13.06.22

Recap of papers read

Technische Universität München + Johns Hopkins Whiting School of Engineering

On-device Real-time Hand Gesture Recognition

- Good:
  - Reinforces that NNs are better for borderline cases than “classic” heuristics-based methods
  - Introduces the idea building a 3D model to estimate pose before training an network on the estimated pose to classify gestures

- Bad:
  - Hand gestures are different from head gestures
  - Thus: lots of the explanations are irrelevant for the thesis (i.e. how to determine the key points for a hand model)
  - No explanation for why their NN looks the way it does (3 fully connected layers with 50 neurons each)

Sung, G., Sokal, K., Uboweja, E., Bazarevsky, V., Baccash, J., Bazavan, E. G., ... & Grundmann, M. (2021). On-device Real-time Hand Gesture Recognition. *arXiv preprint arXiv:2111.00038*.

Deep multimodal representation learning: A survey

- Good:
  - Very thorough analysis of current research in multimodal representation learning
  - Introduces distinction between joint representation, coordinated representation and encoder-decoder type networks
  - Specifically talks about the problem of missing data (and “filling in” data)
  - Good explanations for types of multimodal learning
  - Introduces the concept of attention mechanism (focusing on meaningful features)

- Bad:
  - Lots of irrelevant information (for the thesis)

Guo, W., Wang, J., & Wang, S. (2019). Deep multimodal representation learning: A survey. IEEE Access, 7, 63373-63394.

Learning transferable visual models from natural language supervision

- Good:
    - Includes link to code
    - Introduces idea of using natural language supervision to determine if data from two modalities belong together
    - Their model (CLIP) can be used for a lot of different image classification tasks
- Bad:
    - Very long
    - Application to my problem not that clear
    - Model pre-trained with loads of data (millions of images)
    - Lots of irrelevant information (like scaling, zero-shot learning application etc.)

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.

Multimodal deep learning models for early detection of Alzheimer’s disease stage

- Good:
  - Reinforces that DL strategies are more reliant than “classic” ML approaches (like SVMs, decision trees, random forests and kNN)
  - Reinforces that multi-modality data is better than single-modality data
  - Combines a lot of very different modalities (i.e. clinical data and imaging data)
  - Detailed explanation and evaluation of networks used
  - Offers method for dealing with missing data (by masking it with zeros)

- Bad:
  - Masking with zeros is simple, but there might be better ways to deal with missing data
  - Naive approach to combining modalities (concatenation)

Venugopalan, J., Tong, L., Hassanzadeh, H. R., & Wang, M. D. (2021). Multimodal deep learning models for early detection of Alzheimer’s disease stage. Scientific reports, 11(1), 1-13.

Cross-modal scene networks

- Good:
  - Concerns multi-modal learning and domain adaptation
  - Their model performs well even if data for some modalities is missing
  - Detailed explanation of network(s)
  - Finetuning by freezing later layers instead of first layers
    - finetuning to a modality
  - Very detailed explanations of different methods with evaluations concerning performance for different modalities

- Bad:
  - Their data is not paired between modalities
    - very different approach
  - Chosen modalities are closer to each other (except for text descriptions)

Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., & Torralba, A. (2017). Cross-modal scene networks. IEEE transactions on pattern analysis and machine intelligence, 40(10), 2303-2314.

Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training

- Good:
  - Tries to solve the exact same problem as me
  - Introduces regularization parameter to avoid negative knowledge transfer
  - Simultaneous training of both modality networks
  - Detailed explanation of mathematical basis as well as networks (and hyperparameters)
  - Implemented in TensorFlow
  - Promising results (improvement of recognition over unimodal model)

- Bad:
  - Hand gesture recognition instead of head gesture recognition

Abavisani, M., Joze, H. R. V., & Patel, V. M. (2019). Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1165-1174).