Exploring deep learning for multimodal understanding

Lao, M.

Bibliography

[1] Davenport, T.H., Ronanki, R., et al.: Artificial intelligence for the real world. Harvard business review **96** (2018) 108–116

[2] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521** (2015) 436–444

[3] Kamilaris, A., Prenafeta-Boldú, F.X.: Deep learning in agriculture: A survey. Computers and electronics in agriculture **147** (2018) 70–90

[4] Lu, D., Weng, Q.: A survey of image classification methods and techniques for improving classification performance. International journal of Remote sensing **28** (2007) 823–870

[5] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 558–567

[6] Hafiz, A.M., Bhat, G.M.: A survey on instance segmentation: state of the art. International journal of multimedia information retrieval **9** (2020) 171–189

[7] Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 9157–9166

[8] Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE (2023)

[9] Wu, X., Sahoo, D., Hoi, S.C.: Recent advances in deep learning for object detection. Neurocomputing **396** (2020) 39–64

[10] Almeida, F., Xexéo, G.: Word embeddings: A survey. arXiv preprint arXiv:1901.09069 (2019)

[11] Wang, B., Wang, A., Chen, F., Wang, Y., Kuo, C.C.J.: Evaluating word embedding models: Methods and experimental results. APSIPA transactions on signal and information processing **8** (2019)

[12] Kale, M., Rastogi, A.: Text-to-text pre-training for data-to-text tasks. arXiv preprint arXiv:2005.10433 (2020)

[13] Yu, P., Fei, H., Li, P.: Cross-lingual language model pretraining for retrieval. In: Proceedings of the Web Conference 2021. (2021) 1029–1039

[14] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. (2020) 38–45

[15] Ruder, S., Peters, M.E., Swayamdipta, S., Wolf, T.: Transfer learning in natural language processing. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials. (2019) 15–18

[16] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: A survey. Information **10** (2019) 150

[17] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning–based text classification: a comprehensive review. ACM computing surveys (CSUR)

54 (2021) 1–40

[18] Zeng, C., Li, S., Li, Q., Hu, J., Hu, J.: A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. Applied Sciences **10** (2020) 7640

[19] Zhang, Z., Yang, J., Zhao, H.: Retrospective reader for machine reading comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 35. (2021) 14506–14514

[20] Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine **34** (2017) 96–108

[21] Blikstein, P.: Multimodal learning analytics. In: Proceedings of the third international conference on learning analytics and knowledge. (2013) 102–106

[22] Peng, Y., Huang, X., Zhao, Y.: An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. IEEE Transactions on circuits and systems for video technology **28** (2017) 2372–2385

[23] Huang, C., Luo, X., Zhang, J., Liao, Q., Wang, X., Jiang, Z.L., Qi, S.: Explore instance similarity: An instance correlation based hashing method for multi-label cross-model retrieval. Information Processing & Management **57** (2020) 102165

[24] Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) **51** (2019) 1–36

[25] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence **39** (2016) 652–663

[26] Deng, C., Wu, Q., Wu, Q., Hu, F., Lyu, F., Tan, M.: Visual grounding via accumulated attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7746–7755

[27] Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 4683–4693

[28] Kocasari, U., Dirik, A., Tiftikci, M., Yanardag, P.: Stylemc: multi-channel based fast text-guided image generation and manipulation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2022) 895–904

[29] Kress, G., Selander, S.: Multimodal design, learning and cultures of recognition. The internet and higher education **15** (2012) 265–268

[30] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. (2015) 2425–2433

[31] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding **163** (2017) 21–40

[32] Kafle, K., Kanan, C.: Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding **163** (2017) 3–20

[33] Geman, D., Geman, S., Hallonquist, N., Younes, L.: Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences **112** (2015) 3618–3623

[34] Suk, H.I., Lee, S.W., Shen, D., Initiative, A.D.N., et al.: Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. NeuroImage **101** (2014) 569–582

[35] Xu, J., Yao, T., Zhang, Y., Mei, T.: Learning multimodal attention lstm networks for video captioning. In: Proceedings of the 25th ACM international conference on Multimedia. (2017) 537–545

[36] Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems **16** (2010) 345–379

[37] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130** (2022) 2337–2348

[38] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 16816–16825

[39] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI conference on artificial intelligence. Volume 34. (2020) 13041–13049

[40] Chen, F.L., Zhang, D.Z., Han, M.L., Chen, X.Y., Shi, J., Xu, S., Xu, B.: Vlp: A survey on vision-language pre-training. Machine Intelligence Research **20** (2023) 38–56

[41] Chen, C., Anjum, S., Gurari, D.: Grounding answers for visual questions asked by visually impaired people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 19098–19107

[42] Baker, K., Parekh, A., Fabre, A., Addlesee, A., Kruiper, R., Lemon, O.: The spoon is in the sink: Assisting visually impaired people in the kitchen. In: Proceedings of the Reasoning and Interaction Conference (ReInAct 2021). (2021) 32–39

[43] Di Nuovo, A., Conti, D., Trubia, G., Buono, S., Di Nuovo, S.: Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability. Robotics **7** (2018) 25

[44] Han, D.M., Lim, J.H.: Design and implementation of smart home energy management systems based on zigbee. IEEE Transactions on Consumer Electronics **56** (2010) 1417–1425

[45] de Gelder, E., Paardekooper, J.P.: Assessment of automated driving systems using real-life scenarios. In: 2017 IEEE Intelligent Vehicles Symposium (IV), IEEE (2017) 589–594

[46] Bakator, M., Radosav, D.: Deep learning and medical diagnosis: A review of literature. Multimodal Technologies and Interaction **2** (2018) 47

[47] Lund, B.D., Wang, T.: Chatting about chatgpt: how may ai and gpt impact academia and libraries? Library Hi Tech News (2023)

[48] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 6904–6913

[49] Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 6700–6709

[50] Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV). (2018) 670–685

[51] Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., et al.: Vizwiz: nearly real-time answers to visual questions. In: Proceedings of the 23nd annual ACM symposium on User interface software and technology. (2010) 333–342

[52] Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. Advances in neural information processing systems **28** (2015)

[53] Malinowski, M., Fritz, M.: Towards a visual turing challenge. arXiv preprint arXiv:1410.8027 (2014)

[54] Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 4995–5004

[55] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123** (2017) 32–73

[56] Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D.: Yin and yang: Balancing

and answering binary visual questions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 5014–5022

[57] Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2901–2910

[58] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778

[59] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. (2015) 1440–1448

[60] Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)

