



Universiteit
Leiden
The Netherlands

Multi modal representation learning and cross-modal semantic matching

Wang, X.

Citation

Wang, X. (2022, June 24). *Multi modal representation learning and cross-modal semantic matching*. Retrieved from <https://hdl.handle.net/1887/3391031>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3391031>

Note: To cite this publication please use the final published version (if applicable).

BIBLIOGRAPHY

- [1] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: vol. 9. 8. MIT Press, 1997, pp. 1735–1780.
- [2] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.
- [3] Ross Girshick. “Fast R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1440–1448.
- [4] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *Proceedings of the International Conference on Learning Representations*. 2015.
- [5] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.
- [6] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1724–1734. DOI: 10.3115/v1/D14-1179.
- [7] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
- [8] Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. “Comparative analysis on cross-modal information retrieval: a review”. In: *Computer Science Review* 39 (2021), p. 100336.
- [9] Yunchao Gong et al. “A multi-view embedding space for modeling internet images, tags, and their semantics”. In: *International Journal of Computer Vision* 106.2 (2014), pp. 210–233.
- [10] Viresh Ranjan, Nikhil Rasiwasia, and CV Jawahar. “Multi-label Cross-modal Retrieval”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 4094–4102.
- [11] Peter Anderson et al. “Bottom-up and top-down attention for image captioning and visual question answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6077–6086.

- [12] Chenxi Liu et al. “Attention correctness in neural image captioning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.
- [13] Bo Dai and Dahua Lin. “Contrastive learning for image captioning”. In: *arXiv Preprint arXiv: 1710.02534* (2017).
- [14] Liwei Wang et al. “Improving weakly supervised visual grounding by contrastive knowledge Distillation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14090–14100.
- [15] Yongfei Liu et al. “Relation-aware Instance Refinement for Weakly Supervised Visual Grounding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5612–5621.
- [16] Damien Teney et al. “Tips and tricks for visual question answering: Learnings from the 2017 challenge”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4223–4232.
- [17] Tianyu Yu et al. “Cross-modal omni interaction modeling for phrase grounding”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 1725–1734.
- [18] Tanmay Gupta et al. “Contrastive Learning for Weakly Supervised Phrase Grounding”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.
- [19] Long Chen et al. “Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding”. In: *arXiv Preprint arXiv: 2009.01449* (2020).
- [20] Sibe Yang, Guanbin Li, and Yizhou Yu. “Relationship-embedded representation learning for grounding referring expressions”. In: *arXiv Preprint arXiv: 1906.04464* (2019).
- [21] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European Conference on Computer Vision*. Springer. 2014, pp. 740–755.
- [22] Ranjay Krishna et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations”. In: *International Journal of Computer Vision* 123.1 (2017), pp. 32–73.
- [23] Bryan A Plummer et al. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2641–2649.
- [24] Peter Young et al. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 67–78.
- [25] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *arXiv Preprint arXiv: 1310.4546* (2013).
- [26] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv Preprint arXiv: 1301.3781* (2013).

- [27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.
- [28] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://www.aclweb.org/anthology/N19-1423>.
- [29] Xuejing Liu et al. “Adaptive reconstruction network for weakly supervised referring expression grounding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2611–2620.
- [30] Shaoqing Ren et al. “Faster R-CNN: Towards real-time object detection with region proposal networks”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 91–99.
- [31] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2961–2969.
- [32] Ya Jing et al. “Learning Aligned Image-Text Representations Using Graph Attentive Relational Network”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 1840–1852.
- [33] Christiane Fellbaum. “WordNet”. In: *Theory and Applications of Ontology: Computer Applications*. Springer, 2010, pp. 231–243.
- [34] Robyn Speer, Joshua Chin, and Catherine Havasi. “Conceptnet 5.5: An open multilingual graph of general knowledge”. In: *Thirty-first AAAI Conference on Artificial Intelligence*. 2017.
- [35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding”. In: *arXiv Preprint arXiv: 1807.03748* (2018).
- [36] Hassan Akbari et al. “Multi-level multimodal common semantic space for image-phrase grounding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12476–12486.
- [37] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. “Canonical correlation analysis: An overview with application to learning methods”. In: *Neural Computation* 16.12 (2004), pp. 2639–2664.
- [38] David Grangier and Samy Bengio. “A Discriminative Kernel-Based Approach to Rank Images from Text Queries”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.8 (2008), pp. 1371–1384. DOI: 10.1109/TPAMI.2007.70791.

