


## Week 2 Discussion


[Dave DeBarr](#)


[All Sections](#)


46

Unread





 **Subscribed**

 **Reply**

○

[Dave DeBarr](#)

Apr 13, 2021

⋮

Setting up CUDA, CUDNN, and TensorFlow libraries at home: will double check that these commands work tomorrow ...

<https://www.cross-entropy.net/ML530/setup.txt>

Note: the version of TensorFlow that you want to use dictates the version of CUDA and CUDNN that you need to install:

<https://www.tensorflow.org/install/gpu>

 **Reply** 

○

[Andrew Feldman](#)

Apr 14, 2021

⋮

Hi Dave, which version of Tensorflow does the DSVM use? I saw that the Python version is 3.7

 **Reply** 

○

[Andrew Feldman](#)

Apr 14, 2021

⋮

I ask because tensorflow-gpu (as I understand it) was rolled into tensorflow for versions 2.x, which makes me think the DSVMs have v1.x. Asking because I set up my own environment with 2.x

 **Reply** 

○

[Dave DeBarr](#)

Apr 14, 2021

⋮

Let's see if replying via email creates a reply in the discussion [fingers crossed] ...

The tensorflow.keras stuff wasn't added until version 2.x.

To check, you can ...

```
import tensorflow
print(tensorflow.__version__)
```

That's two '\_' characters on both sides of "version".

If you can't access it, I'll check tonight.

My goal for "pip install tensorflow-gpu" is to make sure we notice if we don't have the correct common unified device architecture (cuda) libraries available.

Dave

On Wed, Apr 14, 2021 at 9:16 AM Andrew Feldman <[notifications@instructure.com](mailto:notifications@instructure.com)> wrote:

Andrew Feldman posted a new comment on the thread Week 2 Discussion for MLEARN 530 A Sp 21: Deep Learning:

I ask because tensorflow-gpu (as I understand it) was rolled into tensorflow for versions 2.x, which makes me think the DSVMs have v1.x. Asking because I set up my own environment with 2.x



Andrew Feldman

Join the conversation using the link below, or comment by replying to this message. When allowed, if you need to include an attachment, please log in to Canvas and reply to the discussion.



[Click here to join the conversation](#) | [Update your notification settings](#)

← [Reply](#) 👍

○



[Dave DeBarr](#)

Apr 14, 2021

Here are the versions I get for python (3.7.9), tensorflow (2.3.1), and keras (2.4.0) ...

```
(py37_tensorflow) deeplearning@ML-RefVm-871628:~$ python --version
Python 3.7.9
(py37_tensorflow) deeplearning@ML-RefVm-871628:~$ python
Python 3.7.9 (default, Aug 31 2020, 12:42:55)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import tensorflow
2021-04-15 03:01:58.282301: I tensorflow/stream_executor/platform/default/dso_loader.cc:48] Successfully opened dy
>>> tensorflow.__version__
'2.3.1'
>>> tensorflow.keras.__version__
'2.4.0'
>>> exit()

(py37_tensorflow) deeplearning@ML-RefVm-871628:~$
```

The latest version of tensorflow is 2.4.1, but I think 2.3.1 is probably fine for this class.

← [Reply](#) 👍

○



[Andrew Feldman](#)

Apr 14, 2021

Hi Dave, I am seeing that this week's assignment takes a great bit of time - it has been running for the better part of the day for me.

I see this error message, could this be a sign that my GPU is not being utilized?

**2021-04-14 12:42:30.307684: W tensorflow/stream\_executor/platform/default/dso\_loader.cc:55] Could not load dynamic library 'libnvinfer.so.6'; dLError: libnvinfer.so.6: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH: /usr/local/cuda/extras/CUPTI/lib64:/usr/local/cuda/lib64:/usr/local/nvidia/lib:/usr/local/nvidia/lib64**

**2021-04-14 12:42:30.307783: W tensorflow/stream\_executor/platform/default/dso\_loader.cc:55] Could not load dynamic library 'libnvinfer\_plugin.so.6'; dLError: libnvinfer\_plugin.so.6: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH: /usr/local/cuda/extras/CUPTI/lib64:/usr/local/cuda/lib64:/usr/local/nvidia/lib:/usr/local/nvidia/lib64**  
**2021-04-14 12:42:30.307791: W**

tensorflow/compiler/tf2tensorrt/utils/py\_utils.cc:30] Cannot dlopen some TensorRT libraries. If you would like to use Nvidia GPU with TensorRT, please make sure the missing libraries mentioned above are installed properly.

