REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[2] Y. Liu *et al.*, “Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models.” arXiv, Apr. 08, 2023. doi: 10.48550/arXiv.2304.01852.

[3] S. Antol *et al.*, “VQA: Visual Question Answering,” presented at the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433. Accessed: Aug. 03, 2022. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html

[4] C. Zhou *et al.*, “A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT.” arXiv, May 01, 2023. doi: 10.48550/arXiv.2302.09419.

[5] S. Huang *et al.*, “Language Is Not All You Need: Aligning Perception with Language Models.” arXiv, Mar. 01, 2023. doi: 10.48550/arXiv.2302.14045.

[6] Y. Li, Z. Li, K. Zhang, R. Dan, and Y. Zhang, “ChatDoctor: A Medical Chat Model Fine-tuned on LLaMA Model using Medical Domain Knowledge.” arXiv, Apr. 18, 2023. doi: 10.48550/arXiv.2303.14070.

[7] T. Li *et al.*, “CancerGPT: Few-shot Drug Pair Synergy Prediction using Large Pre-trained Language Models.” arXiv, Apr. 17, 2023. doi: 10.48550/arXiv.2304.10946.

[8] S. Wu *et al.*, “BloombergGPT: A Large Language Model for Finance.” arXiv, Mar. 30, 2023. doi: 10.48550/arXiv.2303.17564.

[9] K. Alibabaei *et al.*, "A Review of the Challenges of Using Deep Learning Algorithms to Support Decision-Making in Agricultural Activities," *Remote Sens.*, vol. 14, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/rs14030638.

[10] "Artificial Intelligence in Agriculture: A Review." https://ieeexplore.ieee.org/abstract/document/9432187/ (accessed Jan. 27, 2023).

[11] S. Manmadhan and B. C. Kovoor, "Visual question answering: a state-of-the-art review," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5705–5745, Dec. 2020, doi: 10.1007/s10462-020-09832-7.

[12] C. Jackulin and S. Murugavalli, "A comprehensive review on detection of plant disease using machine learning and deep learning approaches," *Meas. Sens.*, vol. 24, p. 100441, Dec. 2022, doi: 10.1016/j.measen.2022.100441.

[13] W. Albattah, M. Nawaz, A. Javed, M. Masood, and S. Albahli, "A novel deep learning method for detection and classification of plant diseases," *Complex Intell. Syst.*, vol. 8, no. 1, pp. 507–524, Feb. 2022, doi: 10.1007/s40747-021-00536-1.

[14] Y. Borhani, J. Khoramdel, and E. Najafi, "A deep learning based approach for automated plant disease classification using vision transformer," *Sci. Rep.*, vol. 12, no. 1, Art. no. 1, Jul. 2022, doi: 10.1038/s41598-022-15163-0.

[15] V. Gandhi, G. Kumar, and R. Marsh, "Agroindustry for rural and small farmer development: issues and lessons from India," *Int. Food Agribus. Manag. Rev.*, vol. 2, no. 3, pp. 331–344, Sep. 1999, doi: 10.1016/S1096-7508(01)00036-2.

[16] A. Trunk, H. Birkel, and E. Hartmann, "On the current state of combining human and artificial intelligence for strategic organizational decision making,"

Bus. Res., vol. 13, no. 3, pp. 875–919, Nov. 2020, doi: 10.1007/s40685-020-00133-x.

[17] A. Terhorst and A. Morshed, "AgroKnowledgeBase (AKB) for plant diseases: Poppy plant use case," Oct. 2013, pp. 1–6. Accessed: Apr. 21, 2023. [Online]. Available: http://eprints.rclis.org/21017/

[18] H. Song, L. Dong, W.-N. Zhang, T. Liu, and F. Wei, "CLIP Models are Few-shot Learners: Empirical Studies on VQA and Visual Entailment." arXiv, Mar. 14, 2022. doi: 10.48550/arXiv.2203.07190.

[19] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "KAT: A Knowledge Augmented Transformer for Vision-and-Language." arXiv, May 05, 2022. doi: 10.48550/arXiv.2112.08614.

[20] H. Orchi, M. Sadik, and M. Khaldoun, "On Using Artificial Intelligence and the Internet of Things for Crop Disease Detection: A Contemporary Survey," *Agriculture*, vol. 12, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/agriculture12010009.

[21] A. Ahmad, D. Saraswat, and A. El Gamal, "A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools," *Smart Agric. Technol.*, vol. 3, p. 100083, Feb. 2023, doi: 10.1016/j.atech.2022.100083.

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015. doi: 10.48550/arXiv.1409.1556.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." arXiv, Dec. 10, 2015. doi: 10.48550/arXiv.1512.03385.

[24] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.” arXiv, Sep. 11, 2020. doi: 10.48550/arXiv.1905.11946.

[25] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.

[26] “Papers with Code - Reliable Deep Learning Plant Leaf Disease Classification Based on Light-Chroma Separated Branches.” https://paperswithcode.com/paper/reliable-deep-learning-plant-leaf-disease (accessed May 08, 2023).