[61] Dey, R., Salem, F.M.: Gate-variants of gated recurrent unit (gru) neural networks. In: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS), IEEE (2017) 1597–1600

[62] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[63] Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. Neurocomputing **452** (2021) 48–62

[64] Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Attention branch network: Learning of attention mechanism for visual explanation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 10705–10714

[65] Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M.: Attention mechanisms in computer vision: A survey. Computational Visual Media **8** (2022) 331–368

[66] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)

[67] Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems **29** (2016)

[68] Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. Advances in neural information processing systems **31** (2018)

[69] Li, L., Gan, Z., Cheng, Y., Liu, J.: Relation-aware graph attention network for visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 10313–10322

[70] Zhu, C., Zhao, Y., Huang, S., Tu, K., Ma, Y.: Structured attentions for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1291–1300

[71] Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding **163** (2017) 90–100

[72] Peng, L., Yang, Y., Wang, Z., Huang, Z., Shen, H.T.: Mra-net: Improving vqa via multi-modal relation attention network. IEEE Transactions on Pattern Analysis and Machine Intelligence **44** (2020) 318–329

[73] Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)

[74] Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 6281–6290

[75] Zhou, Y., Ren, T., Zhu, C., Sun, X., Liu, J., Ding, X., Xu, M., Ji, R.: Trar: Routing the attention spans in transformer for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 2074–2084

[76] Xin, B., Huang, J., Zhou, Y., Lu, J., Wang, X.: Interpretation on deep multimodal fusion for diagnostic classification. In: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE (2021) 1–8

[77] Sui, J., Adali, T., Yu, Q., Chen, J., Calhoun, V.D.: A review of multivariate methods for multimodal fusion of brain imaging data. Journal of neuroscience methods **204** (2012) 68–81

[78] Kong, S., Fowlkes, C.: Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 365–374

[79] Wei, X., Zhang, Y., Gong, Y., Zhang, J., Zheng, N.: Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 355–370

[80] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2016) 457–468

[81] Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)

[82] Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision. (2017) 1821–1830

[83] Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision. (2017) 2612–2620

[84] Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems **29** (2018) 5947–5959

[85] Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 4971–4980

[86] Kervadec, C., Antipov, G., Baccouche, M., Wolf, C.: Roses are red, violets are blue... but should vqa expect them to? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 2776–2785

[87] Dancette, C., Cadene, R., Teney, D., Cord, M.: Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 1574–1583

[88] Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12700–12710

[89] Ramakrishnan, S., Agrawal, A., Lee, S.: Overcoming language priors in visual question answering with adversarial regularization. Advances in Neural Information Processing Systems **31** (2018)

[90] Guo, Y., Nie, L., Cheng, Z., Tian, Q., Zhang, M.: Loss re-scaling vqa: revisiting the language prior problem from a class-imbalance view. IEEE Transactions on Image Processing **31** (2021) 227–238

[91] Guo, Y., Cheng, Z., Nie, L., Liu, Y., Wang, Y., Kankanhalli, M.: Quantifying and alleviating the language prior problem in visual question answering. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. (2019) 75–84

[92] Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In:

Proceedings of the IEEE/CVF international conference on computer vision. (2019) 2591–2600

[93] Han, X., Wang, S., Su, C., Huang, Q., Tian, Q.: Greedy gradient ensemble for robust visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 1584–1593

[94] Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 10800–10809

[95] Abbasnejad, E., Teney, D., Parvaneh, A., Shi, J., Hengel, A.v.d.: Counterfactual vision and language learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 10044–10054

[96] Guo, Y., Nie, L., Cheng, Z., Ji, F., Zhang, J., Del Bimbo, A.: Adavqa: Overcoming language priors with adapted margin cosine loss. arXiv preprint arXiv:2105.01993 (2021)

[97] Chen, L., Zheng, Y., Xiao, J.: Rethinking data augmentation for robust visual question answering. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI, Springer (2022) 95–112

[98] Si, Q., Liu, Y., Meng, F., Lin, Z., Fu, P., Cao, Y., Wang, W., Zhou, J.: Towards robust visual question answering: Making the most of biased samples via contrastive learning. arXiv preprint arXiv:2210.04563 (2022)

[99] Gat, I., Schwartz, I., Schwing, A., Hazan, T.: Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. Advances in Neural Information Processing Systems **33** (2020) 3197–3208

[100] Jing, C., Wu, Y., Zhang, X., Jia, Y., Wu, Q.: Overcoming language priors in vqa via decomposed linguistic representations. In: Proceedings of the AAAI conference on artificial intelligence. Volume 34. (2020) 11181–11188

[101] Kv, G., Mittal, A.: Reducing language biases in visual question answering with visually-grounded question encoder. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, Springer (2020) 18–34

[102] Qiao, Y., Yu, Z., Liu, J.: Rankvqa: Answer re-ranking for visual question answering. In: 2020 IEEE international conference on multimedia and expo (ICME), IEEE (2020) 1–6

[103] Si, Q., Lin, Z., Zheng, M., Fu, P., Wang, W.: Check it again: Progressive visual question answering via visual entailment. arXiv preprint arXiv:2106.04605 (2021)

[104] Qiao, T., Dong, J., Xu, D.: Exploring human-like attention supervision in visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 32. (2018)

[105] Wu, J., Mooney, R.: Self-critical reasoning for robust visual question answering. Advances in Neural Information Processing Systems **32** (2019)

[106] Cadene, R., Dancette, C., Cord, M., Parikh, D., et al.: Rubi: Reducing unimodal biases for visual question answering. Advances in neural information processing systems **32** (2019)

[107] Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. arXiv preprint arXiv:1909.03683 (2019)

[108] Cho, J.W., Kim, D.j., Ryu, H., Kweon, I.S.: Generative bias for visual question answering. arXiv preprint arXiv:2208.00690 (2022)

[109] Liang, Z., Hu, H., Zhu, J.: Lpf: a language-prior feedback objective function for de-biased visual question answering. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. (2021) 1955–1959

[110] Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Mutant: A training paradigm for out-of-distribution generalization in visual question answering. arXiv preprint arXiv:2009.08566 (2020)

[111] Wen, Z., Xu, G., Tan, M., Wu, Q., Wu, Q.: Debiased visual question answering from feature and sample perspectives. Advances in Neural Information Processing Systems **34** (2021)

3784–3796

[112] Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: Avqa: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia. (2022) 3480–3491

[113] Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-avqa: Grounded audio-visual question answering on 360deg videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 2031–2041

