CSE 555 Fall 2020:
Project Description

September 9, 2020

1 Introduction

The goal of this project is to develop a learning system based on what you have learned in the pattern recognition class, which operates on a given dataset. Please feel free to confront and solve issues that are not covered in class. There are 4 projects to choose from (described below), each of which consists of 2 parts. For the first part, you are expected to solve one baseline, which is for you to get familiar with the whole pipeline and build necessary fundamentals. The second part should be your main focus, which is an extension to your baseline.

2 Collaboration

The project should be carried out by teams of 2-3 students (preferred is 2). We do expect that projects done with 3 students have more impressive write-up, results, and novelty than projects done with 2 students.

3 Freedom

You may select one of the projects from the list below. You will need to inform TA (you project choice, team details) Lin Huang about your choices by Friday, Sept 18, 2020. You will have significant freedom in designing what you will do for your baseline and extension. The techniques in terms of data pre-processing, what algorithms to use, training strategy, and evaluation, *etc.* are up to you. Don't be afraid to think outside of the box. There will be no strict answer to each of the project. You are welcome to come to the course TA's (Lin Huang) office hours for discussion any related questions.

4 Project Description

4.1 Project A: Multimodal Fake News Classification

Fake news has altered society in negative ways in politics and culture. It has adversely affected both online social network systems as well as offline communities and conversations. Using automatic machine learning classification models is an efficient way to combat the widespread dissemination of fake news. While research in the area of fake news detection is of high importance for society, most of the existing methods opt for uni-mode approaches, as analyzing the multimodal content is often more complicated and availability of a reasonably sized dataset poses extra challenge.

For this problem, the goal for you is to explore some baseline methods discussed in [1], where both text and image mode component of a sample are represented using their respective mode-specific feature descriptor. The mode specific features are later concatenated directly to obtain a multimodal feature representative for each sample and then propose improvement. More specifically:

- Baseline: The baseline you will implement will be the methods described by the authors of the dataset [1]. Please feel free to choose your preferred word embeddings to represent the text component. You may also utilize several concepts learned in the course (like dimensionality reduction etc.) to demonstrate the effects on the results.

- Extension: Based on your understanding, you are expected to improve or come up with a different learning structure to better handle the problem.

4.1.1 Datasets

- Fakeddit Dataset [1, 2]

4.1.2 Resources

You may follow the relevant resources [1, 3, 4, 5, 6] for your implementation.

4.2 Project B: Point Cloud Analysis

A point cloud is an important type of 3D geometric and irregular data format, corresponding to a unordered set of vectors. While most algorithms in Pattern Recognition focus on input data with regular format like RGB images, depth maps, 3D voxels, or speech sequence, it is also necessary to come up with algorithms which can directly reason about input data with irregular format.

For this problem, the goal for you is to explore (geometric) deep learning architectures, which can explicitly consume the input 3D point cloud and extract essential features for 3D object classification. Specifically:

- Baseline: The baseline you will implement is to utilize the classic geometric learning framework, called PointNet [7] for 3D object classification. As the very first deep learning structure to handle the 3D point sets, you should investigate its benefits as well as drawbacks for modeling unordered point sets.
- Extension: After implementation of the baseline, you should understand the whole pipeline for point cloud classification. Based on your understanding, you are expected to improve or come up with a different learning structure to better handle this classification task.

4.2.1 Datasets

- ModelNet40 [8] for 3D object classification.

4.2.2 Resources

Please follow the relevant resources [9, 10, 7, 11, 12, 13, 14, 15] for your implementation.

4.3 Project C: Multimodal Emotion Recognition in a Multispeaker Video

Emotion detection in conversations has been widely analyzed in the Natural Language Processing community. However, emotion detection in multi-speaker conversations instead of traditional two-speaker conversations in existing studies is significantly under explored. In this project, you will develop a multimodal (text transcript, face expression, and visual component) recognition framework to predict the speaker's emotion in every utterance, which demonstrates significant speaker sensitive dependence.

For this problem, the goal is to develop and evaluate a multimodal recognition framework compared to its uni-mode (text transcript and video frame visual component) counter parts and propose improvement. More specifically:

- Baseline: The baseline you will implement will be a multimodal recognition framework compared that analyze both text transcript and video frame visual component together within an integrated setting to offer a more accurate prediction on the speaker's expression in a given utterance. Please feel free to choose your preferred word embeddings to represent the text component. You may also utilize several concepts learned in the course (like dimensionality reduction etc.) to demonstrate the effects on the results.
- Extension: Based on your understanding, you are expected to improve the baseline multimodal framework to come up with a different learning structure to better handle the problem.

