Example perplexity and class label cleaning

HE Feifan

Abstract

The previous work trained 500 image classifiers with different strengths on the ImageNet dataset. According to the results of each image on these classifiers, we define the perplexity of each image in the training set and validation set of the ImageNet. Perplexity can be regarded as a measure of the quality of sample. My work is to set some thresholds to clean ImageNet and retrain the model, explore the impact of sample quality on the classifier. Try to improve the accuracy of the model with high-quality dataset and to find the misclassified images in ImageNet. We found the relationship between the accuracy of the classifier and the perplexity of the training set, and located a part of the misclassified images in the ImageNet with high accuracy.

1 Introduction

1.1 How is ImageNet collected

First, query on the Internet according to the synonyms of wordnet, and combine the pictures found in the same synonym set into one category. The data set generated in this way is only about 10% accurate, which is a very rough dataset and need to be cleaned. After that, the images are posted on the crowdsourcing website with a brief introduction and ask people if the picture belongs to the corresponding tag. These taggers are not experts. Almost everyone can participate in the tagging of the data set. Although there are some algorithm to ensure the accuracy of data tags during the cleaning process, there are still some wrong or inaccurate tags[1].

1.2 What's wrong with ImageNet

1.As mentioned above, some pictures are mislabeled.

2.There is only one label for each picture, but the content in the picture is very rich. There is not only one subject on the image or the content of the picture is difficult to describe in one word. This kind of picture can not be said to be mislabeled, it can only be said that the label is inappropriate[2].

These problems appear on both train and validation set. Wrong labels on the train set will lead to negative effects of training, and errors on the validation set will cause inaccurate evaluation of the classifiers.

2 Related word

[3] evaluates the quality of a single sample by tracking the influence of a single sample on the change of a certain layer of parameters during the training process. The mislabelled sample is equivalent to an outlier, which will cause a large gradient change(equivalent to Change in the opposite direction).

[2] defined a more scientific and robust pipeline to relabel a part of the ImageNet(ReaL), corrected the wrong labels, changed one label per image to one picture corresponding to multiple labels, captures the diversity

and multiplicity of content in the ImageNet dataset. This relabeled dataset makes up the errors in the original ImageNet and expands the utility of ImageNet in evaluating visual models.

Because the single label per image cannot accurately describe the diversity and multiplicity of the image, some scholars have proposed to use the knowledge distillation[4] to automatically generate soft labels as an effective training supervision to make up for the shortcomings brought by the single label per image. Some studies have found that integrating the predictions of multiple teacher or student models can generate more helpful training supervision and further improve the performance of the model[5][6][7]. In addition to the knowledge distillation series of algorithms, there is another type of method called label refinery. They often use a pre-trained tagger to re-annotate ImageNet. The tagger is generally a deeper network model trained on a larger data set. This not only requires additional models, but also relies on additional larger scales dataset and training resources[8].

3 Data Cleaning Based on Perplexity

We have selected ten popular image networks (VGG16, ResNet50, ResNet101, InceptionV3, Xception, DenseNet121, DenseNet169, DenseNet201, EfficientNetB0, EfficientNetB2). According to the difference in number of images used and training epochs, we generate 500 classifiers with different strengths. There is no strict definition of the strength of classifier. Fewer data used for training, fewer training epoch, weaker the classifier will be. The classifiers can be compared to students. Different model structures correspond to different students' learning methods. The number of images in training sets corresponds to what the students have learned, and the training epochs corresponds to how much students understand what they have learned. The complexity of a single sample is determined by the results of multiple strong and weak classifiers, it is like we need to test an question by trying it on different students to determine how complexity or how hard it is. The perplexity of sample comes from its performance on these 500 classifiers.

3.1 Definition

Let C be a population of N classifiers for classifying examples into M classed.

X-perplexity: For an labeled example (x, y) w.r.t C is defined as $\Phi_X(x) = \frac{1}{N} \sum_{i=1}^{N} 1(C_i \neq y)$, where $C_i(x) = \arg\max_y P_i(y|x)$ is the class assignment function. In words, it is the fraction of the classifiers that misclassifies the example.