- [39] Raman Arora and Karen Livescu. “Kernel CCA for multi-view learning of acoustic features using articulatory measurements”. In: *Symposium on machine learning in speech and language processing*. 2012.
- [40] Douglas B Lenat. “CYC: A large-scale investment in knowledge infrastructure”. In: *Communications of the ACM* 38.11 (1995), pp. 33–38.
- [41] Sheng Guo et al. “CurriculumNet: Weakly supervised learning from large-scale web images”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 135–150.
- [42] Xin Huang and Yuxin Peng. “Deep cross-media knowledge transfer”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8837–8846.
- [43] Youtian Du and Kai Yang. “Learning semantic correlation of web images and text with mixture of local linear mappings”. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. 2015, pp. 1259–1262.
- [44] Xing Xu et al. “Cross-modal attention with semantic consistence for image-text matching”. In: *IEEE Transactions on Neural Networks and Learning Systems* 31.12 (2020), pp. 5412–5425.
- [45] Michael Gutmann and Aapo Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *Proceedings of the thirteenth international Conference on Artificial Intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 297–304.
- [46] John Lafferty and Chengxiang Zhai. “Document language models, query models, and risk minimization for information retrieval”. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001, pp. 111–119.
- [47] Angeliki Lazaridou et al. “Compositionally derived representations of morphologically complex words in distributional semantics”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013, pp. 1517–1526.
- [48] Kevin Lund and Curt Burgess. “Producing high-dimensional semantic spaces from lexical co-occurrence”. In: *Behavior Research Methods, Instruments, & Computers* 28.2 (1996), pp. 203–208.
- [49] Marco Baroni and Alessandro Lenci. “Distributional memory: A general framework for corpus-based semantics”. In: *Computational Linguistics* 36.4 (2010), pp. 673–721.
- [50] Ronan Collobert and Jason Weston. “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In: *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 160–167.

- [51] Piotr Bojanowski et al. “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.
- [52] Satwik Kottur et al. “Visual word2vec (Vis- W2V): Learning visually grounded word embeddings using abstract scenes”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4985–4994.
- [53] Cewu Lu et al. “Visual relationship detection with language priors”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 852–869.
- [54] Radityo Eko Prasajo, Mouna Kacimi, and Werner Nutt. “Modeling and summarizing news events using semantic triples”. In: *European Semantic Web Conference*. Springer. 2018, pp. 512–527.
- [55] Qiu hao Lu and Youtian Du. “Wikipedia-based Entity Semantifying in Open Information Extraction”. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 765–770.
- [56] Daniela Gerz et al. “Simverb-3500: A large-scale evaluation set of verb similarity”. In: *arXiv Preprint arXiv: 1608.00869* (2016).
- [57] Ali Diba et al. “Weakly supervised cascaded convolutional networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 914–922. ISBN: 9781538604571. DOI: 10 . 1109 / CVPR . 2017 . 545.
- [58] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 779–788.
- [59] Augustus Buonviri et al. “Survey of Challenges in Labeled Random Finite Set Distributed Multi-Sensor Multi-Object Tracking”. In: *2019 IEEE Aerospace Conference*. IEEE. 2019, pp. 1–12.
- [60] Wei Du, Ronald Phlypo, and Tülay Adalı. “Adaptive Feature Selection and Feature Fusion for Semi-supervised Classification”. In: *Journal of Signal Processing Systems* 91.5 (2019), pp. 521–537.
- [61] Hakan Bilen and Andrea Vedaldi. “Weakly supervised deep detection networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2846–2854.
- [62] Fang Wan et al. “Min-entropy latent model for weakly supervised object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1297–1306.
- [63] Keren Ye et al. “Cap2Det: Learning to amplify weak caption supervision for object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 9686–9695.

