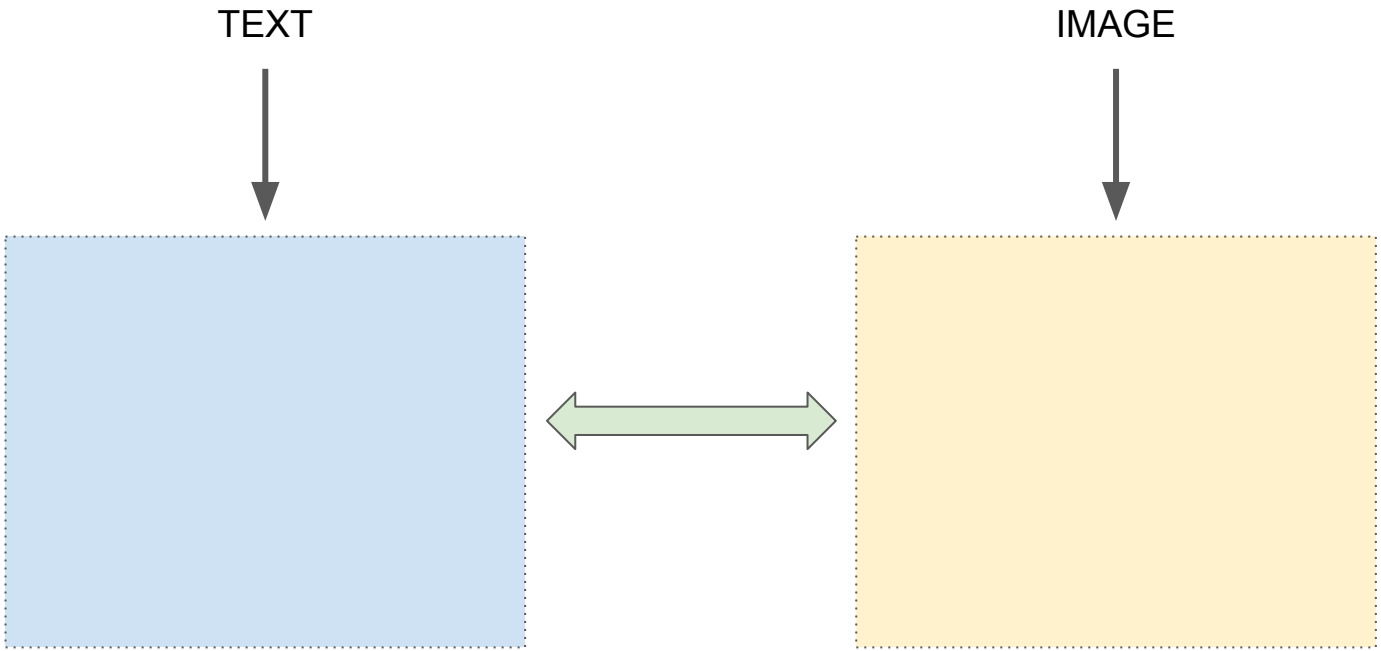
TEXT

IMAGE


TEXT
IMAGE
features
features


TEXT
IMAGE


TEXT
IMAGE


TEXT
IMAGE


TEXT
IMAGE

Motivation

- Human learning
- Concept - grouned knowledge
- Text and images contain complementary information

Approaches

1. Fusion of Embeddings
2. Canonical Correlation Analysis
3. Deep Learning

Fusion approach

Visual Information in Semantic Representation

Yansong Feng and Mirella Lapata (2010)

- big scale, fully automatic approach (without human involvement)
- based on topic models (= documents are mixture of topics)
- assume that the images and texts have been generated by shared topics
- LDA topic modeling

Michelle Obama fever hits the UK

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact. She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase. Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