[114] Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 19108–19118

[115] Zhuang, Y., Xu, D., Yan, X., Cheng, W., Zhao, Z., Pu, S., Xiao, J.: Multichannel attention refinement for video question answering. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **16** (2020) 1–23

[116] Miyanishi, T., Kawanabe, M.: Watch, listen, and answer: Open-ended videoqa with modulated multi-stream 3d convnets. In: 2021 29th European Signal Processing Conference (EUSIPCO), IEEE (2021) 706–710

[117] Biesialska, M., Biesialska, K., Costa-Jussa, M.R.: Continual lifelong learning in natural language processing: A survey. arXiv preprint arXiv:2012.09823 (2020)

[118] Poquet, O., De Laat, M.: Developing capabilities: Lifelong learning in the age of ai. British Journal of Educational Technology **52** (2021) 1695–1708

[119] Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., Díaz-Rodríguez, N.: Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. Information fusion **58** (2020) 52–68

[120] Belouadah, E., Popescu, A., Kanellos, I.: A comprehensive study of class incremental learning algorithms for visual tasks. Neural Networks **135** (2021) 38–54

[121] Yan, S., Xie, J., He, X.: Der: Dynamically expandable representation for class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 3014–3023

[122] Belouadah, E., Popescu, A.: Il2m: Class incremental learning with dual memory. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 583–592

[123] Churamani, N., Kara, O., Gunes, H.: Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition. IEEE Transactions on Affective Computing (2022)

[124] Gunasekara, N., Gomes, H., Bifet, A., Pfahringer, B.: Adaptive online domain incremental continual learning. In: Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part I, Springer (2022) 491–502

[125] Kundu, J.N., Venkatesh, R.M., Venkat, N., Revanur, A., Babu, R.V.: Class-incremental domain adaptation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, Springer (2020) 53–69

[126] Verwimp, E., De Lange, M., Tuytelaars, T.: Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 9385–9394

[127] Yoon, J., Madaan, D., Yang, E., Hwang, S.J.: Online coreset selection for rehearsal-based continual learning. arXiv preprint arXiv:2106.01085 (2021)

[128] Buzzega, P., Boschini, M., Porrello, A., Calderara, S.: Rethinking experience replay: a bag of tricks for continual learning. In: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE (2021) 2180–2187

[129] Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. Advances in neural information processing systems **32** (2019)

[130] Michieli, U., Zanuttigh, P.: Continual semantic segmentation via repulsion-attraction of

sparse and disentangled latent representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2021) 1114–1124

[131] Boschini, M., Bonicelli, L., Buzzega, P., Porrello, A., Calderara, S.: Class-incremental continual learning into the extended der-verse. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[132] Monaikul, N., Castellucci, G., Filice, S., Rokhlenko, O.: Continual learning for named entity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 35. (2021) 13570–13577

[133] Singh, P., Mazumder, P., Rai, P., Namboodiri, V.P.: Rectification-based knowledge retention for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 15282–15291

[134] Wan, T.S., Chen, J.C., Wu, T.Y., Chen, C.S.: Continual learning for visual search with backward consistent feature embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 16702–16711

[135] Del Chiaro, R., Twardowski, B., Bagdanov, A., van de Weijer, J.: Ratt: Recurrent attention to transient tasks for continual image captioning. Proceedings of the International Conference on Neural Information Processing Systems **33** (2020)

[136] Nguyen, G., Jun, T.J., Tran, T., Yalew, T., Kim, D.: Contcap: A scalable framework for continual image captioning. arXiv preprint arXiv:1909.08745 (2019)

[137] Greco, C., Plank, B., Fernández, R., Bernardi, R.: Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. arXiv preprint arXiv:1906.04229 (2019)

[138] Lei, S.W., Gao, D., Wu, J.Z., Wang, Y., Liu, W., Zhang, M., Shou, M.Z.: Symbolic replay: Scene graph as prompt for continual learning on vqa task. arXiv preprint arXiv:2208.12037 (2022)

[139] Srinivasan, T., Chang, T.Y., Pinto Alva, L., Chochlakis, G., Rostami, M., Thomason, J.: Climb: A continual learning benchmark for vision-and-language tasks. Advances in Neural Information Processing Systems **35** (2022) 29440–29453

[140] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., He, B.: A survey on federated learning systems: vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering (2021)

[141] Yang, Q.: Advances and open problems in federated learning. Foundations and Trends in Machine Learning (2021)

[142] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics, PMLR (2017) 1273–1282

[143] Haddadpour, F., Mahdavi, M.: On the convergence of local descent methods in federated learning. arXiv preprint arXiv:1910.14425 (2019)

[144] Khaled, A., Mishchenko, K., Richtárik, P.: Tighter theory for local sgd on identical and heterogeneous data. In: International Conference on Artificial Intelligence and Statistics, PMLR (2020) 4519–4529

[145] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, B., et al.: Towards federated learning at scale: System design. Proceedings of machine learning and systems **1** (2019) 374–388

[146] Konečnỳ, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 (2016)

[147] Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of fedavg on non-iid data. arXiv preprint arXiv:1907.02189 (2019)

[148] Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623 (2021)

[149] Kulkarni, V., Kulkarni, M., Pant, A.: Survey of personalization techniques for federated learning. In: 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), IEEE (2020) 794–797

[150] Liang, P.P., Liu, T., Ziyin, L., Allen, N.B., Auerbach, R.P., Brent, D., Salakhutdinov, R., Morency, L.P.: Think locally, act globally: Federated learning with local and global representations. arXiv preprint arXiv:2001.01523 (2020)

[151] Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning: A meta-learning approach. arXiv preprint arXiv:2002.07948 (2020)

[152] Yu, T., Bagdasaryan, E., Shmatikov, V.: Salvaging federated learning by local adaptation. arXiv preprint arXiv:2002.04758 (2020)

[153] Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Federated learning for vision-and-language grounding problems. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 11572–11579

[154] Smola, A.J., Gretton, A., Borgwardt, K.: Maximum mean discrepancy. In: 13th international conference, ICONIP. (2006) 3–6

[155] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40** (2017) 1452–1464

[156] Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 6541–6549

[157] Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., et al.: Deep learning for computer vision: A brief review. Computational intelligence and neuroscience **2018** (2018)

[158] Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. Journal of the American Medical Informatics Association **18** (2011) 544–551

[159] Yampolskiy, R.V.: Turing test as a defining feature of ai-completeness. Artificial Intelligence, Evolutionary Computing and Metaheuristics: In the Footsteps of Alan Turing (2013) 3–17