2021-04-14 12:42:31.072271: I tensorflow/stream\_executor/platform/default/dso\_loader.cc:44] Successfully opened dynamic library libcuda.so.1

2021-04-14 12:42:31.099097: I tensorflow/core/common\_runtime/gpu/gpu\_device.cc:1555] Found device 0 with properties: pciBusID: 0000:17:00.0 name: GeForce RTX 2080 SUPER computeCapability: 7.5

coreClock: 1.845GHz coreCount: 48 deviceMemorySize: 7.79GiB deviceMemoryBandwidth: 462.00GiB/s

2021-04-14 12:42:31.099291: W tensorflow/stream\_executor/platform/default/dso\_loader.cc:55] Could not load dynamic library 'libcudart.so.10.1'; dlerror: libcudart.so.10.1: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH: /usr/local/cuda/extras/CUPTI/lib64:/usr/local/cuda/lib64:/usr/local/nvidia/lib:/usr/local/nvidia/lib64

2021-04-14 12:42:31.099363: W tensorflow/stream\_executor/platform/default/dso\_loader.cc:55] Could not load dynamic library 'libcublas.so.10'; dlerror: libcublas.so.10: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH: /usr/local/cuda/extras/CUPTI/lib64:/usr/local/cuda/lib64:/usr/local/nvidia/lib:/usr/local/nvidia/lib64

2021-04-14 12:42:31.099421: W tensorflow/stream\_executor/platform/default/dso\_loader.cc:55] Could not load dynamic library 'libcufft.so.10'; dlerror: libcufft.so.10: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH: /usr/local/cuda/extras/CUPTI/lib64:/usr/local/cuda/lib64:/usr/local/nvidia/lib:/usr/local/nvidia/lib64

2021-04-14 12:42:31.099480: W tensorflow/stream\_executor/platform/default/dso\_loader.cc:55] Could not load dynamic library 'libcurand.so.10'; dlerror: libcurand.so.10: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH: /usr/local/cuda/extras/CUPTI/lib64:/usr/local/cuda/lib64:/usr/local/nvidia/lib:/usr/local/nvidia/lib64

2021-04-14 12:42:31.099539: W tensorflow/stream\_executor/platform/default/dso\_loader.cc:55] Could not load dynamic library 'libcusolver.so.10'; dlerror: libcusolver.so.10: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH: /usr/local/cuda/extras/CUPTI/lib64:/usr/local/cuda/lib64:/usr/local/nvidia/lib:/usr/local/nvidia/lib64

2021-04-14 12:42:31.099606: W tensorflow/stream\_executor/platform/default/dso\_loader.cc:55] Could not load dynamic library 'libcusparsesolver.so.10'; dlerror: libcusparsesolver.so.10: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH: /usr/local/cuda/extras/CUPTI/lib64:/usr/local/cuda/lib64:/usr/local/nvidia/lib:/usr/local/nvidia/lib64

2021-04-14 12:42:31.102281: I tensorflow/stream\_executor/platform/default/dso\_loader.cc:44] Successfully opened dynamic library libcudnn.so.7

2021-04-14 12:42:31.102299: W tensorflow/core/common\_runtime/gpu/gpu\_device.cc:1592] Cannot dlopen some GPU libraries. Please make sure the missing libraries mentioned above are installed properly if you would like to use GPU. Follow the guide at <https://www.tensorflow.org/install/gpu> for how to download and setup the required libraries for your platform.

Skipping registering GPU devices...