Fusion approach

Visual Information in Semantic Representation

Image processing - SIFT + K-means algorithm

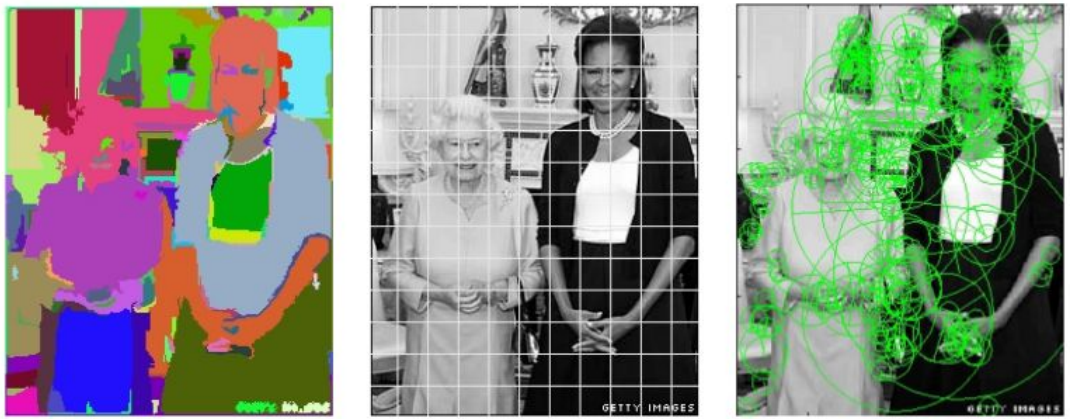


Figure 1: Image partitioned into regions of varying granularity using (a) the normalized cut image segmentation algorithm, (b) uniform grid segmentation, and (c) the SIFT point detector.

Fusion approach

Visual Information in Semantic Representation

Experiments and results

- word similarity and word association

Model	Word Association	Word Similarity
UpperBnd	0.400	0.545
MixLDA	0.123	0.318
TxtLDA	0.077	0.247

- three observations

(Kernel) Canonical Correlation Analysis approach

- CCA: a way of inferring information from cross-covariance matrices,
- finds maximally correlated linear subspaces (manifolds)
- these manifolds are seen as common space, where each document is represented by the projections of its features
- kCCA: removes linearity constraint by using “kernel trick”

(Kernel) Canonical Correlation Analysis approach

Aggregating Image and Text Quantized Correlated Components

Thi Quynh Nhi Tran, Hervé Le Borgne, Michel Crucianu (2016)

- kCCA problem: coarse association between modalities

Average Distance	Pascal VOC07	FlickR 8K
$d_{\text{intramodality}}(I)$	1.18 ± 0.16	1.17 ± 0.13
$d_{\text{intramodality}}(T)$	1.11 ± 0.19	0.75 ± 0.13
$d_{\text{intermodality}}(\text{sample})$	1.39 ± 0.07	1.02 ± 0.12
$d_{\text{intermodality}}(\text{overall})$	1.42 ± 0.06	1.28 ± 0.10

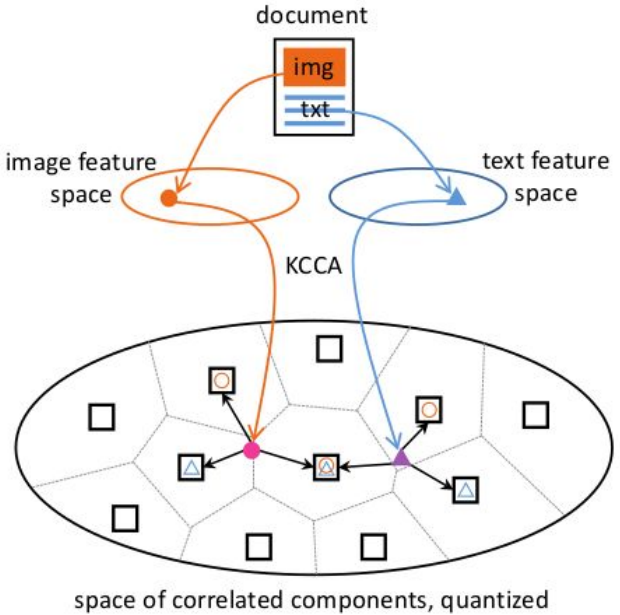
Table 1. Average distances between projections on KCCA space.

(Kernel) Canonical Correlation Analysis approach

Aggregating Image and Text Quantized Correlated Components

- solution: MACC algorithm:

1. Codebook learning: k-means clustering on mixture of features
2. Each document is then encoded by its differences from nearest codewords
3. Aggregation by sum pooling



(Kernel) Canonical Correlation Analysis approach

Aggregating Image and Text Quantized Correlated Components

- image features: VGG-Net, 4096 dim, L2 norm
- text features: Word2Vec, 300 dim, L2 norm
- Results:

Approach	R@1	R@5	R@10
Socher *et al.* [25]	6.1	18.5	29
Hodosh *et al.* [12]	7.6	20.7	30.1
Karpathy *et al.* [17]	11.8	32.1	44.7
Chen *et al.* [4]	17.3	42.5	57.4
KCCA(VGG+W2V)	26.1	53.7	65.6
MACC	**27.6**	**55.6**	**69.4**

Table 7. Image retrieval results on FlickR 8K.

