CS6240: Semester 2: 2023/24
List of Papers to Read and Present

Lecture 2 (Visual Object Recognition, Detection & Vision-Language Models) (Invited Speaker: Ji Wei):

- **Image Recognition**

  (Must-Read) A Dosovitskiy, L Beyer, A Kolesnikov et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

  (To-Read) Z Liu, Y Lin, Y Cao et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV 2021.

  (Must-Read, Best Paper) K He, X Zhang, S Ren & J Sun. Deep Residual Learning for Image Recognition. CVPR 2016.

- **Object Detection:**

  (Must-Read) Z Liu, H Hu, Y Lin, Z Yao, Z Xie, Y Wei, J Ning, Y Cao, Z Zhang, L Dong, F Wei &B Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. CVPR 2022.

  (Must-Read) S Ren, K He, R Girshick, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. TPAMI 2016.

  (To-Read) N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov & S Zagoruyko. End-to-End Object Detection with Transformers. ECCV 2020.

- **Vision-Language Models:**

  (Must-Read) X Chen, X Wang, S Changpinyo, et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. ICLR 2023.

  (Must-Read) W Wang, H Bao, L Dong et al. Image as a Foreign Language: BEiT Pre-training for Vision and Vision-Language Tasks. CVPR 2023.

  (To-Read) L Ma, ZD Lu, LF Shang & H Li. Multimodal Convolutional Neural Networks for Matching Image and Sentence. ICCV 2015.

Lecture 3 (Semantic and Temporal Segmentation and Relation Grounding):

- **Semantic Segmentation: Presenter: Cheng Yi; Reader: Wu Yihang**

  (Must-Read) A Kirillov, E Mintun, N Ravi, et al. Segment anything. arXiv 2023.

  (To-Read) K He, G Gkioxari, P Dollár & R Girshick (2017). Mask R-CNN. ICCV 2017.

- **Temporal Segmentation: Presenter: Ngyuyen Thong Thanh; Reader: Dai Yuhe**

  (Must-Read) Z Hou, W Zhong, L Ji, D Gao, K Yan, et al. CONE: An Efficient Coarse-to-fiNE Alignment Framework for Long Video Temporal Grounding. ACL 2023.

  (Must-Read) LA Hendricks, O Wang, E Shechtman, J Sivic, T Darrell & B Russell. Localizing Moments in Video with Temporal Language. EMNLP 2018.

- **Relation Grounding: Presenter: Zheng Jingnan; Reader: Dibyadip Chatterjee**

  (Must-Read) Y Cong, MY Yang & B Rosenhahn. RelTR: Relation Transformer for Scene Graph Generation. TPAMI. 2023.

  (To-Read) B Dai, Y Zhang & D Lin. Detecting Visual Relationships with Deep Relational Networks. CVPR 2017.

Lecture 4 (Cross-modal Alignment and Multimodal Scene Graph):

- **Cross-modal Alignment: Presenter: Chai Zenghao; Reader: Stefan Putra Lionar**

  (Must-Read) J Li, D Li, S Savarese & S Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023.

  (Must-Read): A Radford, JW Kim, C Hallacy, et al. Learning Transferable Visual Models from Natural Language Supervision. ICML 2021.

  (To-Read): L Qu, M Liu, J Wu, Z Gao & L Nie. Dynamic Modality Interaction Modeling for Image-Text Retrieval. SIGIR 2021.

- **Multimodal Scene Graph: Presenter: Dibyadip Chatterjee; Reader: He Qiyuan**

  (Must-Read) J Yang, W Peng, X Li et al. Panoptic Video Scene Graph Generation. CVPR 2023.

  (To-Read) K Tang, Y Niu, J Huang et al. Unbiased Scene Graph Generation From Biased Training. CVPR 2020.

  (First Dataset, Must-Read) R Krishna, Y Zhu, O Groth, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. IJCV 2017.

Lecture 5 (Few-Shot, Meta and Causal Learning):

- **Few-show Learning: Presenter: Mehdi Yamini; Reader: Liu Nian**

  (Must-Read) X Liu, Y Zheng, Z Du, et al. GPT understands, too. AI Open 2023.

  (Must-Read) O Vinyals, C Blundell, T. Lillicrap, K Kavukcouglu & D Wierstra. Matching Networks for One Shot Learning. NeurIPS 2016.

  (To-Read) F Sung, Y Yang, L Zhang, T Xiang, P. Torr & T Hospedales. Learning to Compare: Relation Network for Few-Shot Learning. CVPR 2018.

- **Meta Learning: Presenter: Bai Jinbin; Reader: Qin Hangyu**

  (Must-Read): J Snell, K Swersky & RS Zemel. Prototypical Networks for Few-Shot Learning. NeurIPS 2017.

  (Must Read): C Finn, P Abbeel & S Levine: Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML 2017.

- **Causal Learning: Presenter: Xing Naili; Reader: Sui Yuan**

  (Must-Read) Y Niu, K Tang, H Zhang, et al. Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

  (To-Read) X Yang, H Zhang, G Qi & J Cai. Causal Attention for Vision-Language Tasks. CVPR 2021.