C-perplexity: For a given example x, C-perplexity of x w.r.t C is defined as the geometric mean of its perplexity of the probability distribution on each classifier: $\Phi_C(x) = \left[\prod_{i=1}^{N} 2^{H(P_i(y|x))}\right]^{\frac{1}{N}}$. The perplexity of the probability distribution is defined to be $2^{H(P_i(y|x))}$. The entropy $H(P_i(y|x)) = -\sum_{y \in M} P_i(y|x)\log_2 P_i(y|x)$ is a measure of how uncertain the classifier is when classifying x, where $P_i(y|x)$ is the probability distribution over the M classes computed by classifier i.

Range of X-perplexity is $[0, 1]$, range of C-perplexity is $[1, M]$. Smaller the value of perplexity, better the example is. Fig.1 is the perplexity distribution of the ImageNet train set and validation set.

Intuitively speaking, C-perplexity can represent the inherent properties of the image. It has nothing to do with the label of the image. Not rigorous say, it is related to the number of subjects on the image and whether the content on the image is easy to describe. The more subjects on the picture, the more difficult it is to describe, the greater the value of C-perplexity. X-perplexity represents the error rate of the classifiers on

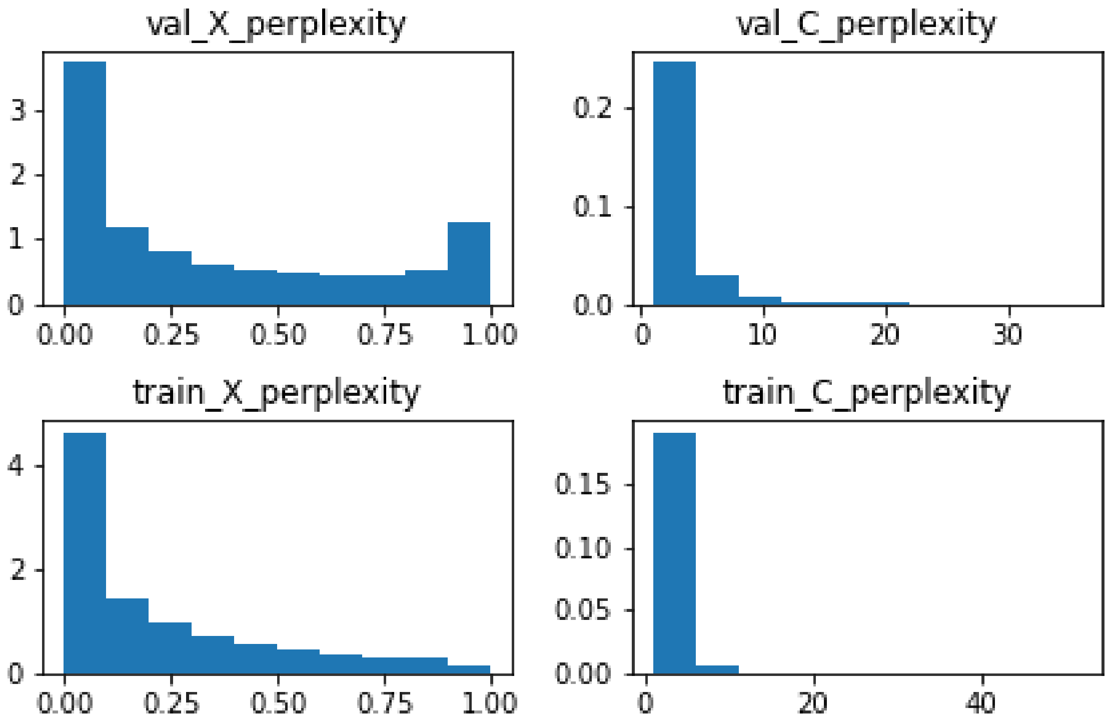


Fig. 1. perplexity distribution

this picture, which is only available for labeled data. Here are some examples:

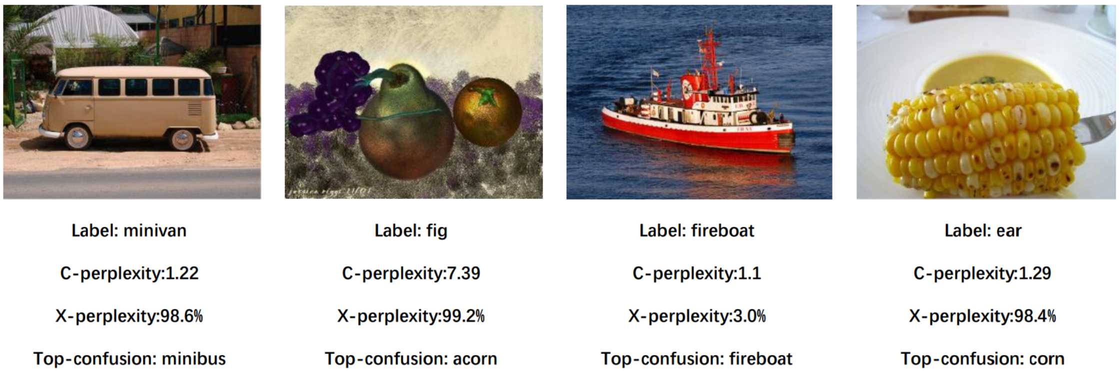


Fig. 2. Examples About Perplexity

3.2 How Dose the Perplexity of Train Set Affect Classifier's Accuracy

Since bad labels on the ImageNet training set will affect the performance of the classifiers, bad labels on the validation set will affect the evaluation of the classifiers. We have obtained the perplexity of each sample label. We can filter out the bad samples based on these perplexity, and use the good examples to train the classifier to improve the performance of the classifier. We selected ResNet101 for the experiment, and selected several perplexity thresholds, as shown in the following table1:

According to the set threshold, we select 500 images that meet the threshold for each category, if not

Model	Baseline	Model1	Model2	Model3
Threshold	nan	Cp<10 Xp<0.7	Cp<4 Xp<0.5	Cp<3 Xp<0.3
The number of images under the threshold	1281167	1173211	1036257	886844
Percentage	100%	91.6%	80.9%	69.2%

Table 1: **Threshold information**

enough, we resample the images that meet the threshold. In this way, we have generated a total of 500,000 filtered clean images. All the classifiers have the same other parameters during the training process. Use the adam optimizer, lr: 0.001, batchsize:128, and train 40 epochs.

The training results are as follows:.

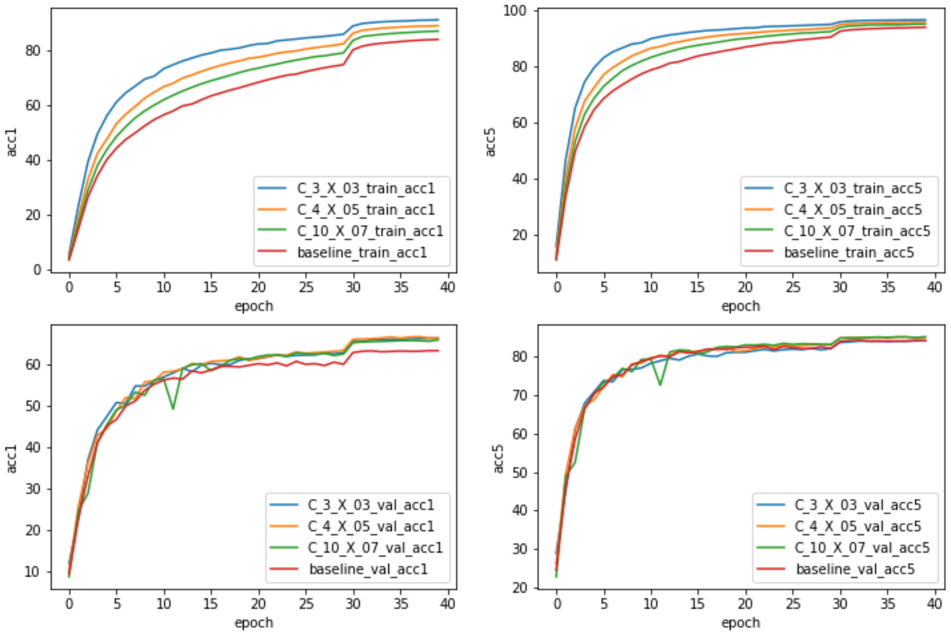


Fig. 3. result analysis

During the training process, cleaner dataset always have higher acc1 and acc5. On the validation set, there is no difference on accuracy between the cleaned datasets(66%), but they are all 3% higher than the baseline.

The picture of loss is as follow: This figure shows that cleaner datasets will converge faster, which is consistent with out definition. Intuitively speaking, cleaner data is easier to learn by the classifier, so there is a smaller loss.