2021-04-14 12:42:31.102481: I tensorflow/core/platform/cpu\_feature\_guard.cc:142] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX2 AVX512F FMA

2021-04-14 12:42:31.127582: I tensorflow/core/platform/profile\_utils/cpu\_utils.cc:94] CPU Frequency: 3799900000 Hz

2021-04-14 12:42:31.128761: I tensorflow/compiler/xla/service/service.cc:168] XLA service 0x5ae1ad0 initialized for platform Host (this does not guarantee that XLA will be used). Devices:

2021-04-14 12:42:31.128794: I tensorflow/compiler/xla/service/service.cc:176] StreamExecutor device (0): Host, Default Version

2021-04-14 12:42:31.366996: I tensorflow/compiler/xla/service/service.cc:168] XLA service 0x5acee10 initialized for platform CUDA (this does not guarantee that XLA will be used). Devices:

2021-04-14 12:42:31.367050: I tensorflow/compiler/xla/service/service.cc:176] StreamExecutor device (0): GeForce RTX 2080 SUPER, Compute Capability 7.5

2021-04-14 12:42:31.367239: I tensorflow/core/common\_runtime/gpu/gpu\_device.cc:1096] Device interconnect StreamExecutor with strength 1 edge matrix:

2021-04-14 12:42:31.367254: I tensorflow/core/common\_runtime/gpu/gpu\_device.cc:1102]

[Search space summary]

|-Default search space size: 11

> dense\_depth (Int)

|-default: None

I ran nvidia-smi to check GPU utilization, and I see

Wed Apr 14 12:01:06 2021

```
+-----+
| NVIDIA-SMI 435.21      Driver Version: 435.21      CUDA Version: 10.1      |
|-----+-----+-----+
| GPU Name      Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|  Memory-Usage | GPU-Util  Compute M. |
|=====+=====+=====+
|  0  GeForce RTX 208... Off | 00000000:17:00.0 Off |             N/A             |
```

```
| 0% 39C P8 3W / 250W | 146MiB / 7979MiB | 0% Default |
+-----+
+-----+
| Processes:                      GPU Memory |
| GPU  PID  Type  Process name      Usage   |
+-----+
| 0    976   G   /usr/lib/xorg/Xorg        14MiB |
| 0    1037  G   /usr/bin/gnome-shell      4MiB |
| 0    6117  C   python              115MiB |
+-----+
```

I have bolded the python process above, it suggests that my GPU is being utilized. Is 115MiB a good indication that the GPU is being used? I was hoping to see something more like a "% utilization" analogous to CPU utilization but I can't find that.

I am one of the people who let their DSVM hours expire, so I am running on my own machine with NVIDIA RTX 2080 Ti (as you can see above), Python 3.7, Tensorflow 2.x, CUDA 10.x (I set up my environment before you published the guidance :))

Any help on this is appreciated, thanks :)

[← Reply](#)



[Dave DeBarr](#)

Apr 14, 2021

I'm thinking it should complete in under 30 minutes on the VM. Will ask Rudy about adding more time to your VM. Please remember to shut it off when not in use.

It's fine to see 0% gpu-util (utilization) every once in a while, but it should be fairly decent non-zero usage.

Are you running the gnome desktop on your machine, for your user interface?

Tensorflow typically allocates all available memory by default, so 115 megabytes is definitely less than the 8 gigabytes of memory for your 2080 card.

How much memory does your machine have? Is there any chance you've run out of memory, and it's swapping pages from memory to disk?

What does your output look like for "free -m"?

Which version of tensorflow are you using?

```
import tensorflow
print(tensorflow.__version__)
```

[← Reply](#)



[Dave DeBarr](#)

Apr 14, 2021

Mea culpa (my bad): I ignored your error message. Yeah, you're not configured correctly.

**Could not load dynamic library 'libnvinfer.so.6'; dlerror: libnvinfer.so.6: cannot open shared object file: No such file or directory; LD\_LIBRARY\_PATH:**

[← Reply](#)



[Dave DeBarr](#)

Apr 14, 2021

How did you install the cuda toolkit?

