Advanced Topics in Deep Learning **28 Oct, 2020**

Final Project.

Instructor: Yuan Yao *Due: 23:59 Sunday 13 Dec, 2020*

1 Requirement

This project as a warm-up aims to explore feature extractions using existing networks, such as pre-trained deep neural networks and scattering nets, in image classifications with traditional machine learning methods.

1. Pick up ONE (or more if you like) favourite dataset below to work. If you would like to work on a different problem outside the candidates we proposed, please email course instructor about your proposal.
2. Team work: we encourage you to form small team, up to FOUR persons per group, to work on the same problem. Each team must submit:

   (a) ONE report, *with a clear remark on each person's contribution.* The report can be in the format of either a *technical report within 8 pages*, e.g. NIPS conference style (preferred format)

   `https://nips.cc/Conferences/2019/PaperInformation/StyleFiles`

   Python (Jupyter) Notebooks with a detailed documentation, or a *poster*, e.g.

   `https://github.com/yuany-pku/2017_math6380/blob/master/project1/DongLoXia_poster.pptx`

   (b) *ONE short presentation video within 10 mins*, e.g. in Youtube or Bilibili link. You may submit your presentation slides together with the video link to help understanding.
3. In the report, show your proposed scientific questions to explore and main results with a careful analysis supporting the results toward answering your problems. Remember: scientific analysis and reasoning are more important than merely the performance tables. Separate source codes may be submitted through email as a zip file, GitHub link, or as an appendix if it is not large.
4. Submit your report by email or paper version no later than the deadline, to the following address (deeplearning.math@gmail.com) with Title: Math 6380P: Project 2.

2 Kaggle in-class Contest: Nexperia Image Classification II

Nexperia (`https://www.nexperia.com/`) is one of the biggest Semi-conductor company in the world. They produce billions of semi-conductors every year. Meanwhile, a lot of unqualified devices are mixed with the good ones. Mass production makes it difficult for human workers to examine all of the products. Therefore, we would like to use modern machine learning methods, particularly deep learning, to help Nexperia pick out as many defect devices as possible while preserving the good ones, thus improving their yield rate.

Nexperia provided a dataset for Kaggle in-class contest that aims to classify images of semiconductor devices into two main classes, good and defect. For example, Fig. 1 shows a good example and a bad example. The Nexperia image dataset in the Kaggle contest contain 34457 train images (27420 good and 7039 bad) and 3830 test images with similar good-to-bad ratio. The key is to detect as many defect images as possible while not sacrificing too many passed ones. So on Kaggle contest, we adopt Area-Under-the-Curve (AUC) for ROC as the evaluation rule. Note that AUC values are in the range of $[0.5, 1]$, the higher, the better.

We note that this real world dataset may contain noisy labels, especially the images labeled as "good" possibly being "bad" ones in fact. We do not have ground truth on which labels are wrong, but you may pay additional attention to this issue. Bonus credits will be given to such explorations.

Checking the following Kaggle website for more details.

- `https://www.kaggle.com/c/semi-conductor-image-classification-second-stage`

To participate the contest, you need to login your Kaggle account first, then open the following invitation link and accept the Kaggle contest rule to download the data:

`https://www.kaggle.com/t/5cbb376414c24ba5a9a9183ac73d648f`

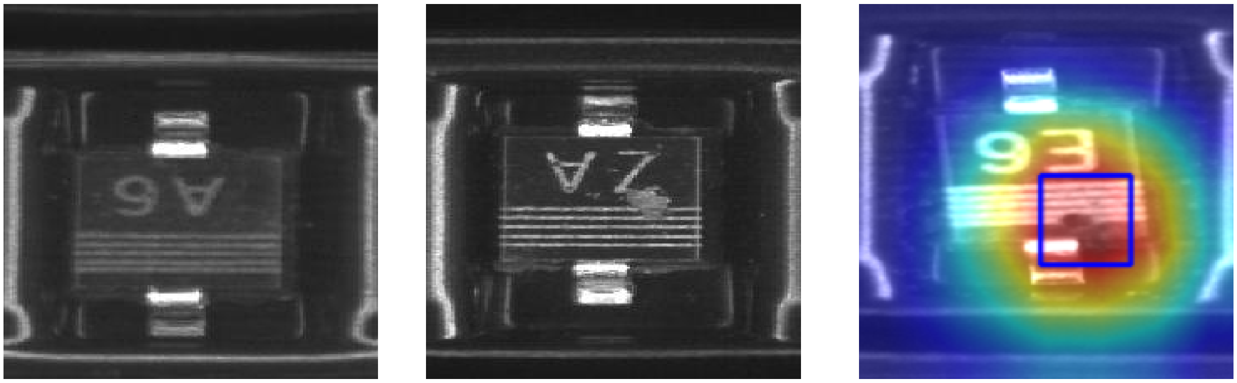


Figure 1: Examples of Nexperia image data. Left: a good example; Middle: a bad example; Right: a visualization based on heatmap

3 Kaggle Contest: Home Credit Default Risk

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data–including telco and transactional information–to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Visit the following website to join the competition.