Lecture 6 (Image/ Video QA, and Reasoning):

- **Image QA and Reasoning: Presenter: Wu Yihang; Reader: Cheng Yi**

  (Classic SOTA-1) P Anderson, X He, C Buehler, D Teney, M Johnson, S Gould, & L Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. CVPR 2018.

  (Classic SOTA-2) J Lu, J Yang, D Batra, et al. Hierarchical Question-Image Co-Attention for Visual Question Answering. NeurIPS 2016.

(Popular Dataset, To-Read) D A Hudson & C D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. CVPR 2019.

- **Video QA and Reasoning: Presenter: He Yingzhi; Reader: Yannis Mohamed Christian Montreuil**

  (Must-Read) A J Piergiovanni, K Morton, W Kuo, et al. Video Question Answering with Iterative Video-Text Co-Tokenization. ECCV 2022.

  (Must-Read) A Yang, A Miech, J Sivic, I Laptev & C Schmid. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. CVPR 2021.

  (Survey: To-Read): Y Zhong, W Ji, J Xiao, Y Li, W Deng & TS Chua. Video Question Answering: Datasets, Algorithms and Challenges. EMNLP 2022.

  (First Dataset, Must-Read): J Xiao, X Shang, A Yao & TS Chua. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. CVPR 2021

Lecture 7 (Diffusion Models for MM Generation):

- **Diffusion Models for Image Generation: Presenter: He Qiyuan; Reader: Chai Zenghao**

  (Must-Read) J Ho, A Jain & P Abbeel. Denoising Diffusion Probabilistic Models. NeurIPS 2020

  (Must-Read) J Song, C Meng & S Ermon. Denoising Diffusion Implicit Models. ICLR 2021.

  (To-Read) Y Song, J Sohl-Dickstein, D. P. Kingma et al. Score-Based Generative Modeling through Stochastic Differential Equations. ICLR 2021.

- **Condition-based Diffusion Models: Presenter: Chen Xihao; Reader: Nguyen Thong Thanh**

  (Must-Read) X Shen, C Du, T Pang, et al. Fine Tuning Text-to-Image Diffusion Models for Fairness. ICLR 2024.

  (Must-Read) R Rombach, A Blattmann, D Lorenz, et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022.

  (To-Read, Best Paper) L Zhang, A Rao & M Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. ICCV 2023.

- **Image/Video Editing & Personalization: Presenter: Lin Xinyu; Reader: Zheng Jingnan**

  (Must-Read) A Hertz, R Mokady & J Tenenbaum, et al. Prompt-to-Prompt Image Editing with Cross Attention Control. ICLR 2023.

  (To-Read) H Ouyang, Q Wang, Y Xiao, et al. CoDeF: Content Deformation Fields for Temporally Consistent Video Processing. arXiv 2023.

  (Must-Read) N Ruiz, Y Li & V Jampani, et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR 2023.

Lecture 8 (Large Multimodal Foundation Model):

- **Large Visual Foundation Models: Presenter: Qin Hangyu; Reader: He Yingzhi**

  (Must-Read) H Liu, C Li, Q Wu, et al. Visual Instruction Tuning. NeurIPS 2023.

  (SOTA) R Dong, C Han, Y Peng, et al. DreamLLM: Synergistic Multimodal Comprehension and Creation. ICLR 2024.

  (Must-Read) D Zhu, J Chen, X Shen, et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. ICLR 2024.

  (Background) H Touvron, T Lavril, G Izacard, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv 2023.

  (Background) T Brown, B Mann, N Ryder, et al. Language Models are Few-Shot Learners. arXiv 2022.

- **Pixel Grounding Large Multimodal Models: Presenter: Stefan Putra Lionar; Reader: Cao Xiao**

  (SOTA) Y. Yuan, W. Li, J. Liu, et al. Osprey: Pixel Understanding with Visual Instruction Tuning. Preprint arXiv 2023.

  (Must-Read) H. Rasheed, M. Maaz, S. Shaji, et al. Glamm: Pixel grounding large multimodal model. Preprint arXiv 2023.

  (To-Read) Z. Ren, Z. Huang, Y. Wei, et al. PixelLM: Pixel Reasoning with Large Multimodal Model. Preprint arXiv 2023.

Lecture 9 (Multimodal Dialogues & NExT-GPT):

- **Multimodal Dialogues: Presenter: Sun Pengzhan; Reader: Xing Naili**

  (SOTA) T Gong, C Lyu, S Zhang, et al. Multimodal-GPT: A vision and language model for dialogue with humans. Preprint arXiv 2023.

  (Must-Read) Q Sun, Y Wang, C Xu, et al. Multimodal dialogue response generation. ACL 2022.

  (To Read) K Shuster, EM Smith, D Ju & J Wesron. Multi-Modal Open-Domain Dialogue. EMNLP 2021.

  (To Read) T L Wu, S Kottur, A Madotto, et al. SIMMC-VR: A Task-oriented Multimodal Dialog Dataset with Situated and Immersive VR Streams. ACL 2023.