[← Reply](#)

 [Dave DeBarr](#)  
Apr 14, 2021


Okay, I had to update the PATH and LD\_LIBRARY\_PATH statements for the setup script (changed "cuda-11" to "cuda"):

<https://www.cross-entropy.net/ML530/setup.txt>


I verified that it works:

<https://www.cross-entropy.net/ML530/setup-console.txt>

← [Reply](#) 

 [Andrew Feldman](#)  
Apr 15, 2021

Thanks for the feedback. Let me try actually following your setup instructions, since it seems things were not quite right in my setup, and the Hyper-banding did take the better part of a day :)

← [Reply](#) 

 [Andrew Feldman](#)  
Apr 14, 2021

As we are starting to investigate a variety of DNN activation functions, I led me to recall malicious adversarial examples as a topic of DNN research <https://towardsdatascience.com/breaking-neural-networks-with-adversarial-attacks-f4290a9a45aa>. Back when I was considering whether to do a masters, I had reached out to a lab at my Alma Mater that was investigating the robustness of DNNs to malicious adversarial examples. Therefore I read a number of papers such as these

Robustness to adversarial examples for ReLU networks

<https://arxiv.org/pdf/1804.09699>

Robustness to adversarial examples for general activation functions (this came later, of course)

<https://papers.nips.cc/paper/2018/file/d04863f100d59b3eb688a11f95b0ae60-Paper.pdf>

It is an interesting concept, that malicious adversarial examples will typically succeed by choosing the smallest-possible image modification that leads the DNN to misclassify. Because small modifications will be noticeable to the DNN but not to the human eye. And so this lab developed a technique that can certify a DNN as being invulnerable to adversarial examples if the sum-of-squares magnitude of the image modification is less than a certain amount.

← [Reply](#) 

 [Dave DeBarr](#)  
Apr 14, 2021

For what it's worth: during the second to last class, we'll look at a couple of examples of generative adversarial networks: where we alternatively train a generative model to generate observations and a discriminative model to distinguish "real" observations from the "generated" observations (one batch for the discriminator; one batch for the generator; another batch for the discriminator; another batch for the generator; lather; rinse; repeat). Both will use crossentropy as the loss function, though the generator is only trained on the generated observations. We'll have lots more to say about this later.

← [Reply](#) 

 [Eric Kuo](#)  
Apr 15, 2021

Hi Dave,

I keep getting a connection timeout error when I try to download the sample code. Any ideas what might be causing the problem?

```
(py37_tensorflow) deeplearning@ML-RefVm-871628:~$ wget https://www.cross-entropy.net/ML530/fashion-search.py.txt
--2021-04-15 15:55:15-- https://www.cross-entropy.net/ML530/fashion-search.py.txt
Resolving www.cross-entropy.net... 107.180.57.14
Connecting to www.cross-entropy.net[107.180.57.14]:443... failed: Connection timed out.
Retrying.

--2021-04-15 15:57:27-- (try: 2) https://www.cross-entropy.net/ML530/fashion-search.py.txt
Connecting to www.cross-entropy.net[107.180.57.14]:443... failed: Connection timed out.
Retrying.

--2021-04-15 15:59:39-- (try: 3) https://www.cross-entropy.net/ML530/fashion-search.py.txt
Connecting to www.cross-entropy.net[107.180.57.14]:443... failed: Connection timed out.
Retrying.

--2021-04-15 16:01:53-- (try: 4) https://www.cross-entropy.net/ML530/fashion-search.py.txt
Connecting to www.cross-entropy.net[107.180.57.14]:443... ^C
```

Eric

Edited by [Eric Kuo](#) on Apr 15 at 4:23pm[← Reply](#) [Dave DeBarr](#)

Apr 15, 2021

It looks like shared hosting across the continent is having trouble :( [I believe our virtual machines are on the East coast; CoViD has made GPU machines a somewhat scarce resource]

I would try adding the "--timeout=2" option, so it fails quickly:

```
wget --timeout=2 https://www.cross-entropy.net/ML530/fashion-search.py.txt
```

