COMP4432 Machine Learning

End-of-term Take-home Assessment
(Due Friday 11:59pm, 6 May 2022)

Please accept the following declaration

"I hereby declare that I attended this assessment and prepared the answers without any communication with other classmate(s) or person(s)."

by typing your student ID (18085481D) and your name (_ZHANG Caiqi_).

Instructions:
- Answer all questions.
- **Interpret the questions logically, show your steps and write down your assumption(s) when necessary**.
- Please submit your answer to L@PU before the due date.
- When needed, handwritten answers are acceptable.
- Late Submission Policy
    - 3-hour "grace period" is given.
    - After that, no late submission
- Plagiarism Policy
    - Both giver and receiver subject to the same penalty below
    - All the students involved not only will receive 0 marks for this assessment, but also will have an additional 50% penalty applied, e.g., 5 marks for a 10-mark assessment.

Part A: ML Job Interview Questions [Total: 25 marks]

For the questions in this part, you are supposed to answer them in a machine learning job interview setting so that your knowledge and competence are well demonstrated. Your answers should be **concise**, **up to the point**, and **using your own words**; otherwise, marks could be deducted.

1. How do you compare supervised learning with unsupervised learning in terms of (i) type of problems being addressed, (ii) type of data assumed, and (iii) difficulty of data collection?

   In terms of the type of problems being addressed, unsupervised learning is mainly for clustering problems, dimension reduction problems etc. These kinds of questions require the model to find the pattern and feature by itself. Supervised learning is mainly for classification and regression problems. As for the types of data assumed, unsupervised learning does require each data point to have a label, but supervised learning requires a label for each data point. Because the need of labels, it is more difficult of supervised learning methods to collect data. The label may be from human annotation or some semi-automatic ways.

2. Use 1-minute equivalent words to explain the use of and difference between training data, testing data, unseen data and validation data.

   During the training process, training data are fed to the model. The model will then try to learn the pattern or some other behaviors from training data. After that, we can use the validation data to evaluate the model. The validation data are also very useful to adjust the hyperparameters. After several rounds, we can get a model with the best result in validation data and then we need to evaluate it using test set as the real-world application. Unseen data are the data that are not shown to the model during training process. In this sense, all the data that are not used to train the model can be treated as unseen data. For each round, this includes both validation and test data.

3. A SME (small and medium-sized enterprise) is interested in deploying deep learning in its business. Use 1-minute equivalent words to explain to its CEO the necessary requirements for the success of deep learning. You may assume that the CEO has some background in traditional machine learning but he/she is new to deep learning.

   There are several main requirements for the success of deep learning. First, the company should have enough computation power, more specifically GPU. As the deep learning models usually have millions of parameters and several layers, it is necessary to ensure the computation power. Second, deep learning usually needs more data to learn the behaviors or the patterns. The quality and quantity of data are very important for the results of deep learning. Third, as the deep learning models consists of many layers, the flexibility is higher. Professional deep learning engineers are required to design an effective model.

4. What is overfitting? Explain how cross validation can avoid overfitting.

   In short. overfitting means that the model can perform well in the training set but poorly on the unseen data. It also indicated that the model does not have a good generalization ability. Cross-validation divides the dataset into several subset and each subset can be used as validation set. Therefore, it can provide more diverse validation data. When cross-validation is used for model selection, the model with the best generalization ability (i.e., least prone to overfitting) can be selected from multiple models. In this sense, cross-validation is a means to avoid the occurrence of overfitting.

5. An investment company is very interested in applying reinforcement learning to investment decisions. Briefly explain how it works. In particular, give an example of environment, state, action, reward and policy in this application.

Reinforcement learning agent can perceive and interpret its environment, take actions and learn through trial and error. Investment decisions are also a continuous process that requires evolving strategies, getting feedback from the market and trying to optimize trading strategies over time. Therefore, it is very natural to employ reinforcement learning to the investment decisions. To be more specific, the **environment** is the investment market/stock market etc. **States** refers to the observed state of the financial market, such as stock price, volume, some relative value factors and technical indicator factors, etc. The **action** is to buy or sell any financial product or the changes in the weight of each asset class in the portfolio. The **reward** is the profit of the investment. The **policy** is the investment strategy that the agent employs to determine the next action based on the current market state, which maps states to actions.