[160] Manjunatha, V., Saini, N., Davis, L.S.: Explicit bias discovery in visual question answering models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 9562–9571

[161] Jolly, S., Kapoor, S.: Can pre-training help vqa with lexical variations? In: Findings of the Association for Computational Linguistics: EMNLP 2020. (2020) 2863–2868

[162] Selvaraju, R.R., Tendulkar, P., Parikh, D., Horvitz, E., Ribeiro, M.T., Nushi, B., Kamar, E.: Squinting at vqa models: Introspecting vqa models with sub-questions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10003–10011

[163] Sharma, V., Kalra, A., Vaibhav, S.C., Patel, L., Morency, L.P.: Attend and attack: Attention guided adversarial attacks on visual question answering models. In: Proc. Conf. Neural Inf. Process. Syst. Workshop Secur. Mach. Learn. Volume 2. (2018)

[164] Zhang, M., Maidment, T., Diab, A., Kovashka, A., Hwa, R.: Domain-robust vqa with diverse datasets and methods but no target labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 7046–7056

[165] Li, R., Xu, C., Guo, Z., Fan, B., Zhang, R., Liu, W., Zhao, Y., Gong, W., Wang, E.: Ai-vqa: Visual question answering based on agent interaction with interpretability. In: Proceedings of the 30th ACM International Conference on Multimedia. (2022) 5274–5282

[166] Whitehead, S., Petryk, S., Shakib, V., Gonzalez, J., Darrell, T., Rohrbach, A., Rohrbach, M.: Reliable visual question answering: Abstain rather than answer incorrectly. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI, Springer (2022) 148–166

[167] Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning,

PMLR (2019) 7472–7482

[168] Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: Robust statistics: theory and methods (with R). John Wiley & Sons (2019)

[169] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)

[170] Srivastava, Y., Murali, V., Dubey, S.R., Mukherjee, S.: Visual question answering using deep learning: A survey and performance analysis. In: Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5, Springer (2021) 75–86

[171] Zou, Y., Xie, Q.: A survey on vqa: Datasets and approaches. In: 2020 2nd International Conference on Information Technology and Computer Application (ITCA), IEEE (2020) 289–297

[172] Gupta, A.K.: Survey of visual question answering: Datasets and techniques. arXiv preprint arXiv:1705.03865 (2017)

[173] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2** (2020) 665–673

[174] Dagaev, N., Roads, B.D., Luo, X., Barry, D.N., Patil, K.R., Love, B.C.: A too-good-to-be-true prior to reduce shortcut reliance. Pattern Recognition Letters **166** (2023) 164–171

[175] Jing, C., Jia, Y., Wu, Y., Liu, X., Wu, Q.: Maintaining reasoning consistency in compositional visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 5099–5108

[176] Tascon-Morales, S., Márquez-Neila, P., Sznitman, R.: Logical implications for visual question answering consistency. arXiv preprint arXiv:2303.09427 (2023)

[177] Heinze-Deml, C., Meinshausen, N.: Conditional variance penalties and domain shift robustness. arXiv preprint arXiv:1710.11469 (2017)

[178] Glenski, M., Ayton, E., Cosbey, R., Arendt, D., Volkova, S.: Towards trustworthy deception detection: Benchmarking model robustness across domains, modalities, and languages. arXiv preprint arXiv:2104.11761 (2021)

[179] Toyama, K.: There are no technology shortcuts to good education. Educational Technology Debate **8** (2011)

[180] Shrestha, R., Kafle, K., Kanan, C.: An investigation of critical issues in bias mitigation techniques. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2022) 1943–1954

[181] Shrestha, R., Kafle, K., Kanan, C.: A negative case analysis of visual grounding methods for vqa. arXiv preprint arXiv:2004.05704 (2020)

[182] Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 6720–6731

[183] Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. Advances in neural information processing systems **33** (2020) 19290–19301

[184] Gupta, V., Li, Z., Kortylewski, A., Zhang, C., Li, Y., Yuille, A.: Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 5078–5088

[185] Si, Q., Meng, F., Zheng, M., Lin, Z., Liu, Y., Fu, P., Cao, Y., Wang, W., Zhou, J.: Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. arXiv preprint arXiv:2210.04692 (2022)

[186] Agarwal, V., Shetty, R., Fritz, M.: Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 9690–9698

[187] Goel, V., Chandak, M., Anand, A., Guha, P.: Iq-vqa: intelligent visual question answering. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II, Springer (2021) 357–370

[188] Akula, A., Jampani, V., Changpinyo, S., Zhu, S.C.: Robust visual reasoning via language guided neural module networks. Advances in Neural Information Processing Systems **34** (2021) 11041–11053

[189] Whitehead, S., Wu, H., Fung, Y.R., Ji, H., Feris, R., Saenko, K.: Learning from lexical perturbations for consistent visual question answering. arXiv preprint arXiv:2011.13406 (2020)

[190] Han, Y., Nie, L., Yin, J., Wu, J., Yan, Y.: Visual perturbation-aware collaborative learning for overcoming the language prior problem. arXiv preprint arXiv:2207.11850 (2022)

[191] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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) 6077–6086

[192] Ray, A., Sikka, K., Divakaran, A., Lee, S., Burachas, G.: Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. arXiv preprint arXiv:1909.04696 (2019)

[193] Shah, M., Chen, X., Rohrbach, M., Parikh, D.: Cycle-consistency for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6649–6658

[194] Yuan, Y., Wang, S., Jiang, M., Chen, T.Y.: Perception matters: Detecting perception failures of vqa models using metamorphic testing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 16908–16917

[195] Wang, R., Qian, Y., Feng, F., Wang, X., Jiang, H.: Co-vqa: Answering by interactive sub question sequence. arXiv preprint arXiv:2204.00879 (2022)

[196] Dua, R., Kancheti, S.S., Balasubramanian, V.N.: Beyond vqa: Generating multi-word answers and rationales to visual questions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 1623–1632

[197] Fedorenko, E., Scott, T.L., Brunner, P., Coon, W.G., Pritchett, B., Schalk, G., Kanwisher, N.: Neural correlate of the construction of sentence meaning. Proceedings of the National Academy of Sciences **113** (2016) E6256–E6262

[198] Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P.: Towards out-of-distribution generalization: A survey. arXiv preprint arXiv:2108.13624 (2021)

[199] Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 122–132

[200] Gómez-Chova, L., Tuia, D., Moser, G., Camps-Valls, G.: Multimodal classification of remote sensing images: A review and future directions. Proceedings of the IEEE **103** (2015) 1560–1584