- [64] Christopher Thomas and Adriana Kovashka. “Predicting the politics of an image using webly supervised data”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 3630–3642.
- [65] Hao Fang et al. “From captions to visual concepts and back”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1473–1482.
- [66] Ishan Misra et al. “Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2930–2939.
- [67] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. “Equal but not the same: Understanding the implicit relationship between persuasive images and text”. In: *arXiv Preprint arXiv: 1807.08205* (2018).
- [68] Weifeng Ge, Sibe Yang, and Yizhou Yu. “Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1277–1286.
- [69] Peng Tang et al. “PCL: Proposal cluster learning for weakly supervised object detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.1 (2018), pp. 176–191.
- [70] Xiaolin Zhang et al. “Adversarial complementary learning for weakly supervised object localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1325–1334.
- [71] Yu Zhang et al. “Weakly supervised fine-grained categorization with part-based image representation”. In: *IEEE Transactions on Image Processing* 25.4 (2016), pp. 1713–1725.
- [72] Youtian Du et al. “Fundamental visual concept learning from correlated images and text”. In: *IEEE Transactions on Image Processing* 28.7 (2019), pp. 3598–3612.
- [73] Yale Song and Mohammad Soleymani. “Polysemous visual-semantic embedding for cross-modal retrieval”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1979–1988.
- [74] Jiu-le TIAN and Wei Zhao. “Words similarity algorithm based on Tongyici Cilin in semantic web adaptive learning system”. In: *Journal of Jilin University (Information Science Edition)* 6.010 (2010).
- [75] Zhendong Dong, Qiang Dong, and Changling Hao. “HowNet and the computation of meaning”. In: (2006).
- [76] Qun Liu. “Word similarity computing based on HowNet”. In: *Computational Linguistics and Chinese Language Processing* 7.2 (2002), pp. 59–76.

- [77] Jonathan Krause et al. “A hierarchical approach for generating descriptive image paragraphs”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 317–325.
- [78] Changliang Li et al. “Measuring word semantic similarity based on transferred vectors”. In: *International Conference on Neural Information Processing*. Springer. 2017, pp. 326–335.
- [79] Maxime Oquab et al. “Is object localization for free?-Weakly-supervised learning with convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 685–694.
- [80] Bolei Zhou et al. “Learning deep features for discriminative localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2921–2929.
- [81] Vadim Kantorov et al. “ContextLocNet: Context-aware deep network models for weakly supervised localization”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 350–365.
- [82] Peng Tang et al. “Multiple instance detection network with online instance classifier refinement”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2843–2851. ISBN: 9781538604571.
- [83] Yunchao Wei et al. “TS2C: Tight box mining with surrounding segmentation context for weakly supervised object detection”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 434–450.
- [84] Yoshua Bengio et al. “Curriculum learning”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 41–48.
- [85] Dingwen Zhang et al. “Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework”. In: *International Journal of Computer Vision* 127.4 (2019), pp. 363–380.
- [86] Miaojing Shi and Vittorio Ferrari. “Weakly supervised object localization using size estimates”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 105–121. ISBN: 9783319464534. DOI: 10 . 1007 / 978 - 3 - 319 - 46454 - 1 _ 7.
- [87] Jiasi Wang, Xinggang Wang, and Wenyu Liu. “Weakly- and semi-supervised Faster R-CNN with curriculum learning”. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE. 2018, pp. 2416–2421.
- [88] Guy Hacoheh and Daphna Weinshall. “On the power of curriculum learning in training deep networks”. In: *arXiv Preprint arXiv: 1904.03626* (2019).
- [89] Christopher D Manning et al. “The Stanford CoreNLP natural language processing toolkit”. In: *Proceedings of 52nd Annual Meeting of the Association for Computational linguistics: system demonstrations*. 2014, pp. 55–60.