Part B: Conceptual and Application Questions [Total: 75 marks]

6. [15 marks] This question is about dimensionality reduction (DR).

   a) Given an unlabeled dataset with 100000 dimensions, a ML engineer designed to apply PCA to this dataset and set the variance ratio to a particular percentage, say 80% (cf. slides 21-26 of the dimensionality reduction lecture notes). How many (PCA) dimensions will be resulted? Note that a conceptual answer rather than a specific numeric answer is expected.

      The answer depends on the dataset. If the variances are distributed in only a few dimensions, there can be very few dimensions left (at least 1). However, if the variances are almost uniformly distributed, more than 90% of the dimensions may be kept.

   b) As a ML engineer/developer, do you think that we can chain two different DR techniques (e.g., a linear DR followed by a nonlinear DR)? Justify your answers.

      It is doable. We can first apply a fast linear DR such as PCA to filter the dimensions with very low variance. Then, we can apply a slow but more efficient nonlinear DR such as LLE to continuously reduce the dimensions of the preceding dataset. In this way, it may achieve better performance than only using LLE but with shorter time.

7. [15 marks] The *agglomerative hierarchical* clustering like *Agglomerative Nesting* (AGNES) and *k*-nearest neighbor method are popular models for unsupervised learning and supervised learning respectively. Think about how they can benefit each other in the following settings.

   a) Empowering *k*-nearest neighbor method by agglomerative hierarchical clustering for **supervised learning tasks** (i.e., datasets with label information). Describe your idea and present a conceptually sound model. What are the potential benefits and shortcomings of your new model?

      We can enhance the KNN in supervised learning tasks in the following way: 1) apply agglomerative hierarchical clustering to cluster the training set into several clusters. The number of clusters is a hyperparameter than can be chosen from several options. 2) for each cluster, we can apply KNN. When a new object comes, we can first see which cluster it belongs to, then apply the KNN within that cluster.

      Discussion: The advantage of the new model is that the classification can be done in a more detailed way in different clusters. The accuracy may be improved. The potential disadvantage is that it may increase the computational cost and make the runtime increase. Also, the final accuracy may, to some extent, depend on the clustering result. If the clustering is not good, the result may be even worse.

   b) Empowering agglomerative hierarchical clustering by *k*-nearest neighbor method for **unsupervised learning tasks** (i.e., datasets without label information). Describe your idea and present a conceptually sound model. What are the potential benefits and shortcomings of your new model?

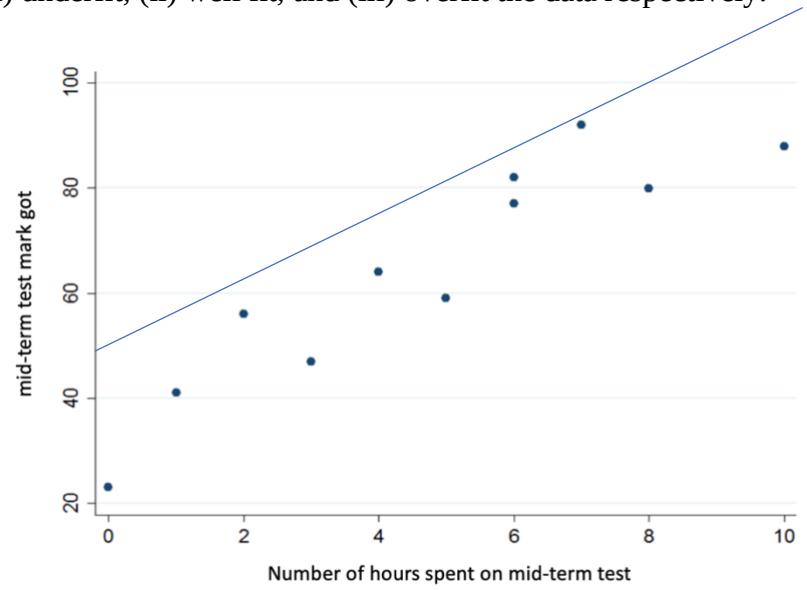
      We can empower the agglomerative hierarchical clustering in unsupervised learning tasks in the following way: 1) We apply the agglomerative hierarchical

clustering only in a subset of the original data. 2) We perform a KNN algorithm to build a classifier to classify all the data. This considers initial clusters to be labels. After that, all the data points with the same pseudo label will be treated as one cluster.