[201] Arras, L., Osman, A., Samek, W.: Clevr-xai: a benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion **81** (2022) 14–40

[202] Garcia, N., Ye, C., Liu, Z., Hu, Q., Otani, M., Chu, C., Nakashima, Y., Mitamura, T.: A dataset and baselines for visual question answering on art. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer (2020) 92–108

[203] Yao, Y., Zhang, J., Shen, F., Hua, X., Xu, J., Tang, Z.: Exploiting web images for dataset construction: A domain robust approach. IEEE Transactions on Multimedia **19** (2017) 1771–1784

[204] Chen, H., Liu, M., Zhao, Y., Yan, X., Yan, D., Cheng, J.: G-miner: an efficient task-oriented graph mining system. In: Proceedings of the Thirteenth EuroSys Conference. (2018) 1–12

[205] Ribeiro, M.T., Guestrin, C., Singh, S.: Are red roses red? evaluating consistency of question-

answering models. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. (2019) 6174–6184

[206] Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Vqa-lol: Visual question answering under the lens of logic. In: ECCV, Springer (2020) 379–396

[207] Jimenez, C.E., Russakovsky, O., Narasimhan, K.: Carets: A consistency and robustness evaluative test suite for vqa. arXiv preprint arXiv:2203.07613 (2022)

[208] Zhang, W., Geng, S., Fu, Z., Zheng, L., Jiang, C., Hong, S.: Metava: Curriculum meta-learning and pre-fine-tuning of deep neural networks for detecting ventricular arrhythmias based on ecgs. arXiv preprint arXiv:2202.12450 (2022)

[209] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)

[210] Zhang, W., Yu, J., Zhao, W., Ran, C.: Dmrfnet: deep multimodal reasoning and fusion for visual question answering and explanation generation. Information Fusion **72** (2021) 70–79

[211] Wu, J., Chen, L., Mooney, R.J.: Improving vqa and its explanations\\by comparing competing explanations. arXiv preprint arXiv:2006.15631 (2020)

[212] Li, Q., Fu, J., Yu, D., Mei, T., Luo, J.: Tell-and-answer: Towards explainable visual question answering using attributes and captions. arXiv preprint arXiv:1801.09041 (2018)

[213] Wu, J., Hu, Z., Mooney, R.J.: Generating question relevant captions to aid visual question answering. arXiv preprint arXiv:1906.00513 (2019)

[214] Li, Q., Tao, Q., Joty, S., Cai, J., Luo, J.: Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 552–567

[215] Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8779–8788

[216] Gan, C., Li, Y., Li, H., Sun, C., Gong, B.: Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. (2017) 1811–1820

[217] You, Z., Ye, J., Li, K., Xu, Z., Wang, P.: Adversarial noise layer: Regularize neural network by adding noise. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 909–913

[218] Chan-Hon-Tong, A.: An algorithm for generating invisible data poisoning using adversarial noise that breaks image classification deep learning. Machine Learning and Knowledge Extraction **1** (2018) 192–204

[219] Agarwal, A., Vatsa, M., Singh, R., Ratha, N.: Cognitive data augmentation for adversarial defense via pixel masking. Pattern Recognition Letters **146** (2021) 244–251

[220] Yang, C.H.H., Liu, Y.C., Chen, P.Y., Ma, X., Tsai, Y.C.J.: When causal intervention meets adversarial examples and image masking for deep neural networks. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 3811–3815

[221] Walmer, M., Sikka, K., Sur, I., Shrivastava, A., Jha, S.: Dual-key multimodal backdoors for visual question answering. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. (2022) 15375–15385

[222] Sheng, S., Singh, A., Goswami, V., Magana, J., Thrush, T., Galuba, W., Parikh, D., Kiela, D.: Human-adversarial visual question answering. Advances in Neural Information Processing Systems **34** (2021) 20346–20359

[223] Li, L., Lei, J., Gan, Z., Liu, J.: Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 2042–2051

[224] Medsker, L.R., Jain, L.: Recurrent neural networks. Design and Applications **5** (2001) 64–67

[225] Zhao, H., Jia, J., Koltun, V.: Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 10076–10085

[226] Peng, L., Yang, Y., Zhang, X., Ji, Y., Lu, H., Shen, H.T.: Answer again: Improving vqa with cascaded-answering model. IEEE Transactions on Knowledge and Data Engineering **34** (2020) 1644–1655

[227] Zhang, M.L., Li, Y.K., Yang, H., Liu, X.Y.: Towards class-imbalance aware multi-label learning. IEEE Transactions on Cybernetics **52** (2020) 4459–4471

[228] Feldmann, A., Whitt, W.: Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. Performance evaluation **31** (1998) 245–279

[229] Zhao, J., Zhang, X., Wang, X., Yang, Y., Sun, G.: Overcoming language priors in vqa via adding visual module. Neural Computing and Applications **34** (2022) 9015–9023

[230] Yang, C., Feng, S., Li, D., Shen, H., Wang, G., Jiang, B.: Learning content and context with language bias for visual question answering. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE (2021) 1–6

[231] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine **35** (2018) 53–65

[232] Lao, M., Guo, Y., Liu, Y., Lew, M.S.: A language prior based focal loss for visual question answering. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE (2021) 1–6

[233] Ouyang, N., Huang, Q., Li, P., Cai, Y., Liu, B., Leung, H.f., Li, Q.: Suppressing biased samples for robust vqa. IEEE Transactions on Multimedia **24** (2021) 3405–3415

[234] Yan, H., Liu, L., Feng, X., Huang, Q.: Overcoming language priors with self-contrastive learning for visual question answering. Multimedia Tools and Applications (2022) 1–16

[235] Niu, Y., Zhang, H.: Introspective distillation for robust question answering. Advances in Neural Information Processing Systems **34** (2021) 16292–16304

[236] Liang, Z., Jiang, W., Hu, H., Zhu, J.: Learning to contrast the counterfactual samples for robust visual question answering. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). (2020) 3285–3292

[237] Zheng, Q., Wang, C., Liu, D., Wang, D., Tao, D.: Cross-modal contrastive learning for robust reasoning in vqa. arXiv preprint arXiv:2211.11190 (2022)

[238] Kervadec, C., Antipov, G., Baccouche, M., Wolf, C.: Estimating semantic structure for the vqa answer space. arXiv preprint arXiv:2006.05726 (2020)

[239] Teney, D., Abbasnejad, E., van den Hengel, A.: Unshuffling data for improved generalization in visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. (2021) 1417–1427