- [90] Nikhil Rasiwasia et al. “A new approach to cross-modal multimedia retrieval”. In: *Proceedings of the 18th ACM International Conference on Multimedia*. 2010, pp. 251–260.
- [91] Xixuan Wu et al. “Cross matching of music and image”. In: *Proceedings of the 20th ACM International Conference on Multimedia*. 2012, pp. 837–840.
- [92] Yue-Ting Zhuang, Yi Yang, and Fei Wu. “Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval”. In: *IEEE Transactions on Multimedia* 10.2 (2008), pp. 221–229.
- [93] Yansong Feng and Mirella Lapata. “Automatic caption generation for news images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.4 (2012), pp. 797–812.
- [94] Lei Wu, Rong Jin, and Anil K Jain. “Tag completion for image retrieval”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.3 (2012), pp. 716–727.
- [95] Meng Wang et al. “Assistive tagging: A survey of multimedia tagging with human-computer joint exploration”. In: *ACM Computing Surveys (CSUR)* 44.4 (2012), pp. 1–24.
- [96] Hai-Feng Guo et al. “Deep multi-instance multi-label learning for image annotation”. In: *International Journal of Pattern Recognition and Artificial Intelligence* 32.03 (2018), p. 1859005.
- [97] Jinhui Tang et al. “Cross-space affinity learning with its application to movie recommendation”. In: *IEEE Transactions on Knowledge and Data Engineering* 25.7 (2012), pp. 1510–1519.
- [98] Jose Costa Pereira et al. “On the role of correlation and abstraction in cross-modal multimedia retrieval”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.3 (2013), pp. 521–535.
- [99] Tao Jiang and Ah-Hwee Tan. “Learning image-text associations”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.2 (2008), pp. 161–177.
- [100] Florent Monay and Daniel Gatica-Perez. “Modeling semantic aspects for cross-media image indexing”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.10 (2007), pp. 1802–1817.
- [101] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. “Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval”. In: *International Conference on Multimedia Modeling*. Springer. 2012, pp. 312–322.
- [102] Aviv Eisenschtat and Lior Wolf. “Linking image and text with 2-way nets”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4601–4611.
- [103] Liwei Wang et al. “Learning two-branch neural networks for image-text matching tasks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2018), pp. 394–407.

- [104] David Grangier and Samy Bengio. “A discriminative kernel-based model to rank images from text queries”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.8 (2008), pp. 1371–1384.
- [105] Tao Jiang and Ah-Hwee Tan. “Learning image-text associations”. In: *IEEE Transactions on Knowledge and Data Engineering* 21.2 (2009), pp. 161–177.
- [106] Pereira J Costa et al. “On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.3 (2014), pp. 521–535.
- [107] Hong Zhang, Yueting Zhuang, and Fei Wu. “Cross-modal correlation learning for clustering on image-audio dataset”. In: *Proceedings of the 15th ACM International Conference on Multimedia*. 2007, pp. 273–276.
- [108] Hong Liu et al. “Supervised matrix factorization for cross-modality hashing”. In: *arXiv Preprint arXiv: 1603.05572* (2016).
- [109] Ting Zhang and Jingdong Wang. “Collaborative quantization for cross-modal similarity search”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2036–2045.
- [110] Ran He et al. “Cross-modal subspace learning via pairwise constraints”. In: *IEEE Transactions on Image Processing* 24.12 (2015), pp. 5543–5556.
- [111] Jinhui Tang et al. “Cross-space affinity learning with its application to movie recommendation”. In: *IEEE Transactions on Knowledge and Data Engineering* 25.7 (2013), pp. 1510–1519.
- [112] Antoine Deleforge, Florence Forbes, and Radu Horaud. “High-dimensional regression with gaussian mixtures and partially-latent response variables”. In: *Statistics and Computing* 25.5 (2015), pp. 893–911.
- [113] Lauren A Hannah, David M Blei, and Warren B Powell. “Dirichlet process mixtures of generalized linear models.” In: *Journal of Machine Learning Research* 12.6 (2011).
- [114] Yan Hua et al. “Cross-modal correlation learning by adaptive hierarchical semantic aggregation”. In: *IEEE Transactions on Multimedia* 18.6 (2016), pp. 1201–1216.
- [115] Liang Zhang et al. “Cross-modal retrieval using multiorordered discriminative structured subspace learning”. In: *IEEE Transactions on Multimedia* 19.6 (2017), pp. 1220–1233.
- [116] Xing Xu et al. “Learning discriminative binary codes for large-scale cross-modal retrieval”. In: *IEEE Transactions on Image Processing* 26.5 (2017), pp. 2494–2507.
- [117] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. “Automatic image annotation and retrieval using cross-media relevance models”. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. 2003, pp. 119–126.