Maybe it is because that those samples that may have a greater impact on the accuracy of the classifier have been cleaned out, that is, those points that have a greater impact on the accuracy of the classifier are all outside the threshold Cp<10 and Xp<0.7.

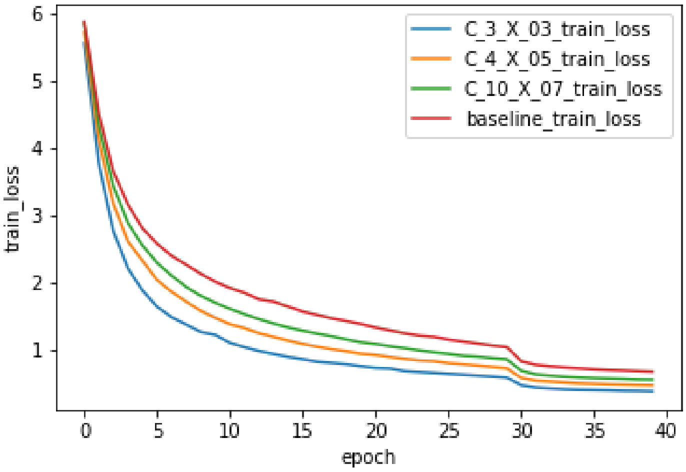


Fig. 4. loss

3.2.1 Results on "Cleaned" Validation Set

As we have mentioned before, incorrect labels on the validation set will affect the evaluation of models. In order to reduce this effect, we compare the differences between the models on different cleaned version validation set. For validation set, we select the same threshold as the training set, get acc1 and acc5 of baseline and the three models as shown in the table below.

Validation set threshold	How many images on validation set selected	Baseline(acc1)	Model1 (acc1)	Model2(acc1)	Model3(acc1)
All	50000	63.40	66.0	66.5	66.38
Cp<10 Xp<0.7	38776	69.3	73.0	74.8	74.8
Cp<4 Xp<0.5	33324	75.5	78.9	80.6	81.0
Cp<3 Xp<0.3	28079	80.3	83.8	85.4	86.4

Validation set threshold	How many images on validation set selected	Baseline(acc5)	Model1 (acc5)	Model2(acc5)	Model3(acc5)
All	50000	84.1	85.2	84.9	84.2
Cp<10 Xp<0.7	38776	88.4	90.0	90.8	90.7
Cp<4 Xp<0.5	33324	91.4	92.8	93.4	93.5
Cp<3 Xp<0.3	28079	93.4	94.4	95.0	95.4

From above experiment and analysis, we can get the following conclusions:

1 Classifiers trained on a cleaned dataset have higher accuracy than the baseline.

2 Although Model3 trained on a cleaner dataset than Model2, their accuracy is almost the same.

3 We filtered out worst 10% of the data on training set in turn(step size 10%), the first step makes the biggest improvement.

3.2.2 Similarity Between Models

In addition to accuracy, I want to compare the similarity between models in more detail, so I defined the following two metrics.

Suppose there a N images in the dataset, y_i is the label of image i in dataset, $P_A(i)$ is the prediction of model A on the image i, $P_B(i)$ is the prediction of model B on the image i.

0/1 similarity:$0/1\_sim(A, B) = \frac{1}{N}\sum_{\text{i=0}}^{N} 1(1(P_A(i) = y_i) = 1(P_B(i) = y_i))$. The 0/1 similarity is the percentage of consistent predictions(right or wrong) given by the two classifiers on the validation set.

prediction similarity:$pred\_sim(A, B) = \frac{1}{N}\sum_{\text{i=0}}^{N} 1(P_A(i) = P_B(i))$. Prediction similarity is the percentage of consistent predictions(label) given by the two classifiers on the validation set.