4.3.1 Datasets

- MELD Dataset [16, 17]

4.3.2 Resources

You may follow the relevant resources [3, 4, 18, 5, 6] for your implementation.

4.4 Project D: Action Recognition

Human action recognition is mainly about obtaining the underlying patterns of motions taking place in the video, which is a standard computer vision problem. Most actions typically require information from multiple consecutive frames to recognize its category. Therefore, the main idea lies in how to jointly make use of the spatial and temporal aspects in order to extract video level representation.

For this problem, the goal for you is to propose techniques, which can fully exploit both information in order to predict accurate human action category (like running or jumping) given a short fragment of a video presenting human activity (with arbitrary video length). Specifically:

- Baseline: The baseline you will implement is to utilize the classic Long-term Recurrent Convolutional Network (LRCN) [19], which combines convolutional layers and recurrent network, LSTM [20] for human action recognition in videos.
- Extension: After implementation of the baseline, you should understand the whole pipeline for action recognition as well as the basic recurrent modeling mechanism for sequential data. Based on your understanding, you are expected to improve or come up with a different techniques to better capture spatiotemporal information.

4.4.1 Datasets

- UCF-101 dataset [21].

4.4.2 Resources

You may follow the relevant resources [19, 22, 23, 24, 20, 25, 26, 27, 28, 29, 30, 31] for your implementation.

4.5 Project E: Cross-modal Representation Learning

Standard Auto-Encoders (AE) are a unsupervised learning algorithms which are mainly used for learning efficient and representative data codings. Typically, AE is designed with a bottleneck inside the network to force a much more compressed and knowledgeable latent representation of the original input. Variational Auto-Encoders (VAEs) [32] have been one of the most popular deep generative models, which can learn consistent as well as continuous latent representations of the given dataset without supervision.

Generally, both can be used for feature extraction, dimensionality reduction, *etc.* Moreover, VAEs can also be used to synthesize unseen data consistently as well as obtain disentangled representation with reasonable regularization.

For this problem, the goal for you is to extend the basic AE (or VAE) framework for learning a cross-modal latent space for human hands. You will use AE (or VAE) to train a single and unified latent space across different hand-related modalities. In other words, different hand modalities (like 3D hand pose, hand depth map, and hand RGB image) corresponding to the same hand pose should be embedded into the same representation, sharing one latent space.

- Baseline: The baseline you will implement is to utilize the standard AE (or VAE) to find latent space for each hand modality, including hand RGB image, 3D hand pose parameterized as 3D joint locations, and hand depth map. This baseline is a standard use of AE (or VAE) for embedding and reconstructing the input dataset. You will end up with 3 AEs (or 3 VAEs) for each modality.

- Extension: After implementation of each baseline, you should understand the intuitions and the mathematics behind AE (or VAEs). Then, you are expected to find a shared latent space across the 3 hand modalities instead of finding one for each. Your final model should be able to embed human hands from different modalities and to reconstruct them either in the same or in a different modality. (Your model is also able to directly estimate hand depth maps as well as 3D hand poses given hand RGB images. If you use VAEs, new samples should also be generated consistently.)

4.5.1 Datasets

You can use either one of the following 2 datasets for implementation.

- Stereo Tracking Benchmark Dataset (STB) [33], which is a real-world dataset.
- Rendered Hand Pose Dataset (RHD) [34], which is a synthetic dataset.

4.5.2 Resources

For this project, it is important for you to dive deep into the mathematics behind AE (or VAE). Note also you only need to choose AE or VAE for this project. Please follow the relevant resources [32, 35, 36, 37, 38, 39, 40, 41, 42, 43] for your implementation.

5 General Requirements & Notes

- For each project, it is better to go through relevant literature and resources listed above first and then implement the baseline before doing the extension part.
- For large dataset, you could use a subset as long as you can make sure it is large enough to verify your points. Please clearly state this type of use of dataset in your report if you do so.
- Be careful to keep your final testing set uncorrupted, by setting it aside at the beginning, or by using cross-validation procedures appropriately.