- [118] Yansong Feng and Mirella Lapata. “Automatic caption generation for news images”. In: *IEEE transactions on Pattern Analysis and Machine Intelligence* 35.4 (2013), pp. 797–812.
- [119] Ruofei Zhang et al. “A probabilistic semantic model for image annotation and multimodal image retrieval”. In: *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*. Vol. 1. IEEE. 2005, pp. 846–851.
- [120] Ying Wu, Qi Tian, and Thomas S Huang. “Discriminant-EM algorithm with application to image retrieval”. In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*. Vol. 1. IEEE. 2000, pp. 222–227.
- [121] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. “Learning cross-modality similarity for multinomial data”. In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 2407–2414.
- [122] Anh Pham et al. “Multi-instance multi-label learning in the presence of novel class instances”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2427–2435.
- [123] Wanxia Lin, Tong Lu, and Feng Su. “A novel multi-modal integration and propagation model for cross-media information retrieval”. In: *International Conference on Multimedia Modeling*. Springer. 2012, pp. 740–749.
- [124] Michalis Lazaridis et al. “Multimedia search and retrieval using multimodal annotation propagation and indexing techniques”. In: *Signal Processing: Image Communication* 28.4 (2013), pp. 351–367.
- [125] Jiao Xue, Youtian Du, and Hanbing Shui. “Semantic correlation mining between images and texts with global semantics and local mapping”. In: *International Conference on Multimedia Modeling*. Springer. 2015, pp. 427–435.
- [126] Dong Liu et al. “Image retagging using collaborative tag propagation”. In: *IEEE Transactions on Multimedia* 13.4 (2011), pp. 702–712.
- [127] Lei Zhang et al. “Full-space local topology extraction for cross-modal retrieval”. In: *IEEE Transactions on Image Processing* 24.7 (2015), pp. 2212–2224.
- [128] Fei Yan and Krystian Mikolajczyk. “Deep correlation for matching images and text”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3441–3450.
- [129] Liwei Wang, Yin Li, and Svetlana Lazebnik. “Learning deep structure-preserving image-text embeddings”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5005–5013.
- [130] Yuxin Peng et al. “CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network”. In: *IEEE Transactions on Multimedia* 20.2 (2018), pp. 405–420.

- [131] Richang Hong et al. “Coherent semantic-visual indexing for large-scale image retrieval in the cloud”. In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4128–4138.
- [132] Bo Wang et al. “Movie question answering: Remembering the textual cues for layered visual contents”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
- [133] Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC Press, 1994.
- [134] Christopher M. Bishop. “Pattern recognition and machine learning”. In: Springer, 2006.
- [135] Jingdong Wang, Jianguo Lee, and Changshui Zhang. “Kernel trick embedded Gaussian mixture model”. In: *International Conference on Algorithmic Learning Theory*. Springer. 2003, pp. 159–174.
- [136] Jeff A Bilmes et al. “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models”. In: *International Computer Science Institute* 4.510 (1998), p. 126.
- [137] Baback Moghaddam and Alex Pentland. “Probabilistic visual learning for object representation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.7 (1997), pp. 696–710.
- [138] Micah Hodosh, Peter Young, and Julia Hockenmaier. “Framing image description as a ranking task: Data, models and evaluation metrics”. In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 853–899.
- [139] Benjamin Klein et al. “Associating neural word embeddings with deep image representations using fisher vectors”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4437–4446.
- [140] Tat-Seng Chua et al. “Nus-wide: A real-world web image database from national university of singapore”. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*. 2009, pp. 1–9.
- [141] Quoc Le and Tomas Mikolov. “Distributed representations of sentences and documents”. In: *International Conference on Machine Learning*. PMLR. 2014, pp. 1188–1196.
- [142] Martin Engilberge et al. “Finding beans in burgers: Deep semantic-visual embedding with localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3984–3993.
- [143] A. Karpathy and L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (2017), pp. 664–676.
- [144] Ivan Vendrov et al. “Order-embeddings of images and language”. In: *arXiv Preprint arXiv: 1511.06361* (2015).
- [145] Quanzeng You, Zhengyou Zhang, and Jiebo Luo. “End-to-end convolutional semantic embeddings”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

