**CS506 Programming for Computing**

**HOS09A– Computer Vision – Text Classification**

5/19/2025 Reviewed by Clark Ngo

05/22/2025 Reviewed by Naveena Moddu

05/22/2025 Reviewed by Ashwin Paturi

School of Technology & Computing (STC)

City University of Seattle (CityU)



**Before You Start**

* **Screenshots may be different from your environment.**
* The directory path shown in screenshots may be different from yours.
* There might be subtle discrepancies along with the steps. Please use your best judgment while going through this cookbook-style tutorial to complete each step.
* Some steps may not be explained in detail. If you are not sure what to do:

1. Consult the resources from the course.
2. If you cannot solve the problem after a few tries (usually 15 -30 minutes), ask a TA for help.

**Learning Outcomes**

Students will be able to:

* Use Machine Learning to classify texts.
* Process, build, and train datasets to make and verify predictions.

**Resources**

* Tensorflow.org
* TechWithTim

**Section 1: Preparing your environment -** Get started with your virtual environment here: <https://cityuseattle.github.io/docs/git/github_codepsace/#codespaces>

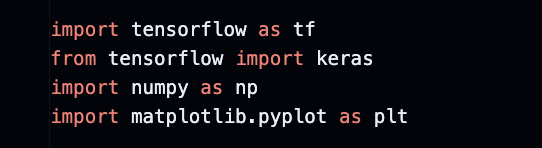
1. Last week, we learned how to

* + 1. Use Machine Learning / deep learning to classify images.
    2. Process, build, and train datasets to make and verify predictions.
    3. Create Convolutional Neural Networks
  1. This week, we will continue to practice working on datasets using embedding layers with text classification. We will work on the IMDB dataset, which contains the text of 50,000 movie reviews.

**Section 2: Text Classification with Keras**

**Import Libraries**

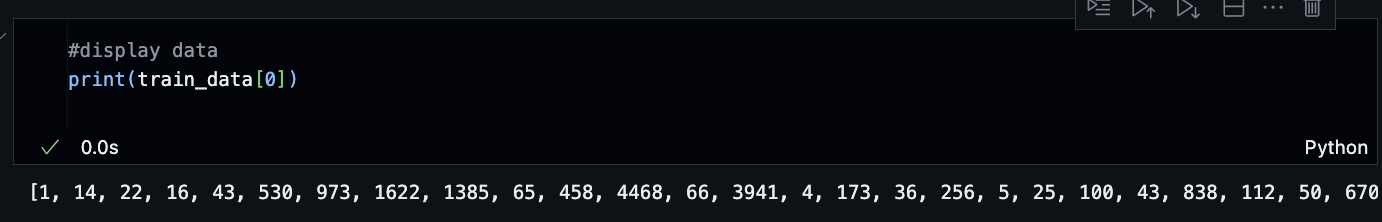
* 1. Create a new file called Text\_Classification.ipynb, and click on the file to open the notebook.
  2. Type the following into the Text\_Classification.ipynb file to install all the libraries needed:



* 1. Type the following to a new block to import the IMDB dataset above:



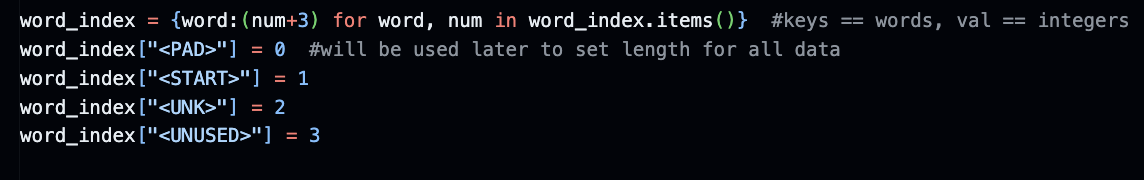
* 1. Let’s print out the dataset to see what we are working with:



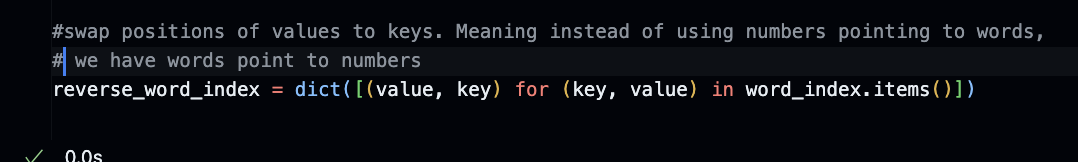
* 1. As you can see, the results are all numbers, which is difficult for us to understand. We are working on movie reviews, so shouldn’t we expect letters instead of numbers? Yes. Each number above points to a word, which is easier for the computer to handle. Let’s transform them into words so that we can read. First, we must create a map between the words and the numbers.
  2. Put the following code into a new block:



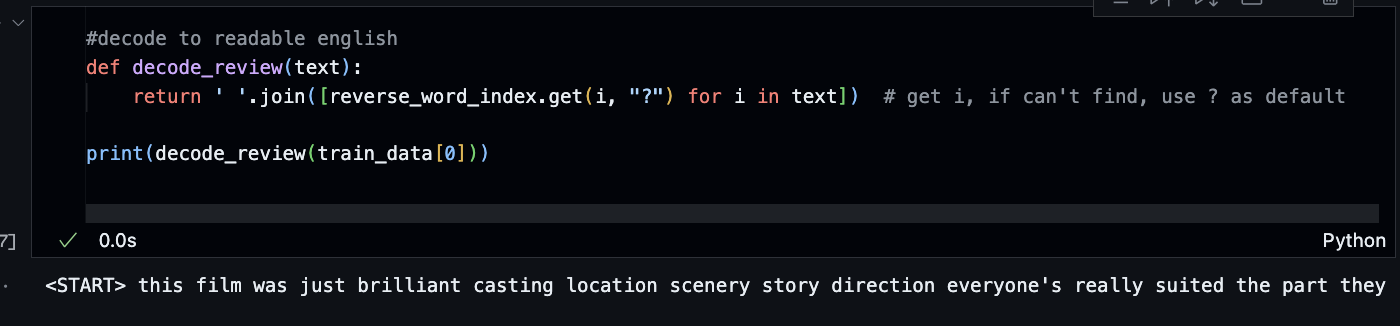
* 1. Put the following code into a new block:



* + 1. <PAD>: will be used for padding. As each review has a different length, we need to set them to a set length to train the model accurately. Let’s say we set the size of each review to be ten words; if one review only contains five words, <PAD> will be filled in until the length is increased to 10.
    2. <START>: will be used to indicate the beginning of the review
    3. <UNK>: to replace the unknown words. Since we only imported 10,000 of the most frequently used words into the datasets, some words will be unknown. If that’s the case, <UNK> will be inserted.
    4. <UNUSED>: to represent the unused words.
  1. Type the following into a new block:



* 1. Now, decode numbers into words and reprint the result. Type the following into a new block:



**Output:**

|  |
| --- |
| <START> this film was just brilliant casting location scenery story direction everyone's really suited the part they played and you could just imagine being there robert <UNK> is an amazing actor and now the same being director <UNK> father came from the same scottish island as myself so i loved the fact there was a real connection with this film the witty remarks throughout the film were great it was just brilliant so much that i bought the film as soon as it was released for <UNK> and would recommend it to everyone to watch and the fly fishing was amazing really cried at the end it was so sad and you know what they say if you cry at a film it must have been good and this definitely was also <UNK> to the two little boy's that played the <UNK> of norman and paul they were just brilliant children are often left out of the <UNK> list i think because the stars that play them all grown up are such a big profile for the whole film but these children are amazing and should be praised for what they have done don't you think the whole story was so lovely because it was true and was someone's life after all that was shared with us all |

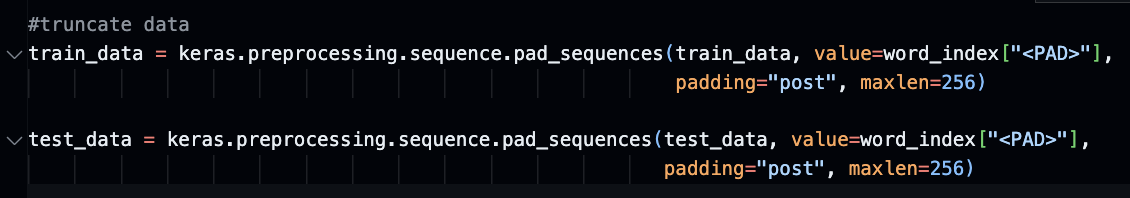
* 1. We can now read the reviews in text. Notice that there are unknown words that were replaced with <UNK>