6 Deliverables

Your final report is required to be between 5-6 pages using the provided template. You can add other necessary supplementary materials (not counted toward the report page limit). You will submit your report as a PDF file, your supplementary material as a separate PDF or ZIP file, and your source code as another ZIP file. All will be submitted in UBlearns. We will provide more submission instructions as the deadline nears. Examples of components to put in your report and supplementary material are listed below:

Report

- Title, Author(s)
- Abstract
- Introduction
- Description of the Baseline methods and how your proposed method improves over those baselines
- Experiments and Results
- Conclusion
- Contributions: Please include a section that describes what each team member has contributed to the project and it is not counted toward your report page limit
- References: Do not miss any references and it is not counted toward your report page limit

Supplementary Material

- More analysis
- More experiments
- More visualization results, videos, or demos

7 Grading Criteria

Since the project is open-ended, and the work you will do is largely up to you. Grading criteria will include: understanding and interpretation (of approach, algorithms used, and results) of the problem which will be evaluated from your writeup (weight: 20%); and Preparedness and clarity in Presentation (weight: 20%); code implementation (weight: 25%) ; final performance and significance of the work (weight: 25%) ; and novelty of the new idea introduced to improve the baseline (weight: 15%). Midterm project report is due on the following week of midterm. The final presentation will be in Zoom, when you would be asked to present your work, show demo, which will be recorded for grading purposes.

Please schedule a biweekly appointment (15 mins) with the TA Lin Huang to discuss your regular progress and receive his feedback.

Exceptional projects will be awarded with Bonus points.

8 Honor Code

Your report and code should reflect your own work done specifically for this class. You may consult any materials like papers, books, or any publicly available resources for implementation, code, and ideas that you might want to use as a basis for your projects, as long as your clearly quote and reference in your write-up and code.

References

[1] K. Nakamura, S. Levy, and W. Y. Wang, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," 2019.

[2] K. Nakamura. [Online]. Available: https://fakeddit.netlify.app/

[3] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[4] "bert-as-service." [Online]. Available: https://github.com/hanxiao/bert-as-service

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[7] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017.

[8] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *CVPR*, 2015.

[9] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[10] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[11] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *NIPS*, 2017.

[12] R. Klokov and V. Lempitsky, "Escape from cells: Deep kd-networks for the recognition of 3d point cloud models," in *CVPR*, 2017.

[13] A. Komarichev, Z. Zhong, and J. Hua, "A-cnn: Annularly convolutional neural networks on point clouds," in *CVPR*, 2019.

[14] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics*, vol. 38, no. 5, pp. 1–12, 2019.

[15] Y. Liu. [Online]. Available: https://github.com/Yochengliu/awesome-point-cloud-analysis

[16] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[17] S. Poria. [Online]. Available: https://affective-meld.github.io/

[18] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations." in *IJCAI*, 2019.

[19] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[22] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[23] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[24] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *ICML*, 2013, pp. 1310–1318.

[25] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[26] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019, pp. 6202–6211.

[27] jinwchoi. [Online]. Available: https://github.com/jinwchoi/awesome-action-recognition

[28] V. Maskara. [Online]. Available: https://towardsdatascience.com/literature-survey-human-action-recognition-cc7c3818a99a

[29] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *arXiv preprint arXiv:1801.07455*, 2018.

[30] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[31] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," *arXiv preprint arXiv:1711.08200*, 2017.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[33] J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, and Q. Yang, "A hand pose tracking benchmark from stereo matching," in *ICIP*, 2017.

[34] C. Zimmermann and T. Brox, "Learning to estimate 3D hand pose from single RGB images," in *ICCV*, 2017.

[35] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[36] I. Shafkat. [Online]. Available: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

[37] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *arXiv preprint arXiv:1812.05069*, 2018.

[38] J. Klys, J. Snell, and R. Zemel, "Learning latent subspaces in variational autoencoders," in *NIPS*, 2018.

[39] M. Rad, M. Oberweger, and V. Lepetit, "Feature mapping for learning fast and accurate 3d pose inference from synthetic images," in *CVPR*, 2018.

[40] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *CVPR*, 2018.

[41] L. Yang, S. Li, D. Lee, and A. Yao, "Aligning latent spaces for 3d hand pose estimation," in *ICCV*, 2019.

[42] L. Yang and A. Yao, "Disentangling latent hands for image synthesis and pose estimation," in *CVPR*, 2019.

[43] T. Theodoridis, T. Chatzis, V. Solachidis, K. Dimitropoulos, and P. Daras, "Cross-modal variational alignment of latent spaces," in *CVPRW*, 2020.