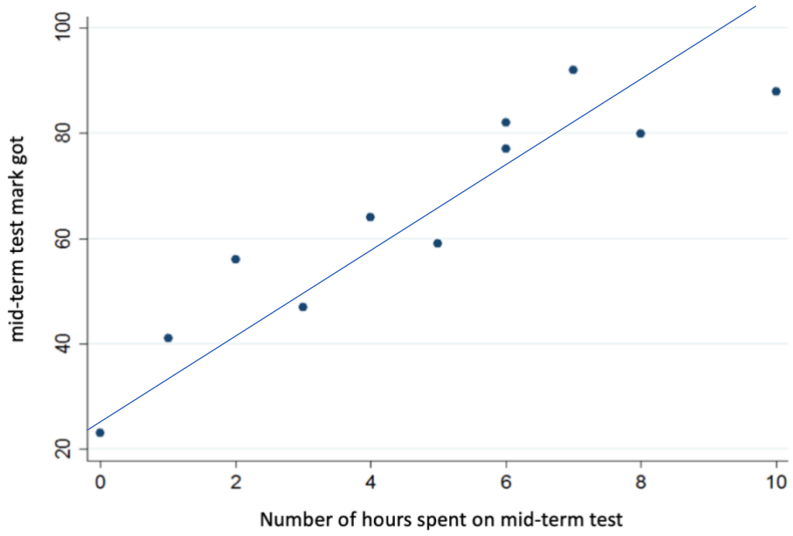
Discussion: The advantages include the following two parts: 1) The accuracy may be improved. [1] proved the efficiency of the algorithm by conducting experiments in several datasets. 2) As the computation cost of the clustering is higher than classification, this method can save more time since we only cluster a subset of the dataset. The possible disadvantage is that it introduces more variables such as the partition rate (i.e., the size of the subset for clustering). The sampling may influence the final result.

Reference:
[1] Mylonas, P., Wallace, M., & Kollias, S. (2004, May). Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering. In *Hellenic Conference on Artificial Intelligence* (pp. 191-200). Springer, Berlin, Heidelberg.

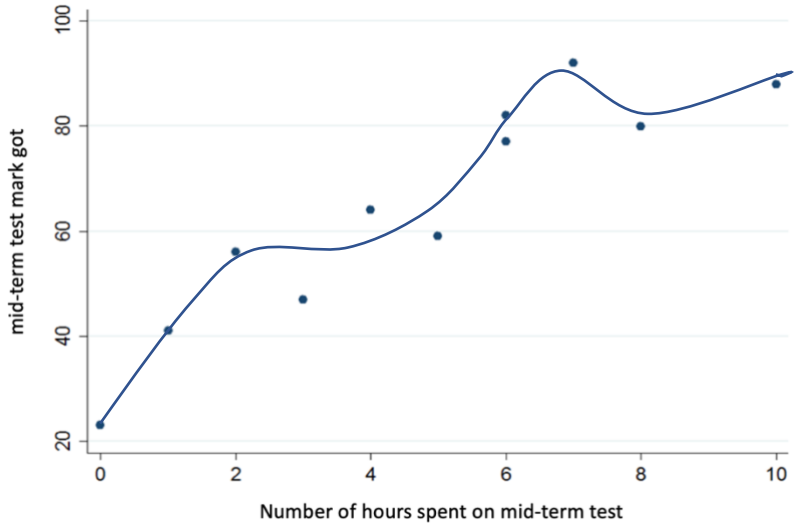
8. [5 marks] For the following regression problem with 11 data points, depicting the number of hours spent on a mid-term test (x-axis) and the mark got (y-axis), show the regression results that (i) underfit, (ii) well fit, and (iii) overfit the data respectively.



Underfit



Well fit

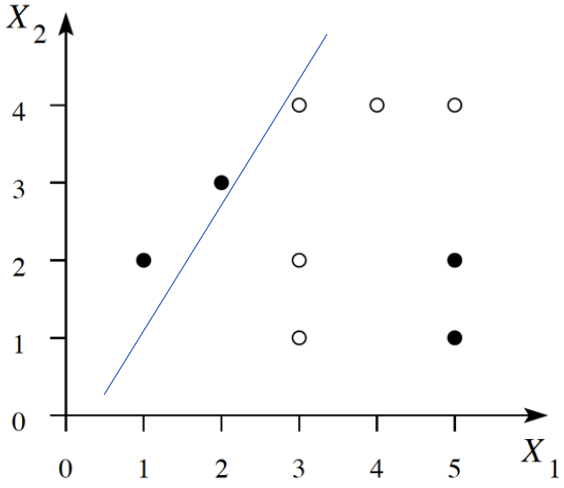


Overfit

9. [15 marks] This question is about neural networks.

a) For the data shown below, show one possible decision plane formed by a perceptron model to obtain the minimum number of misclassifications for the two classes (black dots and white dots). Further show how a multilayer perceptron model with 1 hidden layer can **perfectly** classify all the nine data points into two classes. You are suggested to simply draw the separating planes formed by the hidden neurons and the output neuron to present your answer.

