Final project proposal

Announced on April 3, 2019

1. Important dates

- **Proposal** due date: **May 6**, 2019 (before midnight)
- **Presentation** date: **June 19 & 21**, 2019 (in class)
- **Report** deadline: **June 29**, 2019 (before midnight)

2. Register your team

- Please form a group of **4-6 people and send an email to TA**, including
    - Team members **IDs & names** (please mark the team leader)
    - **List of preferred tasks** in order of preference for your team (from 1st to 9th)
- Students who have no teams after April 10 will be randomly assigned to a team with a task
- Finalized teams and corresponding tasks will be announced on iLMS system

3. Tasks

- There are three domains to choose from: FinTech, biology, and computer vision
    - Each domain consists of three related tasks (a total of 9 tasks)
    - For FinTech and biology domains, specific datasets will be assigned for you to work on; for computer vision, no datasets are specified

3.1 FinTech domain: Amazon

We adopt a dataset from Amazon which includes product reviews and metadata for 142.8 million reviews from May. 1996 to July, 2014 (http://jmcauley.ucsd.edu/data/amazon/). Please use the small dataset Movies and TV [5-core (1,697,533 reviews)]. The website above contains codes for reading in files and basic information for this dataset as follows:

file name: Movies_and_TV_5.json
reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin - ID of the product, e.g. 0000013714
reviewerName - name of the reviewer
helpful - helpfulness rating of the review, e.g. 2/3
reviewText - text of the review
overall - rating of the product (1-5)
summary - summary of the review
unixReviewTime - time of the review (unix time)
reviewTime - time of the review (raw)

3.1.1 Task 1: sentiment analysis

- **You need to use the review to the product to classify the user's feeling (positive/negative) for this product**
- You can use the pre-trained word and phrase embedding vectors
- You can refer to [1-1] for the detail about sentence classification

3.1.2 Task 2: product rating

- **Use review and score of the user to predict the rating of the product that the user hasn't bought**
- In this work, you can refer to [1-2] and [1-3] for more details about product rating

3.1.3 Task 3: product recommendation

- **Use items that a user has bought to predict what the user will buy and generate a product recommendation list for the user**
- [1-4] and [1-5] are papers about recommendation, you can use different models to generate recommendation list

3.1.4 Reference papers

- **[1-1]** Yoon Kim (2014). Convolutional Neural Networks for Sentence Classification.
- **[1-2]** Lei Zheng, Vahid Noroozi & Philip S. Yu (2017). Joint Deep Modeling of Users and Items Using Reviews for Recommendation.
- **[1-3]** Rose Catherinen & William Cohen (2017). TransNets: Learning to Transform for Recommendation
- **[1-4]** Shumpei Okura, Yukihiro Tagami, Shingo Ono & Akira Tajima (2017). Embedding-based News Recommendation for Millions of Users.
- **[1-5]** Badrul Sarwar, George Karypis, Joseph Konstan & John Riedl (2001). Item-Based Collaborative Filtering Recommendation Algorithms.

3.2 Biology domain: DeepSeq

Two datasets are provided here:

- *Structure Protein Sequence dataset* (https://www.rcsb.org/)
- *Protein Secondary Structure dataset* (https://academic.oup.com/bioinformatics/article/19/12/1589/258419)

Structure Protein Sequence dataset

This is a protein data set retrieved from Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB.

The constantly-growing PDB is a reflection of the research that is happening in laboratories across the world. This can make it both exciting and challenging to use the database in research and education. Structures are available for many of the proteins and nucleic acids involved in the central processes of life, so you can go to the PDB archive to find structures for ribosomes, oncogenes, drug targets, and even whole viruses. However, it can be a challenge to find the information that you need, since the PDB archives so many different structures. You will often find multiple structures for a given molecule, or partial structures, or structures that have been modified or inactivated from their native form.

There are two data files. Both are arranged on *structureId* of the protein:

- *pdb_data_no_dups.csv* contains protein metadata which include details on protein classification (protein family), extraction methods, etc. Columns include:
    - *structureId*
    - *classification*
    - *experimentalTechnique*
    - *macromoleculeType*
    - *residueCount*
    - *resolution*
    - *structureMolecularWeight*
    - *crystallizationMethod*
    - *crystallizationTempK*
    - *densityMatthews*
    - *densityPercentSol*
    - *pdbxDetails*
    - *phValue*
    - *publicationYear*
- *data_seq.csv* contains more than 400,000 protein structure sequences. Columns include:
    - *structureId* - Structure ID
    - *chainId* - Chain ID
    - *sequence* - protein sequence
    - *residueCount* - Number of residues (ATCG's)
    - *macromoleculeType* - Type of macromolecule

Protein Secondary Structure dataset

Protein secondary structure can be calculated based on its atoms' 3D coordinates once the protein's 3D structure is solved using X-ray crystallography or NMR. Commonly, DSSP is the tool used for calculating the secondary structure and assigns one of the following secondary structure types (https://swift.cmbi.umcn.nl/gv/dssp/index.html) to every amino acid in a protein:

- C: Loops and irregular elements (corresponding to the blank characters output by DSSP)
- E: β-strand
- H: α-helix
- B: β-bridge
- G: 3-helix
- I: π-helix
- T: Turn
- S: Bend

However, X-ray or NMR is expensive. Ideally, we would like to predict the secondary structure of a protein based on its primary sequence directly, which has had a long history. You can refer to [2-2] for more details about this dataset. Descriptions of its columns are as follows.