`https://www.kaggle.com/c/home-credit-default-risk/`

4 Challenge from Project 1

The basic challenge is

- Feature extraction by scattering net with known invariants;
- Feature extraction by pre-trained deep neural networks, e.g. VGG19, and resnet18, etc.;
- Visualize these features using classical unsupervised learning methods, e.g. PCA/MDS, Manifold Learning, t-SNE, etc.;
- Compute the global mean of features $\Phi_i \in \mathbb{R}^p$ over all samples

  $$\boldsymbol{\mu}_G \triangleq \underset{i,c}{\mathrm{Ave}} \{\boldsymbol{\Phi}_{i,c}\}$$

  class-means

  $$\boldsymbol{\mu}_c \triangleq \underset{i}{\mathrm{Ave}} \{\boldsymbol{\Phi}_{i,c}\}, \quad c = 1, \ldots, C$$

  total covariance matrix

  $$\boldsymbol{\Sigma}_T \triangleq \underset{i,c}{\mathrm{Ave}} \left\{ (\boldsymbol{\Phi}_{i,c} - \boldsymbol{\mu}_G)(\boldsymbol{\Phi}_{i,c} - \boldsymbol{\mu}_G)^\top \right\}$$

  between class covariance

  $$\boldsymbol{\Sigma}_B \triangleq \underset{c}{\mathrm{Ave}} \left\{ (\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^\top \right\}$$

  and within class covariance

  $$\boldsymbol{\Sigma}_W \triangleq \underset{i,c}{\mathrm{Ave}} \left\{ (\boldsymbol{\Phi}_{i,c} - \boldsymbol{\mu}_c)(\boldsymbol{\Phi}_{i,c} - \boldsymbol{\mu}_c)^\top \right\}$$

such that $\mathbf{\Sigma}_T = \mathbf{\Sigma}_B + \mathbf{\Sigma}_W$. Verify the contraction of within class variation (NC1),

$$\mathrm{Tr}\left\{\mathbf{\Sigma}_W \mathbf{\Sigma}_B^{\dagger}\right\}/C;$$

closeness to equal-norms of class-means

$$\mathrm{Std}_c\left(\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2\right) / \, \mathrm{Avg}_c\left(\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2\right),$$

equal-angularity,

$$\mathrm{Std}_c\left(\cos_\mu\left(c, c'\right)\right) = \mathrm{Std}_c[\langle \boldsymbol{\mu}_c - \boldsymbol{\mu}_G, \boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G\rangle \, / \, (\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2 \, \|\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G\|_2)]$$

closeness to maximal-angle equiangularity,

$$\mathrm{Avg}_{c,c'}\left|\cos_\mu\left(c, c'\right) + 1/(C-1)\right|$$

- Image classifications using traditional supervised learning methods based on the features extracted, e.g. LDA, logistic regression, SVM, random forests, etc.;
- *Train the last layer or fine-tune the deep neural networks in your choice;
- Compare the results you obtained and give your own analysis on explaining the phenomena.

Below are two candidate datasets. Challenge marked by * above is only optional.

4.1 MNIST dataset – a Warmup

Yann LeCun's website contains original MNIST dataset of 60,000 training images and 10,000 test images.

`http://yann.lecun.com/exdb/mnist/`

There are various ways to download and parse MNIST files. For example, Python users may refer to the following website:

`https://github.com/datapythonista/mnist`

or MXNET tutorial on mnist

`https://mxnet.incubator.apache.org/tutorials/python/mnist.html`

4.2 Fashion-MNIST dataset

Zalando's Fashion-MNIST dataset of 60,000 training images and 10,000 test images, of size 28-by-28 in grayscale.