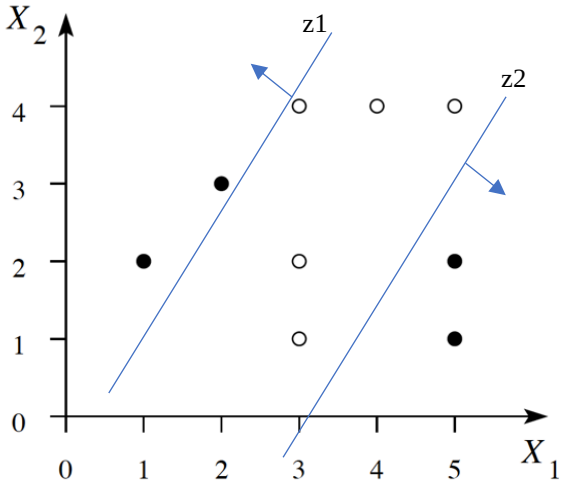
Single layer:



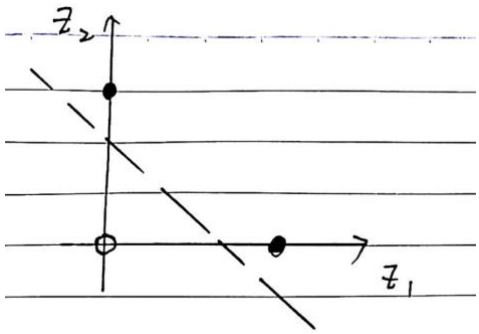
Multilayer:

- The arrow indicates the positive direction (1).

The hidden layer planes are as follows:

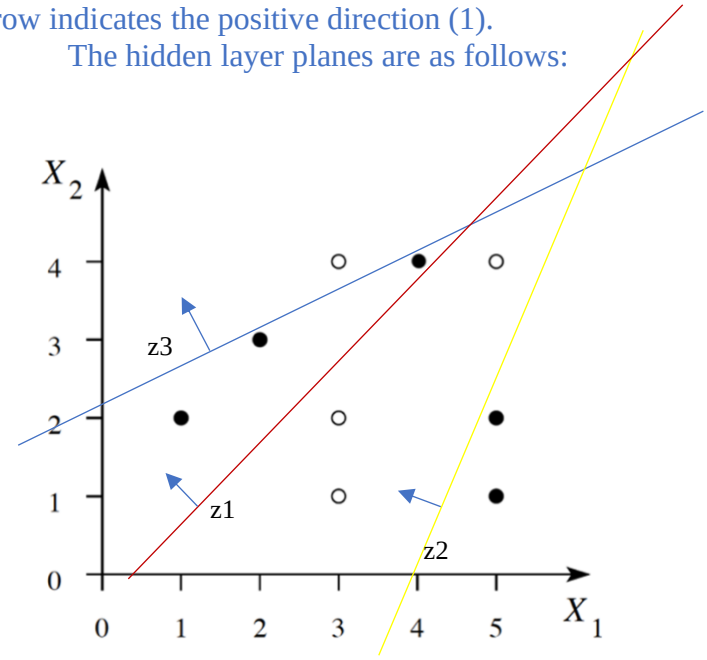


The output neuron plane is as follows:

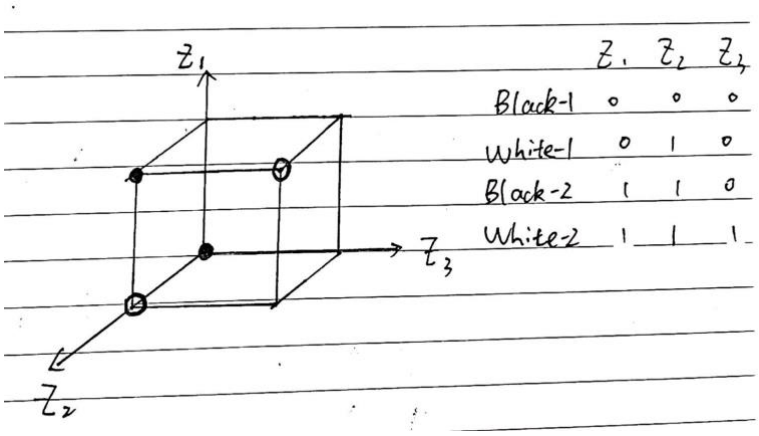


b) One of the data points above is labeled differently as shown below. Show how a multilayer perceptron model with 1 hidden layer can **perfectly** classify all the nine data points into two classes. Again, you are suggested to simply draw the separating planes formed by the hidden neurons and the output neuron to present your answer.

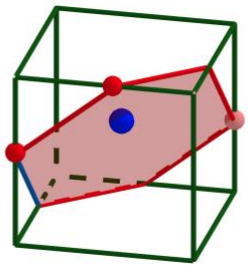
- The arrow indicates the positive direction (1).
  The hidden layer planes are as follows:



The output neuron plane is as follows:



We can easily find a plane to separate the nodes in the following way:



10. [10 marks] Recall that the learning rule based on the following error function for a single layer perceptron using sigmoid output can be described by

$$y^t = sigmoid(w^t x^t)$$

$$E^t(\mathbf{w}|\mathrm{x}^t, \mathrm{r}^t) = -r^t log y^t - (1 - r^t)\log\,(1 - y^t)$$

$$\Delta w_j^t = \eta(r^t - y^t)x_j^t.$$

With respect to the following two added regularization terms (cf. L_1 and L_2 regularization), derive the new learning rules accordingly.

a) $E^t(\mathbf{w}|\mathrm{x}^t, \mathrm{r}^t) = -r^t log y^t - (1 - r^t)\log(1 - y^t) + \lambda \sum_i |w_i|$

According to the derivative sum rule, we do not need to derive from the beginning, we only need to do the derivation to the regularization term.

$$\Delta w_j^t = -\eta \frac{\partial E}{\partial w} = \eta[(r^t - y^t)x_j^t \pm \lambda].$$

$*$ Here w_i may be positive or negative so we add a $\pm$.

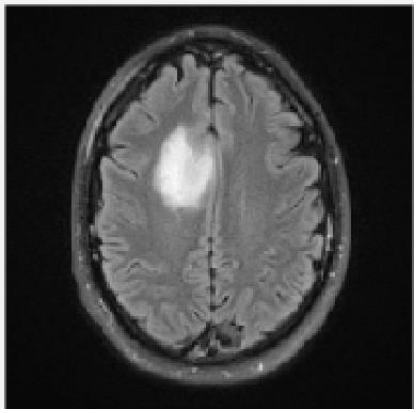
b) $E^t(\mathbf{w}|\mathrm{x}^t, \mathrm{r}^t) = -r^t log y^t - (1 - r^t)\log(1 - y^t) + \lambda \sum_i w_i^2$

According to the derivative sum rule, we do not need to derive from the beginning, we only need to do the derivation to the regularization term.

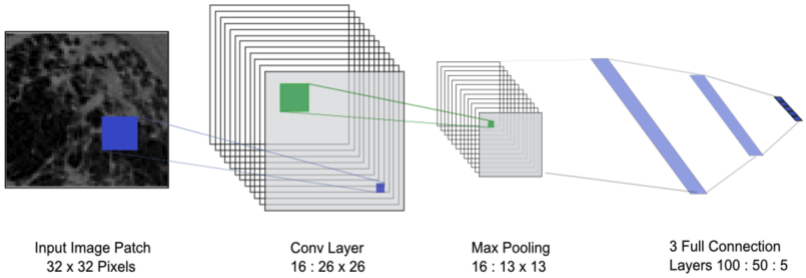
$$\Delta w_j^t = -\eta \frac{\partial E}{\partial w} = \eta[(r^t - y^t)x_j^t - 2\lambda w_j].$$

11. [15 marks] Convolutional neural network (CNN) and autoencoder (AE) have been widely used for different supervised and unsupervised learning tasks. Describe how they can be used in the following applications. Make necessary assumptions when needed.

a) Medical image diagnosis: There exist many medical diagnosis tasks based on the given imaging information like X-ray, MRI, etc. As an example, the task is to diagnosed whether the following brain image has tumor or not. As a ML engineer/developer, you are asked to propose an initial solution based on CNN and/or AE. **Briefly** describe your idea.