[240] Liu, J., Fan, C., Zhou, F., Xu, H.: Be flexible! learn to debias by sampling and prompting for robust visual question answering. Information Processing & Management **60** (2023) 103296

[241] Wu, Y., Zhao, Y., Zhao, S., Zhang, Y., Yuan, X., Zhao, G., Jiang, N.: Overcoming language priors in visual question answering via distinguishing superficially similar instances. arXiv preprint arXiv:2209.08529 (2022)

[242] Teney, D., Abbasnejad, E., Kafle, K., Shrestha, R., Kanan, C., Van Den Hengel, A.: On the value of out-of-distribution testing: An example of goodhart's law. Advances in Neural Information Processing Systems **33** (2020) 407–417

[243] Zhu, X., Mao, Z., Liu, C., Zhang, P., Wang, B., Zhang, Y.: Overcoming language priors with self-supervised learning for visual question answering. arXiv preprint arXiv:2012.11528 (2020)

[244] Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM International Conference on Multimedia. MM '14, New York, NY, USA, Association for Computing Machinery (2014) 7–16

[245] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)

[246] Ren, Shaoqing axnd He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the International Conference on Neural Information Processing Systems. (2015) 91–99

[247] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). (2014) 1532–1543

[248] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9** (1997) 1735–1780

[249] Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv e-prints **abs/1412.3555** (2014) Presented at the Deep Learning workshop at NIPS2014.

[250] Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: The IEEE International Conference on Computer Vision (ICCV). (2015)

[251] Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31** (1966) 279–311

[252] Liu, Y., Guo, Y., Liu, L., Bakker, E.M., Lew, M.S.: Cyclematch: A cycle-consistent embedding network for image-text matching. Pattern Recognition **93** (2019) 365–379

[253] Eisenschtat, A., Wolf, L.: Linking image and text with 2-way nets. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)

[254] Huang, Y., Wu, Q., Wang, L.: Learning semantic concepts and order for image and sentence matching. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 6163–6171

[255] Guo, Y., Liu, Y., De Boer, M.H., Liu, L., Lew, M.S.: A dual prediction network for image captioning. In: 2018 IEEE international conference on multimedia and expo (ICME), IEEE (2018) 1–6

[256] Mingrui, L., Yanming, G., Hui, W., Xin, Z.: Cross-modal multistep fusion network with co-attention for visual question answering. IEEE Access (2018)

[257] Shen, W., Wang, B., Jiang, Y., Wang, Y., Yuille, A.: Multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection. In: The IEEE International Conference on Computer Vision (ICCV). (2017)

[258] Li, J., Liang, X., Li, J., Wei, Y., Xu, T., Feng, J., Yan, S.: Multistage object detection with group recursive learning. IEEE Transactions on Multimedia **20** (2018) 1645–1655

[259] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 2223–2232

[260] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15** (2014) 1929–1958

[261] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[262] Wu, C., Liu, J., Wang, X., Dong, X.: Object-difference attention: A simple relational attention for visual question answering. In: Proceedings of the ACM International Conference on Multimedia. (2018) 519–527

[263] Wu, C., Liu, J., Wang, X., Dong, X.: Chain of reasoning for visual question answering. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., eds.: Advances in Neural Information Processing Systems 31. Curran Associates, Inc. (2018) 275–285

[264] Peng, L., Yang, Y., Wang, Z., Wu, X., Huang, Z.: Cra-net: Composed relation attention network for visual question answering. In: Proceedings of the 27th ACM international conference on multimedia. (2019) 1202–1210

[265] Zhang, Y., Hare, J.S., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. International Conference on Learning Representations (2018)

[266] Cadene, R., Ben-younes, H., Cord, M., Thome, N.: Murel: Multimodal relational reasoning for visual question answering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)

[267] Soujanya, P., Erik, C., Rajiv, B., Amir, H.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37** (2017) 98–125

[268] Jing, C., Wu, Y., Zhang, X., Jia, Y., Wu, Q.: Tvercoming language priors in vqa via decomposed linguistic representations. In: AAAI. (2020)

[269] Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: ICCV. (2017)

[270] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S.: Deep learning for visual understanding: A review. Neurocomputing **187** (2016) 27–48

[271] Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017)

[272] Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. ICML (2009)

[273] Akira, F., Dong, H.P., Daylen, Y., Anna, R., Trevor, D., Marcus, R.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2016)

[274] Ben-younes, H., Cadene, R., Thome, N., Cord, M.: Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). (2019)

[275] Lao, M., Guo, Y., Pu, N., Chen, W., Liu, Y., Lew, M.S.: Multi-stage hybrid embedding fusion network for visual question answering. Neurocomputing **423** (2021) 541–550

[276] Tudor Ionescu, R., Alexe, B., Leordeanu, M., Popescu, M., Papadopoulos, D.P., Ferrari, V.: How hard can it be? estimating the difficulty of visual search in an image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)

[277] Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)

[278] Platanios, E.A., Stretcu, O., Neubig, G., Póczos, B., Mitchell, T.M.: Competence-based curriculum learning for neural machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). (2019)

[279] Zhang, X., Shapiro, P., Kumar, G., McNamee, P., Carpuat, M., Duh, K.: Curriculum learning for domain adaptation in neural machine translation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). (2019)

[280] Li, Q., Huang, S., Hong, Y., Zhu, S.C.: A competence-aware curriculum for visual concepts learning via question answering. In: Proceedings of the European Conference on Computer Vision (ECCV). (2020)

[281] Wu, T., Li, X., Li, Y.F., Haffari, R., Qi, G., Zhu, Y., Xu, G.: Curriculum-meta learning for order-robust continual relation extraction. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). (2021)

[282] Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M.E., Stone, P.: Curriculum learning for reinforcement learning domains: A framework and survey. Journal of Machine Learning Research **21** (2020) 1–50

[283] Tang, Y., Huang, S.: Self-paced active learning: Query the right thing at the right time. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). (2019)

[284] Jiang, L., Meng, D., Mitamura, T., Hauptmann, A.G.: Easy samples first: Self-paced reranking for zero-example multimedia search. In: Proceedings of the 22nd ACM Interna-

tional Conference on Multimedia. ACM MM (2014)

[285] Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Proceedings of the International Conference on Neural Information Processing Systems. NIPS (2010)

[286] Zhang, D., Meng, D., Li, C., Jiang, L., Zhao, Q., Han, J.: A self-paced multiple-instance learning framework for co-saliency detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)