Approach	R@1	R@5	R@10
Socher *et al.* [25]	8.9	29.8	41.1
Karpathy *et al.* [17]	15.2	37.7	50.5
Chen *et al.* [4]	18.5	45.7	58.1
MACC (F8k)	**33.9**	**65.6**	**77.5**
MACC (F30k)	**35.3**	**66.0**	**78.2**

Table 8. Image retrieval results on FlickR 30K. MACC parameters are cross-validated on FlickR 8k (F8k) or FlickR 30k (F30k)

(Kernel) Canonical Correlation Analysis approach

Canonical correlation analysis: An overview with application to learning methods
David R. Hardoon, Sandor Szedmak, John Shawe-Taylor (2004)

Cross-Modal Image Clustering via Canonical Correlation Analysis
Cheng Jin, Wenhui Mao, Ruiqi Zhang, Yuejie Zhang, Xiangyang Xue (2015)

Connecting Modalities: Semi-supervised Segmentation and Annotation of Images Using Unaligned Text Corpora
Richard Socher, Li Fei-Fei (2010)

A Correlation Approach for Automatic Image Annotation
David R. Hardoon, Craig Saunders, Sandor Szedmak, John Shawe-Taylor (2006)

Deep Learning approach

Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

Andrej Karpathy, Armand Joulin, Li Fei-Fei (2014)

Uses:

- fragments of images = objects detected using RCNN
- fragments of sentences = dependency tree relations
- inner product to count fragment similarity

Deep Learning approach

Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

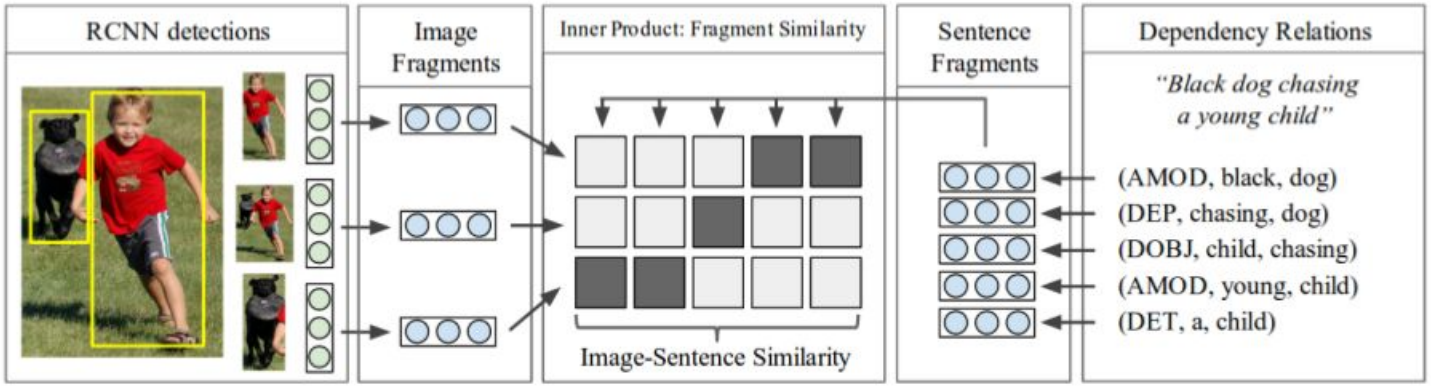


Figure 2: Computing the Fragment and image-sentence similarities. **Left:** CNN representations (green) of detected objects are mapped to the fragment embedding space (blue, Section 3.2). **Right:** Dependency tree relations in the sentence are embedded (Section 3.1). Our model interprets inner products (shown as boxes) between fragments as a similarity score. The alignment (shaded boxes) is latent and inferred by our model (Section 3.3.1). The image-sentence similarity is computed as a fixed function of the pairwise fragment scores.

Deep Learning approach

Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

Objective function is sum of:

- Fragment Alignment Objective - on fragments level
- Global Ranking Objective - on sentence/image level