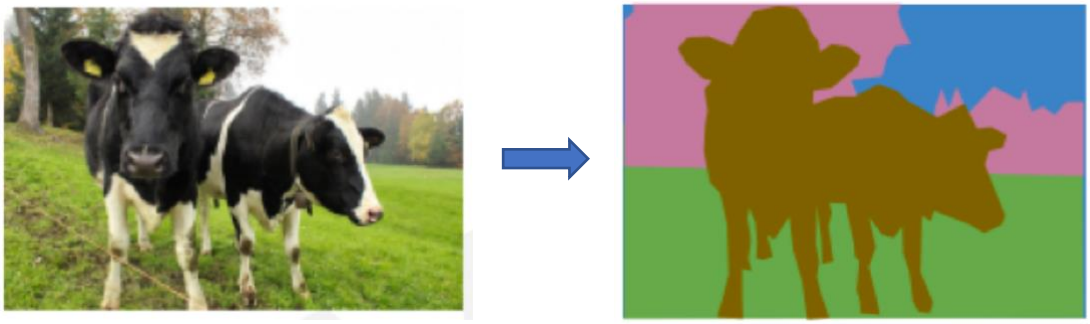


- **Task Type**: Classification. As we are required to diagnose whether the brain image has a tumor or not, it is a supervised classification problem.
- **Model**: CNN. CNN is proved to be effective in image classification problem. It can learn the local and global structures from image data very well.
- **Dataset**: Medical images with labels (whether there is a tumor)
- **Input**: Normalized brain image with unit variance and zero mean.
- **Output**: The possibility of whether there is a tumor (range from 0 to 1)
- **Structure of the model**: First we apply a convolution layer to capture the features in the image. Then a max pooling layer is followed. We can also apply some methods like dropout to utilize the model. To classify the image, we connect 3 linear layers in the end of the model. Finally, the output is the possibility of whether there is a tumor. A very similar structure graph is shown below with the only difference in output layer.

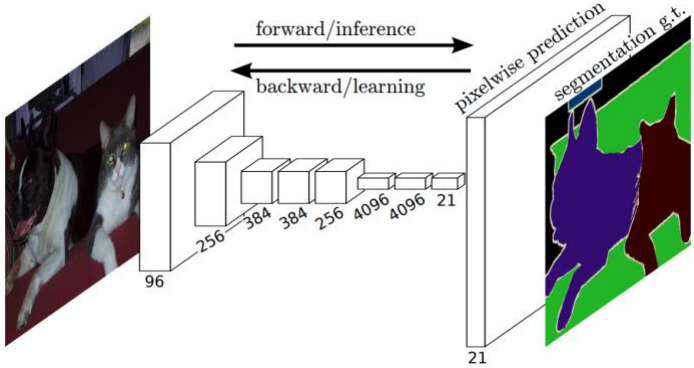


- **Reference**: Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., & Chen, M. (2014, December). Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)* (pp. 844-848). IEEE.

b) Image Segmentation: Image segmentation is very common in computer vision. For the following semantic image segmentation task, i.e. segmenting the given image on the left and producing a segmented image on the right with 4 classes, namely, cow/animal region, sky region, tree region and grass region. You are asked to propose an initial solution based on CNN and/or AE. **Briefly** describe your idea.



- **Task Type**: Classification. The basic idea of segmentation is to classify which object the pixel belongs to. Therefore, we can still use CNN to do the task.
- **Model**: Fully Convolutional Network (FCN). CNN is proved to be effective in image classification problem. Traditional CNN usually connect several linear layers at the end of the model. However, FCN only use convolution layers.
- **Dataset**: PASCAL VOC, NYUDv2, and SIFT Flow
- **Input**: An image of any size
- **Output**: A image (or say a matrix) with a class prediction of every pixel
- **Structure of the model**: Not like other CNN tasks, in this task, we need to predict a class for each pixel in the image. Therefore, we can apply the FCN network, which only consists of convolution layers with different size. An illustration of the model is as follows. By passing several CNN layers, the local and global feature in the image can be captured. Notice that the FCN uses a deconvolution layer to upsample the feature map of the last convolution layer so that it recovers to the same size as the input image, thus allowing a prediction to be generated for each pixel.



- **Reference**: Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

E N D