The heatmap of model similarity is as follow:

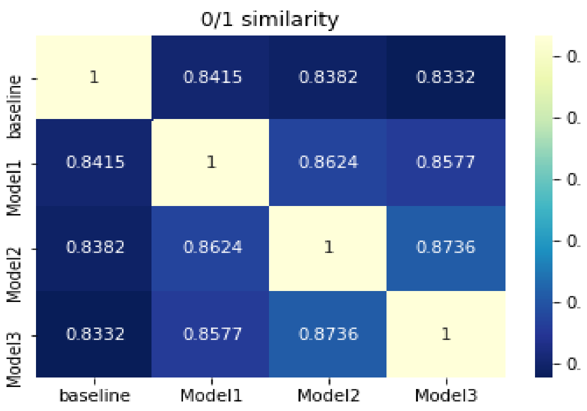


Fig. 5

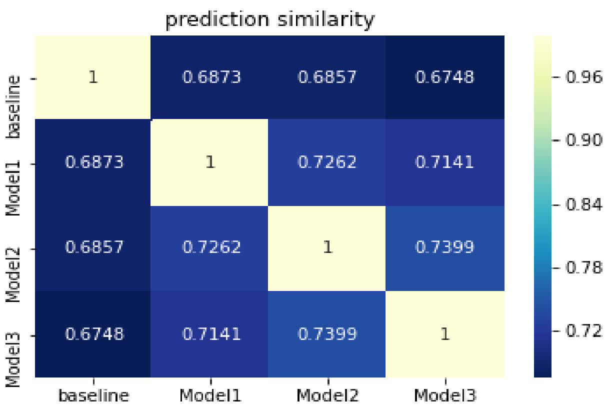


Fig. 6

We can find that the similarity between models trained with cleaned dataset is higher than the similarity with baseline. This is consistent with our previous guess, the model trained with cleaned data is more similar. This show that the perplexity of the images that has a great negative impact on the model is in the range Xp >0.7 or Cp >10.

3.2.3 Similar Experiments with Different Settings

Our pervious experiment on ResNet101 got good results. But this does not verify the universality of this method, so we try it on another classifier and use a smaller dataset to do the similar experiments.

We use ResNet50 as classifier and set different thresholds to filer the training set, also reduce the size of each category to 200. We verify on the original validation set, and get the result as follow:

Train set threshold	Acc1	Acc5
All	50.1	73.2
Cp<10 Xp<0.7	53.87	76.8
Cp<6 Xp<0.95	53.79	76.8
Cp<4 Xp<0.5	54.2	75.6

We can get conclusions that basically similar to the previous experiment. The classifier using cleaned training set has a higher accuracy.

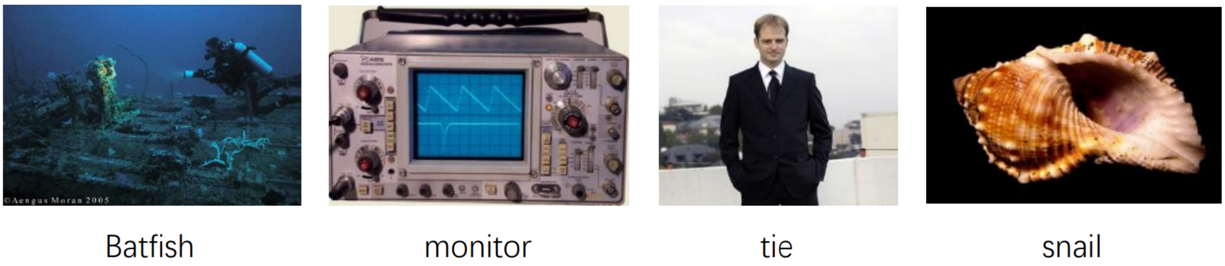
4 Find the Wrong Label and Correct it

In the previous section, when we explored the impact of the cleanliness of the training set on the classifier, we also found that the images with bad label are roughly within the range of Xp >0.7 or Cp >10. Based on this, we have set two sets of threshold. We are pretty confident that a large percentage of images within this threshold are mislabeled.

1 Set A(Model4):Xp>0.95

2 Set B(Model5):Xp>0.95 and Cp<2

Let's continue to use the analogy of students. Set A means that if 95% of the students have done this question wrong, then we think the reference answer of this question is wrong. Set B means that if this is a very easy question and students can give a clear answer, but 95% students have done this question wrong, then we think the reference answer of this question is wrong. We found 4173 images in set A and 1284 images in set B. Here are some examples from Set B with its original label. We can find that a large part of images in set B are mislabeled.