[287] Choi, J., Jeong, M., Kim, T., Kim, C.: Pseudo-labeling curriculum for unsupervised domain adaptation. In: Proceedings of the British Machine Vision Conference (BMVC). (2019)

[288] Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E **69** (2004) 066138

[289] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)

[290] Shinji, W., Takaaki, H., Jonathan, L., John, R.H.: Student-teacher network learning with enhanced features. In: Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2017)

[291] Grand, G., Belinkov, Y.: Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In: Proceedings of the 2nd Workshop on Shortcomings in Vision and Language (SiVL) at NAACL-HLT 2019. (2019)

[292] Teney, D., Abbasnedjad, E., van den Hengel, A.: Learning what makes a difference from counterfactual examples and gradient supervision. In: Proceedings of the European Conference on Computer Vision (ECCV). (2020)

[293] Zhu, L., Xu, Z., Yang, Y., Hauptmann, A.G.: Uncovering the temporal context for video question answering. IJCV **124** (2017) 409–421

[294] Fayek, H.M., Johnson, J.: Temporal reasoning via audio question answering. IEEE-ACM T AUDIO SPE **28** (2020) 2283–2294

[295] Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A.: A survey on causal inference. TKDD **15** (2021) 1–46

[296] Glymour, M., Pearl, J., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)

[297] Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: ICASSP. (2017) 131–135

[298] Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., Gan, C.: Beyond rnns: Positional self-attention with co-attention for video question answering. In: AAAI. Volume 33. (2019) 8658–8665

[299] Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., Huang, H.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: CVPR. (2019) 1999–2007

[300] Schwartz, I., Schwing, A.G., Hazan, T.: A simple baseline for audio-visual scene-aware dialog. In: CVPR. (2019) 12548–12558

[301] Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: ICCV. (2021) 1686–1697

[302] Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100, 000+ questions for machine comprehension of text. In: EMNLP. (2016) 2383–2392

[303] Yu, L., Park, E., Berg, A.C., Berg, T.L.: Visual madlibs: Fill in the blank description generation and question answering. In: ICCV. (2015) 2461–2469

[304] Lee, D., Cheon, Y., Han, W.S.: Regularizing attention networks for anomaly detection in visual question answering. In: AAAI. Volume 35. (2021) 1845–1853

[305] Ye, K., Kovashka, A.: A case study of the shortcut effects in visual commonsense reasoning.

In: AAAI. Volume 35. (2021) 3181–3189

[306] Tanaka, R., Nishida, K., Yoshida, S.: Visualmrc: Machine reading comprehension on document images. In: AAAI. Volume 35. (2021) 13878–13888

[307] Pearl, J.: Causal inference in statistics: An overview. Statistics surveys **3** (2009) 96–146

[308] Keele, L.: The statistics of causal inference: A view from political methodology. Political Analysis **23** (2015) 313–335

[309] Richiardi, L., Bellocco, R., Zugna, D.: Mediation analysis in epidemiology: methods, interpretation and bias. International journal of epidemiology **42** (2013) 1511–1519

[310] Li, Y., Wang, X., Xiao, J., Ji, W., Chua, T.S.: Invariant grounding for video question answering. In: CVPR. (2022) 2928–2937

[311] Yang, X., Zhang, H., Qi, G., Cai, J.: Causal attention for vision-language tasks. In: CVPR. (2021) 9847–9857

[312] Lao, M., Guo, Y., Chen, W., Pu, N., Lew, M.S.: Vqa-bc: Robust visual question answering via bidirectional chaining. In: ICASSP. (2022) 4833–4837

[313] Lao, M., Guo, Y., Liu, Y., Chen, W., Pu, N., Lew, M.S.: From superficial to deep: Language bias driven curriculum learning for visual question answering. In: ACM MM. (2021) 3370–3379

[314] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV, Springer (2020) 121–137

[315] Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., Fu, J.: Seeing out of the box: End-to-end pre-training for vision-language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12976–12985

[316] McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation. Volume 24. Elsevier (1989) 109–165

[317] De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE transactions on pattern analysis and machine intelligence **44** (2021) 3366–3385

[318] Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113** (2019) 54–71

[319] Pu, N., Chen, W., Liu, Y., Bakker, E.M., Lew, M.S.: Lifelong person re-identification via adaptive knowledge accumulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 7901–7910

[320] Hu, W., Lin, Z., Liu, B., Tao, C., Tao, Z.T., Zhao, D., Ma, J., Yan, R.: Overcoming catastrophic forgetting for continual learning via model adaptation. In: International conference on learning representations. (2019)

[321] Javed, K., White, M.: Meta-learning representations for continual learning. Advances in neural information processing systems **32** (2019)

[322] de Masson D'Autume, C., Ruder, S., Kong, L., Yogatama, D.: Episodic memory in lifelong language learning. Advances in Neural Information Processing Systems **32** (2019)

[323] Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., Wang, H.: Ernie 2.0: A continual pre-training framework for language understanding. In: Proceedings of the AAAI conference on artificial intelligence. Volume 34. (2020) 8968–8975

[324] Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems **32** (2020) 604–624

[325] Song, J., Guo, Y., Gao, L., Li, X., Hanjalic, A., Shen, H.T.: From deterministic to generative: Multimodal stochastic rnns for video captioning. IEEE transactions on neural networks and learning systems **30** (2018) 3047–3058

[326] Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 3608–3617

[327] Berriel, R., Lathuillere, S., Nabi, M., Klein, T., Oliveira-Santos, T., Sebe, N., Ricci, E.: Budget-aware adapters for multi-domain learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 382–391

[328] Shrestha, R., Kafle, K., Kanan, C.: Answer them all! toward universal visual question answering models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 10472–10481

[329] Chao, W.L., Hu, H., Sha, F.: Cross-dataset adaptation for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5716–5725

[330] Xu, Y., Chen, L., Cheng, Z., Duan, L., Luo, J.: Open-ended visual question answering by multi-modal domain adaptation. arXiv preprint arXiv:1911.04058 (2019)

[331] Wang, K., Herranz, L., van de Weijer, J.: Continual learning in cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 3628–3638

[332] Yan, S., Hong, L., Xu, H., Han, J., Tuytelaars, T., Li, Z., He, X.: Generative negative text replay for continual vision-language pretraining. In: ECCV. (2022) 22–38

[333] Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision (2021)

[334] Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2017) 2935–2947

[335] Chen, W., Liu, Y., Pu, N., Wang, W., Liu, L., Lew, M.S.: Feature estimations based correlation distillation for incremental image retrieval. IEEE Transactions on Multimedia (2021)