- **Multimodal Instruction Tuning: Presenter: Liu Nian; Reader: Bai Jinbin**

  (Must-Read)J. Han, R. Zhang, W. Shao, et al. Imagebind-llm: Multi-modality instruction tuning. Preprint arXiv 2023.

  (To Read) Z. Xu, Y. Shen, L. Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. Preprint arXiv 2022.

  (To Read) Z. Yin, J. Wang, J. Cao, et al. LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark. Preprint arXiv 2023.

- **NExT-GPT: (Invited Speaker: Yu Shengqiong)**

  (Must-Read) S Wu, H Fei, L Qu, et al. NExT-GPT: Any-to-any Multimodal LLM. Preprint arXiv 2023.

Lecture 10 (Responsible AI: Trust, Safety, Privacy & Biased in MM):

- **Hallucination: Presenter: Cao Xiao; Reader: Chen Xihao**

  (SOTA) S Dhuliawala, M Komeili, J Xu, et al. Chain-of-Verification Reduces Hallucination in Large Language Models. Preprint arXiv 2023.

  (Must-Read) P Manakul, A Liusie, M J F Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. Preprint arXiv 2023.

  (Background) Y Zhang, Y Li, L Cui, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Preprint arXiv 2023.

- **Privacy: Presenter: Dai Yuhe; Reader: Lin Xinyu**

  (SOTA) S Kim, S Yun, H Lee, et al. Propile: Probing privacy leakage in large language models. Preprint arXiv 2023.

  (Must-Read) J Huang, H Shao, K C C Chan. Are Large Pre-Trained Language Models Leaking Your Personal Information? ACL 2022.

  (Background) H Shao, J Huang, S Zheng, et al. Quantifying Association Capabilities of Large Language Models and Its Implications on Privacy Leakage. EACL 2023.

- **Bias: Presenter: Yannis Mohamed Christian Montreuil; Reader: Mehdi Yamini**

  (Must-Read) P Schramowski, M Brack, B Deiseroth, et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. CVPR 2023.

  (Must-Read) Q Li, X Wang, Z Wang, et al. Be causal: De-biasing Social Network Confounding in Recommendation. ACM TKDD 2023.

  (To-Read) A S Luccioni, C Akiki, M Mitchell, et al. Stable bias: Analyzing societal representations in diffusion models. Preprint arXiv 2023.

Lecture 11 (Multimodal Event (& Fashion) Detection & Forecasting):

- **Multimodal Event Detection: Presenter: Sui Yuan; Reader: Sun Pengzhan**

  (SOTA) M. Li, R. Xu, S. Wang, et al. Clip-event: Connecting text and images with event structures. CVPR 2022.

  (To-Read) Li Z, Ding X, Liu T. Constructing Narrative Event Evolutionary Graph for Script Event Prediction. IJCAI 2018.

  (Must-Read) T Zhang, S Whitehead, H Zhang, et al. Improving event extraction via multimodal integration. ACM MM 2017.

- **Multimodal Fashion Forecasting: (Invited Speaker: Ma Yunshan)**

  (Must-Read) U Mall, K Matzen, B Hariharan, et al. Geostyle: Discovering fashion trends and events. ICCV 2019.

  (SOTA) Hsiao W L, Grauman K. From culture to clothing: Discovering the world events behind a century of fashion images. ICCV 2021.

  (Must-Read) Ma Y, Yang X, Liao L, et al. Who, where, and what to wear? extracting fashion knowledge from social media. ACM MM 2019.

Lecture 12 (MM Recommendation & Presentation): (Invited Speaker: Wang Wenjie)

- **Rank-based MM Recommendation:**

(SOTA) Q Li, X Wang, Z Wang, et al. Be-Causal: De-biasing Social Network Confounding in Recommendation. ACM TKDD 2023.

(Must-Read) Z Deng, J Li, Z Guo, et al. Multi-view Multi-aspect Neural Networks for Next-basket Recommendation. ACM SIGIR 2023.

(To-Read) Y Ding, Y Ma, W Wong & TS Chua. Leveraging Two Types of Global Graph for Sequential Fashion Recommendation. ICMR 2021.

(Background) W Wang, Y Zhang, H Li, et al. Causal recommendation: Progresses and future directions. ACM SIGIR 2023.

- **Generative MM Recommendation:**

  (SOTA) B Yin, J Xie, Y Qin, Z Ding, Z Feng, X Li, W Lin. Heterogeneous Knowledge Fusion: A Novel Approach for Personalized Recommendation via LLM. RecSys 2023.

  (Must-Read) W Wang, X Lin, F Feng, et al. Generative recommendation: Towards next-generation recommender paradigm. Preprint arXiv 2023.

  (Background) L Wu, Z Zheng, Z Qiu, H Wang, H Gu, T Shen, C Qin, C Zhu, H Zhu, Q Liu, H Xiong, E Chen. A Survey on Large Language Models for Recommendation. Preprint arXiv 2023.