[27] V. Kodali and D. Berleant, “Recent, Rapid Advancement in Visual Question Answering: a Review,” in *2022 IEEE International Conference on Electro Information Technology (eIT)*, May 2022, pp. 139–146. doi: 10.1109/eIT53891.2022.9813988.

[28] N. Xie, F. Lai, D. Doran, and A. Kadav, “Visual Entailment: A Novel Task for Fine-Grained Image Understanding.” arXiv, Jan. 20, 2019. doi: 10.48550/arXiv.1901.06706.

[29] S. Barra, C. Bisogni, M. De Marsico, and S. Ricciardi, “Visual question answering: Which investigated applications?,” *Pattern Recognit. Lett.*, vol. 151, pp. 325–331, Nov. 2021, doi: 10.1016/j.patrec.2021.09.008.

[30] M. Malinowski and M. Fritz, “Towards a Visual Turing Challenge.” arXiv, May 05, 2015. doi: 10.48550/arXiv.1410.8027.

[31] R. Y. Zakari, J. W. Owusu, H. Wang, K. Qin, Z. K. Lawal, and Y. Dong, “VQA and Visual Reasoning: An Overview of Recent Datasets, Methods and

Challenges.” arXiv, Dec. 26, 2022. Accessed: Jan. 20, 2023. [Online]. Available: http://arxiv.org/abs/2212.13296

[32] “ImageNet classification with deep convolutional neural networks | Communications of the ACM.” https://dl.acm.org/doi/abs/10.1145/3065386 (accessed May 08, 2023).

[33] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN.” arXiv, Jan. 24, 2018. doi: 10.48550/arXiv.1703.06870.

[34] S. Manmadhan and B. C. Kovoor, “Visual question answering: a state-of-the-art review,” *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5705–5745, Dec. 2020, doi: 10.1007/s10462-020-09832-7.

[35] S. Uppal *et al.*, “Multimodal Research in Vision and Language: A Review of Current and Emerging Trends.” arXiv, Dec. 21, 2020. doi: 10.48550/arXiv.2010.09522.

[36] H. Zhu, Z. Wang, Y. Shi, Y. Hua, G. Xu, and L. Deng, “Multimodal Fusion Method Based on Self-Attention Mechanism,” *Wirel. Commun. Mob. Comput.*, vol. 2020, p. e8843186, Sep. 2020, doi: 10.1155/2020/8843186.

[37] K. Liu, Y. Li, N. Xu, and P. Natarajan, “Learn to Combine Modalities in Multimodal Deep Learning.” arXiv, May 29, 2018. doi: 10.48550/arXiv.1805.11730.

[38] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 457–468. doi: 10.18653/v1/D16-1044.

[39] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, “MUTAN: Multimodal Tucker Fusion for Visual Question Answering.” arXiv, May 18, 2017. doi: 10.48550/arXiv.1705.06676.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[41] A. Radford and K. Narasimhan, “Improving Language Understanding by Generative Pre-Training,” 2018. Accessed: May 09, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035

[42] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in Vision: A Survey,” *ACM Comput. Surv.*, vol. 54, no. 10s, p. 200:1-200:41, Sep. 2022, doi: 10.1145/3505244.

[43] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal Learning with Transformers: A Survey.” arXiv, Jun. 13, 2022. doi: 10.48550/arXiv.2206.06488.

[44] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “VisualBERT: A Simple and Performant Baseline for Vision and Language.” arXiv, Aug. 09, 2019. doi: 10.48550/arXiv.1908.03557.

[45] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019. Accessed: Feb. 04, 2023. [Online]. Available: https://papers.nips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html

[46] I. Ilievski and J. Feng, “Multimodal Learning and Reasoning for Visual Question Answering,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 03, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/f61d6947467ccd3aa5af24db320235dd-Abstract.html

[47] F. Gardères, M. Ziaeefard, B. Abeloos, and F. Lecue, “ConceptBert: Concept-Aware Representation for Visual Question Answering,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 489–498. doi: 10.18653/v1/2020.findings-emnlp.44.

[48] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, “KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14111–14121. Accessed: Feb. 04, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Marino_KRISP_Integrating_Implicit_and_Symbolic_Knowledge_for_Open-Domain_Knowledge-Based_VQA_CVPR_2021_paper.html

[49] Y. Ding, J. Yu, B. Liu, Y. Hu, M. Cui, and Q. Wu, “MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-Based Visual Question Answering,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5089–5098. Accessed: Jan. 28, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Ding_MuKEA_Multi

modal_Knowledge_Extraction_and_Accumulation_for_Knowledge-Based_Visual_Question_CVPR_2022_paper.html

[50] X. Wang, Q. He, J. Liang, and Y. Xiao, “Language Models as Knowledge Embeddings.” arXiv, Jun. 25, 2022. doi: 10.48550/arXiv.2206.12617.

[51] *Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding*, (Mar. 22, 2019). Accessed: Aug. 04, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=d3g3pCHqgo8

[52] A. Vaswani *et al.*, “Attention Is All You Need.” arXiv, Dec. 05, 2017. doi: 10.48550/arXiv.1706.03762.

[53] “PlantVillage.” https://plantvillage.psu.edu/diseases (accessed May 09, 2023).

[54] “Wikipedia.” https://www.wikipedia.org/ (accessed May 09, 2023).