Problem: not every image object is mention in the caption and vice versa

Solution: Multiple Instance Learning

Deep Learning approach

Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

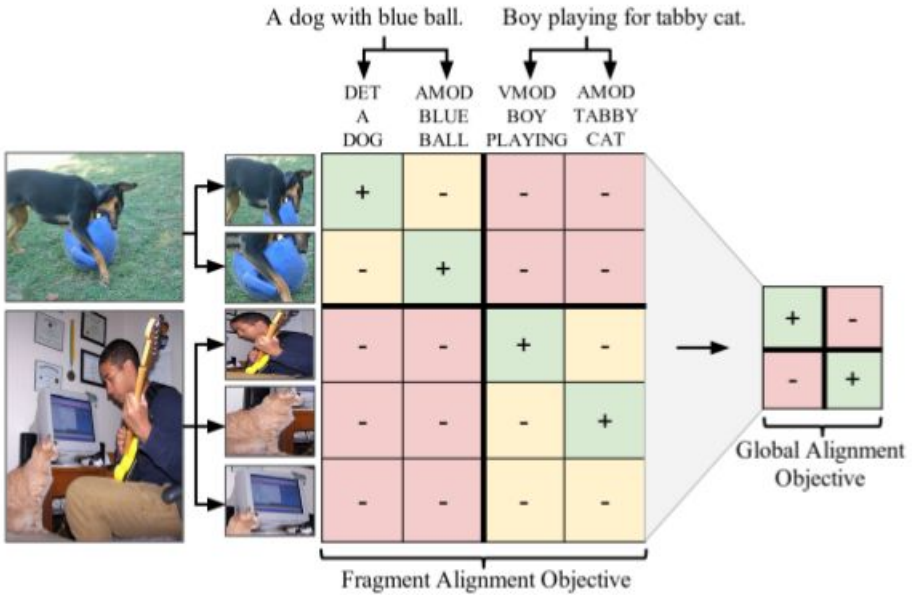


Figure 3: The two objectives for a batch of 2 examples. **Left:** Rows represent fragments v_i, columns s_j. Every square shows an ideal scenario of $y_{ij} = sign(v_i^T s_j)$ in the MIL objective. Red boxes are $y_{ij} = -1$. Yellow indicates members of positive bags that happen to currently be $y_{ij} = -1$. **Right:** The scores are accumulated with Equation 6 into image-sentence score matrix S_{kl}.

Deep Learning approach

Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

Deep Learning approach

Deep Fragment Embeddings for Bidirectional Image Sentence Mapping

Deep Learning approach

Multimodal Learning with Deep Boltzmann Machines
Nitish Srivastava, Ruslan Salakhutdinov (2012)

Learning language through pictures
Grzegorz Chrupała, Ákos Kádár, Afra Alishahi (2015)

Imagined Visual Representations as Multimodal Embeddings
Guillem Collell, Ted Zhang, Marie-Francine Moens (2017)

Learning Deep Structure-Preserving Image-Text Embeddings
Liwei Wang, Yin Li, Svetlana Lazebnik (2016)

Conclusion

Multimodal learning heavily depends on the image and text representations and the selected learning method

Neural network approach dominates state-of-the-art.

Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world

Angeliki Lazaridou and Elia Bruni and Marco Baroni (2014)

Sources

[1] *Shaudhury, S., Dey, L., Verma, I., and Hassan E.* (2017). **Mining Multimodal Data**. Pattern Recognition and Big Data: pp. 581-604.
[2] *Feng, Y., Lapata, M*. (2010). **Visual information in semantic representation**. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, p.91-99. doi:10.1109/cvpr.2016.225
[3] T*ran, T. Q., Borgne, H. L., Crucianu, M*. (2016). **Aggregating Image and Text Quantized Correlated Components**. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[4] *Hardoon, D. R., Szedmak, S., Shawe-Taylor, J.* (2004). **Canonical Correlation Analysis: An Overview with Application to Learning Methods**. Neural Computation, 16(12), 2639-2664. doi:10.1162/0899766042321814
[5] *Jin, C., Mao, W., Zhang, R., et al*. (2015). **Cross-modal image clustering via canonical correlation analysis**. Twenty-Ninth AAAI Conference on Artificial Intelligence .
[6] *Socher, R., Fei-Fei, L*. (2010). **Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora**. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2010.5540112
[7] *Hardoon, D. R., Saunders, C., Szedmak, S., Shawe-Taylor, J*. (2006). **A correlation approach for automatic image annotation**. Springer LNAI (Vol. 4093, pp. 681–692). Berlin: Springer.
[8] *Chrupała, G., Kádár, A., Alishahi, A.* (2015). **Learning language through pictures**. arXiv preprint arXiv:1506.03694.
[9] *Srivastava, N., Salakhutdinov, R.* (2012). **Multimodal learning with deep Boltzmann machines**. NIPS.
[10] *Collell, G., Zhang, T., Moens, M. F.* (2017). **Imagined Visual Representations as Multimodal Embeddings.** AAAI
[11] *Wang, L., Li, Y., Lazebnik, S.* (2016). **Learning deep structure-preserving image-text embeddings**. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5005-5013).
[12] *Lazaridou, A., Bruni, E., Baroni, M.* (2014). **Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world**. ACL (1) (pp. 1403-1414).