- *pdb_id* - the id used to locate its entry on https://www.rcsb.org/
- *chain_code* - when a protein consists of multiple peptides (chains), the chain code is needed to locate a particular one.
- *seq* - the sequence of the peptide
- *sst8* - the eight-state (Q8) secondary structure
- *sst3* - the three-state (Q3) secondary structure
- *len* - the length of the peptide
- *has_nonstd_aa* - whether the peptide contains nonstandard amino acids (B, O, U, X, or Z)

3.2.1 Task 1: protein family

- **Classify protein families based on their sequence of amino acids**
- Please use the first dataset for this task
- This work is based on the current success of deep learning models in natural language processing (NLP) and assumes that the protein sequences can be viewed as language, you can refer to [2-1] and [2-3] for more details

3.2.2 Task 2: macromolecule type

- **Classify macromolecule type based on their sequence of amino acids**
- Please use the first dataset for this task
- In this Task, there will be three different type of macromolecule (Protein, DNA/RNA hybrid, DNA). You can try to use the same way as Task 1 or use different Deep learning model (e.g CNN) for prediction
- You can refer to [2-1], [2-2] and [2-3] for more details

3.2.3 Task 3: secondary structure (Q3,Q8) of a chain

- **Predicting the secondary structure (Q3,Q8) of a chain.**
- Please use the second dataset for this task

- For the purpose of secondary structure prediction, you can select the target (Q3 or Q8) that you want to predict. You can use natural language-based deep learning algorithms for prediction (LSTM, seq2seq). It’s similar that tasks 1 and 2 view sequences as language.
- You can refer to [2-4] for more details

3.2.4 Reference papers

- **[2-1]** Timothy K. Lee, Tuan Nguyen, "Protein Family Classification with Neural Networks", 2016.
- **[2-2]** Nguyen, N. , Tran, V. , Ngo, D. , Phan, D. , Lumbanraja, F. , Faisal, M. , Abapihi, B. , Kubo, M. and Satou, K. (2016) DNA Sequence Classification by Convolutional Neural Network. Journal of Biomedical Science and Engineering, 9, 280-286. doi: 10.4236/jbise.2016.95021
- **[2-3]** Hou J, Adhikari B, Cheng J. DeepSF: deep convolutional neural network for mapping protein sequences to folds. Bioinformatics. 2017;34(8):1295-1303.
- **[2-4]**.Yang Y. et al. (2016) Sixty-five years of long march in protein secondary structure prediction: the final stretch? Brief. Bioinform., DOI: 10.1093/bib/bbw129.

3.3 Computer vision

No specific datasets are assigned for this domain. You can choose ANY dataset that is **CLOSELY** related to the task.

3.3.1 Task 1: style transfer

- **Please follow the structure mentioned in [3-1] and construct three subnetworks: style extraction, content extraction, and image generation networks**
- You can use pre-trained feature maps from VGG-19 as detailed in [3-2]
- You are free to design the following items
    - The structure of three subnetworks
    - Loss function for image generation networks

3.3.2 Task 2: domain adaptation & transfer learning

- **You should first train a model on a dataset then adapt/transfer to another dataset**
- An example mentioned in [3-3]: trained on MNIST and transferred to SVHN
- Various applications can be found in [3-4]

3.3.3 Task 3: image caption generation / visual question answering

- **Given an image, you are required to generate textual information for the image and/or answer a question about the image**
- Please refer to [3-5] & [3-6] for more details

- [3-7] is an example of a reversed version of this task

3.3.4 Reference papers

- **Style transfer**
    - **[3-1]** Gatys, L.A., Ecker, A. S., & Bethge, M. (2016). Image Style Transfer Using Convolutional Neural Networks. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2414-2423).
    - **[3-2]** Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556.
- **Domain adaptation & transfer learning**
    - **[3-3]** Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 17(1), 2096-2030.
    - **[3-4]** Csurka, G. (2017). Domain adaptation for visual applications: A comprehensive survey. arXiv preprint arXiv:1702.05374.
- **Image caption generation**
    - **[3-5]** Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). 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 (pp. 6077-6086).
    - **[3-6]** Lu, X., Wang, B., Zheng, X., & Li, X. (2018). Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing, 56(4), 2183-2195.
    - **[3-7]** Johnson, J., Gupta, A., & Fei-Fei, L. (2018). Image generation from scene graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1219-1228).

4. Scoring criteria

- Final project code/results (10%)
- Final project report (10%)
    - Each student should write your own report
- Final project presentation (Extra 2%)
    - Volunteer to present your project to the class
    - You need to prepare ≤ 10 mins presentation with slides