**Build the model**

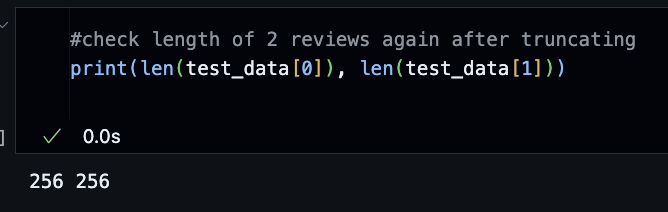
* 1. We must clean up our dataset before building the model. In the previous steps, we created the <PAD> because we want all the reviews to have the same length and will use <PAD> as padding. Before doing so, let’s check out the size of the 2 reviews to see if they differ. Type the following into a new block:



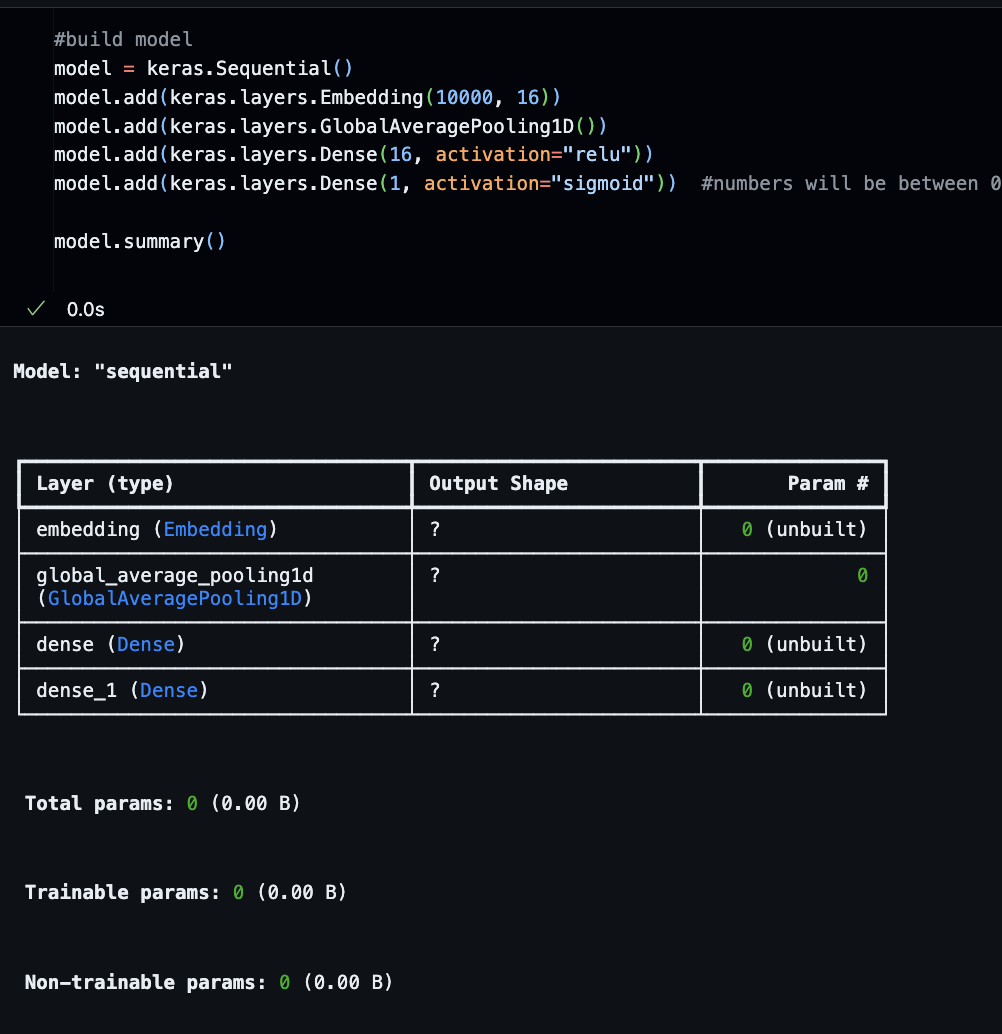
* 1. The review from the train\_data set is 218, while the review from the test\_data set is 260. Let’s set a general cap for the length of the review.
  2. Type the following into a new block:



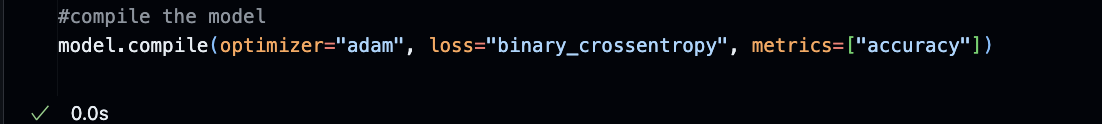
* 1. The code above caps the length of each review to a maximum of 256 words. If the number of words in any review is less than 256, <PAD> will be filled in at the end until the length gets to 256.
  2. Type the following into a new block to make sure the length of our reviews is the same now:



* 1. Right now, both sets have their reviews with the same length of 256. We can now go ahead and build the model
  2. Type the following into a new block to do so:

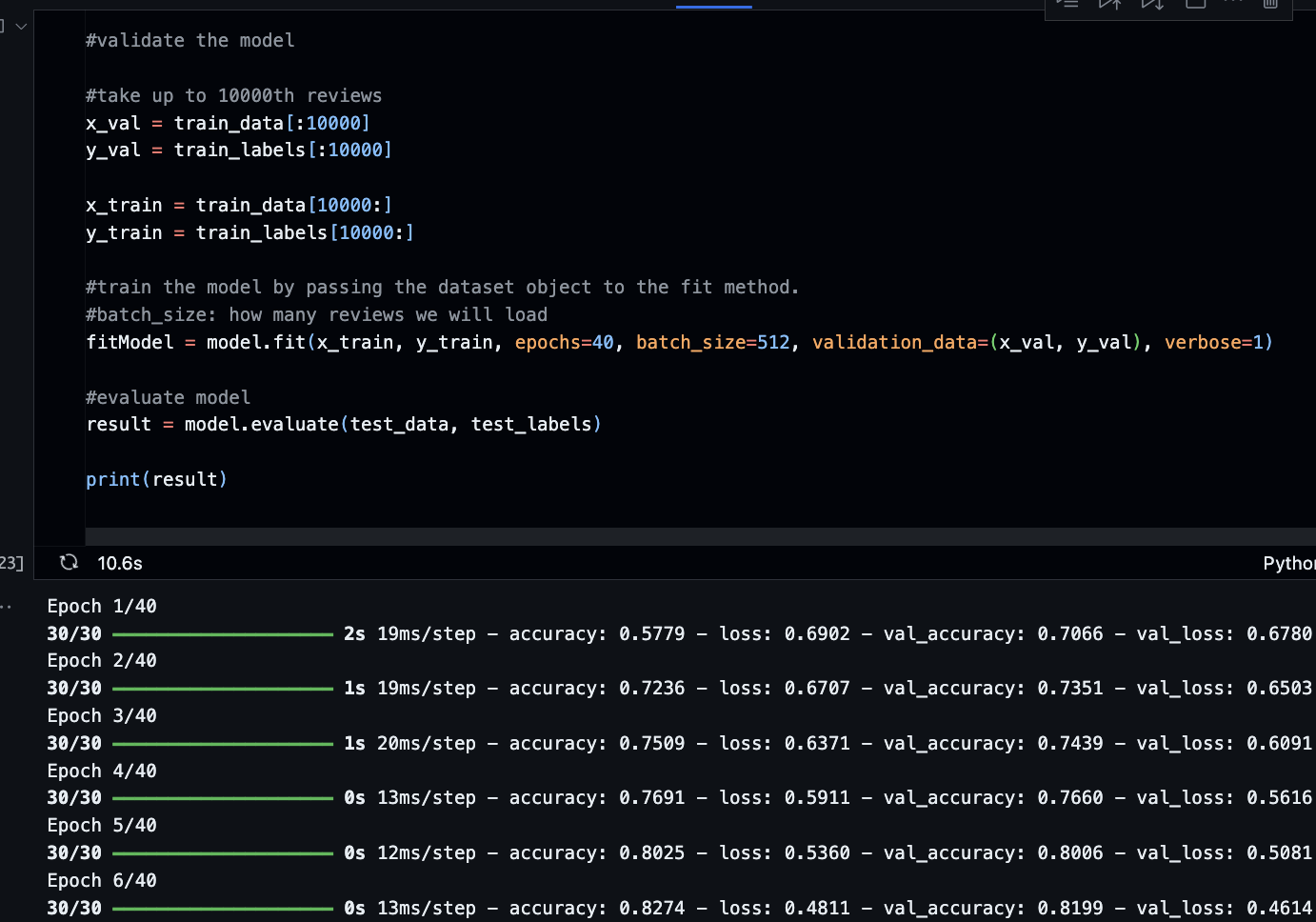


* 1. Type the following into a new block to compile the model:

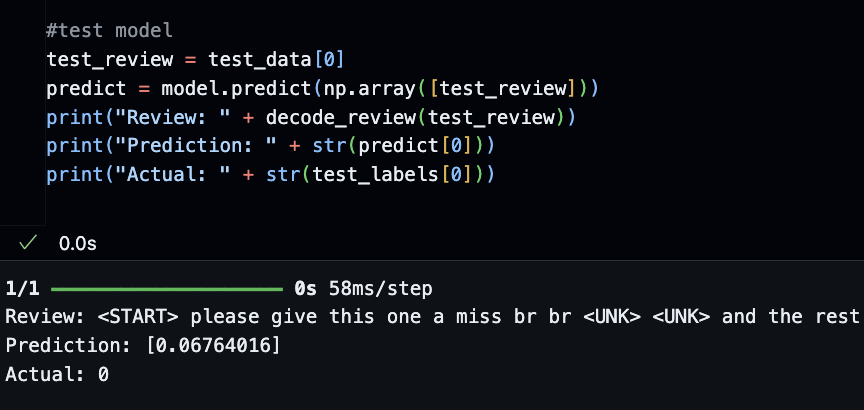


**Evaluate and test the model**

* 1. Type the following into a new block to evaluate our model accuracy and loss:



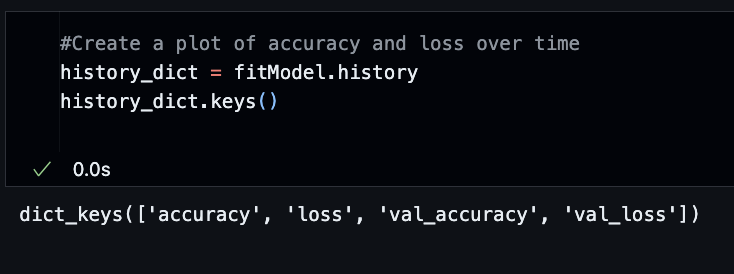
* 1. Based on the result, the accuracy of our model is 87%, while the loss is 36%
  2. Type the following into a new block to test the model:



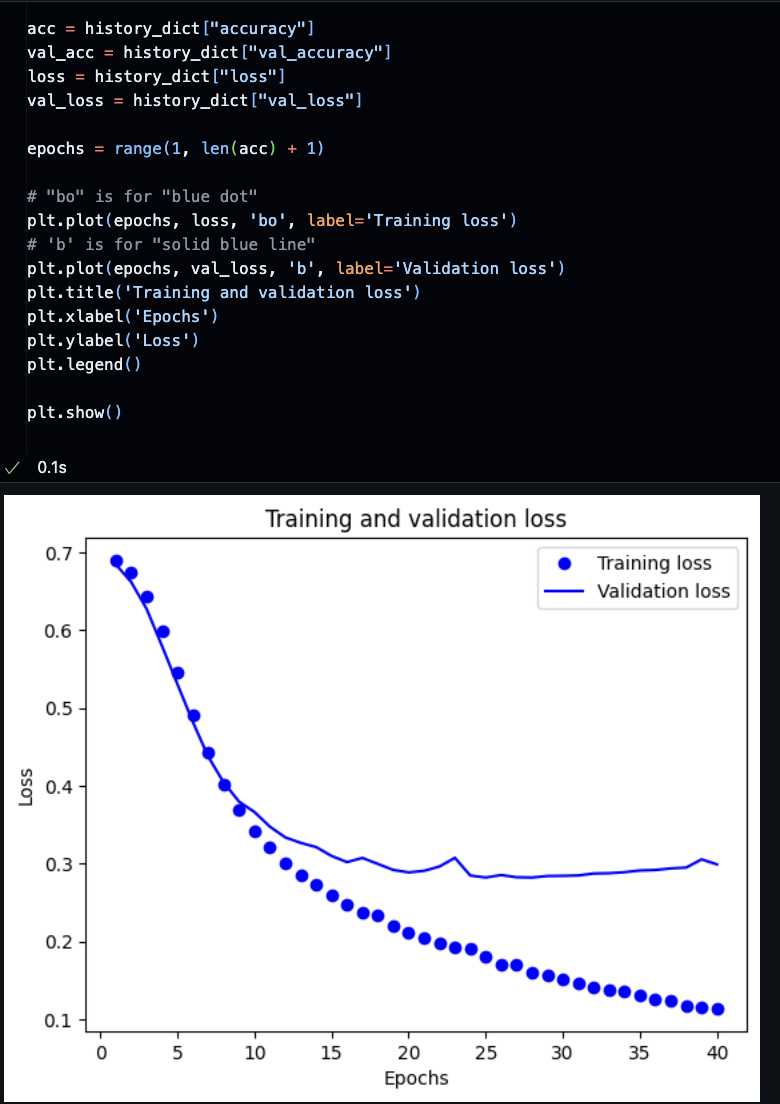
**Output:**

|  |
| --- |
| Review: <START> please give this one a miss br br <UNK> <UNK> and the rest of the cast rendered terrible performances the show is flat flat flat br br i don't know how michael madison could have allowed this one on his plate he almost seemed to know this wasn't going to work out and his performance was quite <UNK> so all you madison fans give this a miss <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> <PAD> |

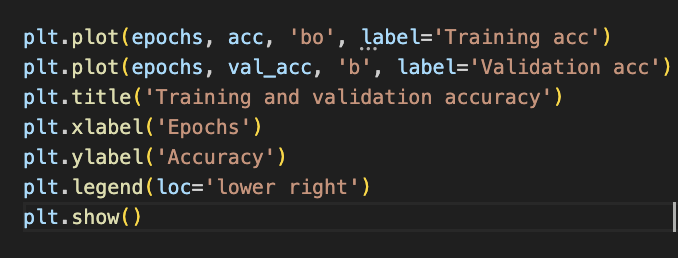
* 1. Now, let’s plot accuracy and loss. Type the following into a new block:

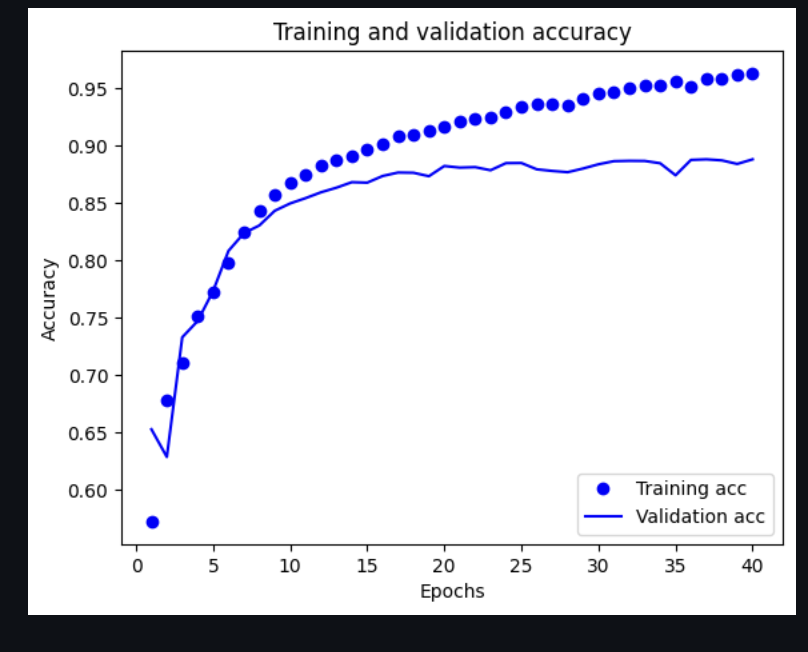


* 1. In a new block, type the following to plot the loss validation:



* 1. Type the following into a new block to plot the accuracy validation:





**Save your Jupyter Notebook with all Output.**

**Section 3: Pushing your work to GitHub**

Follow instructions here: <https://cityuseattle.github.io/docs/git/codespaces_submission/>

1. Go to Source Control on your GitHub Code space and observe the pending changes.
2. Type the message for your changes in the message box at the top. For example,” **Submission for Module09 – Your Name.**”
3. Click on the dropdown beside the commit button and select “**Commit & Push”** to update the changes to your repository's main branch.

Select **Yes** when prompted.