[336] Chen, W., Xu, H., Pu, N., Liu, Y., Lao, M., Liu, L., Wang, W., Lew, M.S.: Lifelong fine-grained image retrieval. IEEE Transactions on Multimedia (2022)

[337] Liu, Y., Hong, X., Tao, X., Dong, S., Shi, J., Gong, Y.: Model behavior preserving for class-incremental learning. IEEE Transactions on Neural Networks and Learning Systems (2022)

[338] Gero, K., Kedzie, C., Reeve, J., Chilton, L.: Low-level linguistic controls for style transfer and content preservation. arXiv preprint arXiv:1911.03385 (2019)

[339] Joyce, J.M.: Kullback-leibler divergence. In: International encyclopedia of statistical science. Springer (2011) 720–722

[340] Yu, L., Yazici, V.O., Liu, X., Weijer, J.v.d., Cheng, Y., Ramisa, A.: Learning metrics from teachers: Compact networks for image embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 2907–2916

[341] Allen, D.M.: Mean square error of prediction as a criterion for selecting variables. Technometrics **13** (1971) 469–475

[342] Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 1365–1374

[343] Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. arXiv preprint arXiv:1905.09418 (2019)

[344] Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning, PMLR (2021) 5583–5594

[345] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[346] Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)

[347] Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX, Springer (2020) 104–120

[348] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)

[349] Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: European Conference on Computer Vision. (2018) 532–547

[350] Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multimodal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. (2018) 0–0

[351] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114** (2017) 3521–3526

[352] Park, D., Hong, S., Han, B., Lee, K.M.: Continual learning by asymmetric loss approximation with single-side overestimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 3335–3344

[353] Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 7096–7104

[354] Zhou, Y., Ji, R., Sun, X., Su, J., Meng, D., Gao, Y., Shen, C.: Plenty is plague: Fine-grained learning for visual question answering. IEEE transactions on pattern analysis and machine intelligence **44** (2019) 697–709

[355] Bara, C.P., Ping, Q., Mathur, A., Thattai, G., MV, R., Sukhatme, G.S.: Privacy preserving visual question answering. arXiv preprint arXiv:2202.07712 (2022)

[356] Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al.: Advances and open problems in federated learning. Foundations and Trends® in Machine Learning **14** (2021) 1–210

[357] Hu, R., Andreas, J., Darrell, T., Saenko, K.: Explainable neural computation via stack neural module networks. In: Proceedings of the European conference on computer vision (ECCV). (2018) 53–69

[358] Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. (2019)

[359] Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 11336–11344

[360] Jiang, J., Liu, Z., Liu, Y., Nan, Z., Zheng, N.: X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. In: Proceedings of the 29th ACM International Conference on Multimedia. (2021) 199–208

[361] Lao, M., Pu, N., Liu, Y., He, K., Bakker, E.M., Lew, M.S.: Coca: Collaborative causal regularization for audio-visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 37. (2023) 12995–13003

[362] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2** (2020) 429–450

[363] Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 10713–10722

[364] Acar, D.A.E., Zhao, Y., Navarro, R.M., Mattina, M., Whatmough, P.N., Saligrama, V.: Federated learning based on dynamic regularization. arXiv preprint arXiv:2111.04263 (2021)

[365] Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., Xu, C.Z.: Feddc: Federated learning with non-iid data via local drift decoupling and correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 10112–10121

[366] Li, D., Wang, J.: Fedmd: Heterogenous federated learning via model distillation. arXiv preprint arXiv:1910.03581 (2019)

[367] Zhang, J., Guo, S., Ma, X., Wang, H., Xu, W., Wu, F.: Parameterized knowledge transfer for personalized federated learning. Advances in Neural Information Processing Systems **34** (2021) 10092–10104

[368] Wu, C., Wu, F., Lyu, L., Huang, Y., Xie, X.: Communication-efficient federated learning via knowledge distillation. Nature communications **13** (2022) 2032

[369] You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., Hsieh, C.J.: Large batch optimization for deep learning: Training bert in 76 minutes. arXiv preprint arXiv:1904.00962 (2019)

[370] Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv:1910.10699 (2019)

[371] Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. (2022) 11953–11962

[372] Li, L., Fan, Y., Tse, M., Lin, K.Y.: A review of applications in federated learning. Computers & Industrial Engineering **149** (2020) 106854

[373] Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N., Jararweh, Y.: Federated learning review: Fundamentals, enabling technologies, and future applications. Information processing & management **59** (2022) 103061

List of Abbreviations

Abbreviation	Full Name / Short Definition
DCNNs	Deep Convolutional Neural Networks / A regularized version of multilayer perceptrons based on convolution kernels
VQA	Visual Question Answering / A multimodal task
AVQA	Audio-Visual Question Answering / A multimodal task
MHEF	Multi-stage Hybrid Embedding Fusion / A multi-stage fusion approache for vision-language interactions for VQA
LP-Focal	Language Prior based Focal Loss / A debiased loss function for VQA
LBCL	Language Bias driven Curriculum Learning / An anti-bias learning framework for VQA
COCA	COllabrative CAusal regularization / An regularization strategy to mitigate multiple shortcut biases
MDL-VQA	Multi-Domain Lifelong VQA / An VQA setting over lifelong learning across multiple visual domains
SCD	Self-Critical Distillation / A replay-free continual learning approach to alleviate catastrophic forgetting
FedVQA	Federated VQA / A VQA setting of federated learning over heterogeneous scenes
FedP3	Federated Pairwise Preference Preserving / A knowledge preserving framework for personalized federated VQA
Faster-RCNN	A real-time visual-object detection network
ViLT	A Transformer for multimodal inputs / Vision-and-Language Transformer Vision-and-Language Transformer Without Convolution or Region Supervision
LwF	Learning without Forgetting / A teacher-student distillation architecture to avoid catastrophic forgetting issues
BCE	Binary Cross Entropy / A metric that tracks incorrect labeling of the data class by a model
SSL	Self-Supervised Learning / a learning process where the model trains itself to learn one part of the input from another part of the input.
OOD	out-of distribution / it is used to describe a dataset where its train and test splits are in different label distributions.
IID	identically distributed (data)
MMD	Maximum Mean Discrepancy
CIL	Class-Incremental Learning
DIL	Domain-Incremental Learning
DKD	Decoupled Knowledge Distillation // Knowledge distillation based on two teacher models
GCNs	Graph Convolutional Networks
MSE	Mean Squared Error / A metric to measure the distances between two features
KLD	Kullback–Leibler divergence / A metric to measure the distance between two distributions