`https://github.com/zalandoresearch/fashion-mnist`

4.3 Identification of Raphael's paintings from the forgeries

The following data, provided by Prof. Yang WANG from HKUST,

`https://drive.google.com/folderview?id=0B-yDtwSjhaSCZ2FqN3AxQ3NJNTA&usp=sharing`

contains a 28 digital paintings of Raphael or forgeries. Note that there are both jpeg and tiff files, so be careful with the bit depth in digitization. The following file

`https://docs.google.com/document/d/1tMaaSIrYwNFZZ2cEJdx1DfFscIfERd5Dp2U7K1ekjTI/edit`

contains the labels of such paintings, which are

1 Maybe Raphael - Disputed

2 Raphael

3 Raphael

4 Raphael

5 Raphael

6 Raphael

7 Maybe Raphael - Disputed

8 Raphael

9 Raphael

10 Maybe Raphael - Disputed

11 Not Raphael

12 Not Raphael

13 Not Raphael

14 Not Raphael

15 Not Raphael

16 Not Raphael

17 Not Raphael

18 Not Raphael

19 Not Raphael

20 My Drawing (Raphael?)

21 Raphael

22 Raphael

23 Maybe Raphael - Disputed

24 Raphael

25 Maybe Raphael - Disputed

26 Maybe Raphael - Disputed

27 Raphael

28 Raphael

There are some pictures whose names are ended with alphabet like A's, which are irrelevant for the project.

The challenge of Raphael dataset is: can you exploit the known Raphael vs. Not Raphael data to predict the identity of those 6 disputed paintings (maybe Raphael)? Textures in these drawings may disclose the behaviour movements of artist in his work. One preliminary study in this project can be: *take all the known Raphael and Non-Raphael drawings and use leave-one-out test to predict the identity of the left out image; you may break the images into many small patches and use the known identity as its class.*

The following student poster report seems a good exploration

`https://github.com/yuany-pku/2017_CSIC5011/blob/master/project3/05.GuHuangSun_poster.pdf`

The following paper by Haixia Liu, Raymond Chan, and me studies Van Gogh's paintings which might be a reference for you:

`http://dx.doi.org/10.1016/j.acha.2015.11.005`

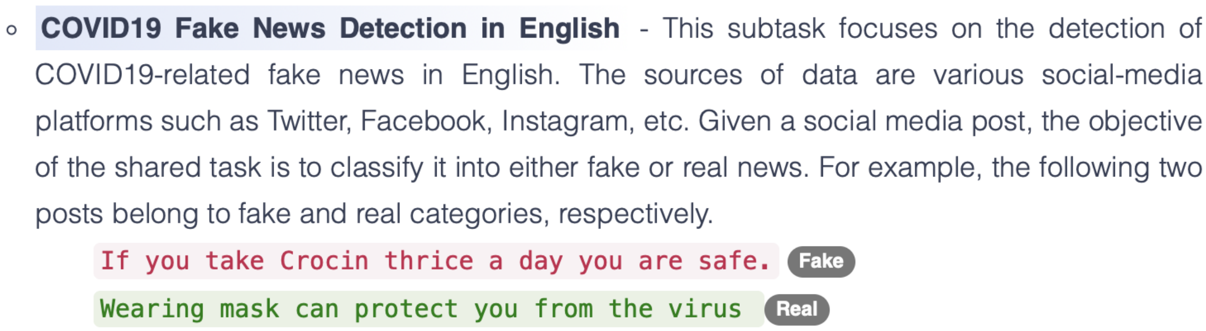
5 Self Proposals

5.1 COVID-19 Fake News Detection

This proposal is made by Yejin BANG, Samuel CAHYAWIJAYA, Etsuko ISHII, Ziwei JI. We would like to propose to work on AAAI 2021 shared task on COVID-19 Fake News Detection (`https://constraint-shared-task-2021.github.io/`) for the final project of MATH6380P.

The data size is as following:

- Train: 6420
- Valid: 2140



○ **COVID19 Fake News Detection in English** - This subtask focuses on the detection of COVID19-related fake news in English. The sources of data are various social-media platforms such as Twitter, Facebook, Instagram, etc. Given a social media post, the objective of the shared task is to classify it into either fake or real news. For example, the following two posts belong to fake and real categories, respectively.

`If you take Crocin thrice a day you are safe.` Fake

`Wearing mask can protect you from the virus` Real

Figure 2: COVID-19 Fake News Detection.

- Test: Not released yet, but will be released on 1 December
- And, the detailed description of dataset is described on screenshot in Fig. 2.