Batfish monitor tie snail

Fig. 7. loss

We change the label of images found in Set A and Set B to its top fusion and retrain the classifier respectively(denoted as Model4 and Model5). We find that the accuracy of Model4 is 2% lower than baseline, accuracy of Model5 is almost equal to the baseline. This show that the images we find is not completely mislabeled, we also introduced a part of wrong labels when we replace the label to top fusion. We need to use a more detailed method to correct the label, or set a smaller range. On the contrary, if we can accurately find misclassified images and correct its label, there will be an improvement of the models.

5 Conclusions and Value of Work

We found that classifiers trained with images in a smaller degree of perplexity have a higher accuracy and converges faster. Therefore, based on perplexity, we can get a cleaned version of ImageNet, making the training process faster and the model more accurate.

A cleaned version of ImageNet might evaluate the classifier more accurate.

We found a rough range of incorrect labels and corrected the labels without manually relabel.

References

[1] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[2] Beyer L, Hnaff O J, Kolesnikov A, et al. Are we done with ImageNet?[J]. arXiv preprint arXiv:2006.07159, 2020.

[3] Pruthi G, Liu F, Kale S, et al. Estimating Training Data Influence by Tracing Gradient Descent[J]. Advances in Neural Information Processing Systems, 2020, 33.

[4] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[5] Tian Y, Krishnan D, Isola P. Contrastive representation distillation[J]. arXiv preprint arXiv:1910.10699, 2019.

[6] Shen Z, He Z, Xue X. Meal: Multi-model ensemble via adversarial learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 4886-4893.

[7] Guo Q, Wang X, Wu Y, et al. Online knowledge distillation via collaborative learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11020-11029.

[8] Yun S, Oh S J, Heo B, et al. Re-labeling ImageNet: from Single to Multi-Labels, from Global to Localized Labels[J]. arXiv preprint arXiv:2101.05022, 2021.

A Minutes

A.1 Minutes of the 1st Project Meeting

Date:2021-1-29

Time:2:30 pm

Place:Zoom

Attending:me and Prof.Zhang

Approval of minutes:

This is first meeting, so there were no minutes to approve.

Discussion Items:

Project background and Q&A

Things to do:

1.Read some paper about this project.

2.Develop a project plan.

Meeting adjournment and next meeting:

The meeting was adjourned at 3:00 PM and the next meeting will be held in March.

A.2 Minutes of the 2nd Project Meeting

Date:2021-3-1

Time:2:00 pm

Place:Zoom

Attending:All the student for IP, Prof. Zhang

Approval of minutes:

This is first formal group meeting, so there were no minutes to approve.

Discussion Items:

1.Clear description of project objective and scope.

2.By student: 10-15 minutes presentation each.

Things to do:

1.Keep doing the experiments and can have an initial result next time.

Meeting adjournment and next meeting:

The meeting was adjourned at 4:30 PM and the next meeting will be held in March.

A.3 Minutes of the 3rd Project Meeting

Date:2021-3-12

Time:4:00 pm

Place:Zoom

Attending:me, LIN Zhi(PHD student) and Prof.Zhang

Approval of minutes:

The minutes of the last meeting were approved without amendment.

Discussion Items:

A clear and detailed experiment plan.

Things to do:

1.Experiment according to the plan.

2.I got some computing resources.

Meeting adjournment and next meeting:

The meeting was adjourned at 4:30 PM and the next meeting will be held in March.

A.4 Minutes of the 4th Project Meeting

Date:2021-3-29

Time:1:00 pm

Place:Zoom

Attending:All the student for IP, Prof. Zhang

Approval of minutes:

The minutes of the last meeting were approved without amendment.

Discussion Items:

1.Progress report, about the results so far and the problems so far.

2.By student: 10-15 minutes presentation each.

Things to do:

1.Keep doing the experiments on a larger dataset and fix the mistakes of the last experiment.

2.Prepare to finish the whole project.

Meeting adjournment and next meeting:

The meeting was adjourned at 5:30 PM and the next meeting will be held in May.

A.5 Minutes of the 5th Project Meeting

Date:2021-5-17

Time:2:00 pm

Place:Zoom

Attending:All the student for IP, Prof. Zhang

Approval of minutes:

The minutes of the last meeting were approved without amendment.

Discussion Items:

1.Final presentation, clearly present the whole project, what I have done, what are the results and the value of work.

By student: 30 minutes presentation each and QA.

Things to do:

1.Write final report and send to prof. Zhang

Meeting adjournment and next meeting:

The meeting was adjourned at 5:30 PM and no further meeting.