- [146] Yu Liu et al. “Learning a recurrent residual fusion network for multimodal matching”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 4107–4116.
- [147] Xinlei Chen et al. “Microsoft COCO captions: Data collection and evaluation server”. In: *arXiv Preprint arXiv: 1504.00325* (2015).
- [148] Stanislaw Antol et al. “VQA: Visual question answering”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2425–2433.
- [149] Alane Suhr et al. “A corpus for reasoning about natural language grounded in photographs”. In: *arXiv Preprint arXiv: 1811.00491* (2018).
- [150] Rowan Zellers et al. “From recognition to cognition: Visual commonsense reasoning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6720–6731.
- [151] Samyak Datta et al. “Align2ground: Weakly supervised phrase grounding guided by image-caption alignment”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2601–2610.
- [152] Farley Lai et al. “Contextual Grounding of Natural Language Entities in Images”. In: *arXiv Preprint arXiv: 1911.02133* (2019).
- [153] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. “G3graphground: Graph-based language grounding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4281–4290.
- [154] Bryan A Plummer et al. “Phrase localization and visual relationship detection with comprehensive image-language cues”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1928–1937.
- [155] Kan Chen, Jiyang Gao, and Ram Nevatia. “Knowledge aided consistency for weakly supervised phrase grounding”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4042–4050.
- [156] A. Neubeck and L. Van Gool. “Efficient Non-Maximum Suppression”. In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 3. 2006, pp. 850–855. DOI: 10.1109/ICPR.2006.479.
- [157] C Lawrence Zitnick and Piotr Dollár. “Edge boxes: Locating object proposals from edges”. In: *European Conference on Computer Vision*. Springer. 2014, pp. 391–405.
- [158] Navaneeth Bodla et al. “Soft-NMS – Improving Object Detection With One Line of Code”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 5561–5569.
- [159] Yihui He et al. “Softer-NMS: Rethinking bounding box regression for accurate object detection”. In: *arXiv Preprint arXiv: 1809.08545* 2.3 (2018).
- [160] Long Chen et al. “Ref-NMS: Breaking proposal bottlenecks in two-stage referring expression grounding”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2. 2021, pp. 1036–1044.

- [161] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9729–9738.
- [162] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1597–1607.
- [163] Zhirong Wu et al. “Unsupervised feature learning via non-parametric instance discrimination”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3733–3742.
- [164] Han Zhang et al. “Cross-modal contrastive learning for text-to-image generation”. In: *arXiv Preprint arXiv: 2101.04702* (2021).
- [165] Zhuowan Li et al. “Context-aware group captioning via self-attention and contrastive features”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3440–3450.
- [166] Xin Huang and Yuxin Peng. “Cross-modal deep metric learning with multi-task regularization”. In: *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2017, pp. 943–948.
- [167] Ashish Vaswani et al. “Attention is all you need”. In: *arXiv Preprint arXiv: 1706.03762* (2017).
- [168] Anna Rohrbach et al. “Grounding of textual phrases in images by reconstruction”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 817–834. ISBN: 9783319464473. DOI: 10 . 1007 / 978 - 3 - 319 - 46448 - 0_49.
- [169] Licheng Yu et al. “Modeling context in referring expressions”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 69–85.
- [170] Junhua Mao et al. “Generation and comprehension of unambiguous object descriptions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 11–20.