Alternatively, you could download from [https://canvas.uw.edu/courses/1432990/files/76180372/download?download\\_frd=1](https://canvas.uw.edu/courses/1432990/files/76180372/download?download_frd=1) (or <https://www.cross-entropy.net/ML530/fashion-search.py.txt>); then upload to your VM using ...

```
"C:\Program Files\PuTTY\pscp.exe" -P PortNumberGoesHere fashion-search.py.txt
deeplearning@HostNameGoesHere:/home/deeplearning/fashion-search.py.txt
```

[← Reply](#) (1 like)[Aeden Jameson](#)

Apr 15, 2021

Hi Dave,

Would you expand on your comments on how to properly compare models at least to the extent of where you might expect us beginners to be at? During class someone had commented "ROC", would also expand on your remarks there as well?

Edited by [Aeden Jameson](#) on Apr 15 at 1:21pm[← Reply](#) [Dave DeBarr](#)

Apr 15, 2021

Part 1 ...

On pages 222-223 (based on printed page numbers; numbers found at the top of the pages) of the 2nd edition of The Elements of Statistical Learning (<https://web.stanford.edu/~hastie/ElemStatLearn/download.html>), they write ...

<begin selective quotes>

Model selection: estimating the performance of different models in order to choose the best one.

Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data.

If we are in a data-rich situation, the best approach for both problems is to randomly divide the dataset into three parts: a training set, a validation set, and a test set. The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model. Ideally, the test set should be kept in a "vault," and be brought out only at the end of the data analysis.

The methods in this chapter are designed for situations where there is insufficient data to split it into three parts. Again it is too difficult to give a general rule on how much training data is enough; among other things, this depends on the signal-to-noise ratio of the underlying function, and the complexity of the models being fit to the data.

The methods of this chapter approximate the validation step either analytically (AIC, BIC, MDL, SRM) or by efficient sample re-use (cross-validation and the bootstrap). Besides their use in model selection, we also examine to what extent each method provides a reliable estimate of test error of the final chosen model.

<end selective quotes>

I would strongly encourage you to avoid using a selection measure on the training data, even a penalized version of the measure [such as the Akaike Information Criterion (AIC; equation 7.27), the Bayesian Information Criterion (BIC; equation 7.35); etc].

If you have the resources and you want to repeatedly split the training data into training and validation sets, in order to perform cross validation for model selection, then that's awesome! Suppose we have 11,000 data points. We could split the data into 9,000 for training, 1,000 for validation, and 1,000 for testing; but it's also possible for us to combine the training and validation together then run 10-fold cross validation [essentially repeating the training/validation process 10 times instead of once].

While it's possible you will run into a tiny data set where you may need to consider using one of the methods discussed in that chapter, I expect this to be a very uncommon circumstance.

← [Reply](#) 👍



[Dave DeBarr](#)

Apr 15, 2021



Part 2 ...

Uh, can you remind what was said about the Receiver Operating Characteristic (ROC) curve? :)

I think it's fine to compare learning algorithms by comparing the Area Under the ROC (AUC) for some validation data; but keep in mind that, in practice, we often have to choose a classification threshold for the model. For example, let's suppose we can only mail out 10,000 offers. Then we want to maximize the response rate given the constraint that we can only mail out 10,000 offers.

P.S. It's possible for AUC to be used for a Precision-Recall (PR) curve as well.

← [Reply](#) 👍



[Abhijit Majumdar](#)

Apr 15, 2021



Hello Dave,

For Machine Learning, I use t-tests to understand whether the performance of one model is statistically significant vis-à-vis the other. I run my cross validation through a certain number of epochs and then collect my metric (which I am using to assess the performance of my model via cross validation) before executing my t-test. Although I know this is not a part of our current syllabus, but would you be able to let me know if my approach is correct :-)? The code, which I use for my t-test, is as follows:

```
import scipy.stats as ss
```

```
import statsmodels.stats.weightstats as ws
```

```
def run_t_test(a, b, alpha=0.05, alternative='two-sided'):
```

```
    """
```

```
    This function executes a student's t-test on the given set of observations.
```

```
    -----
```

```
    :params: a - The first given set of observations.
```

```
    :params: b - The second given set of observations.
```

```
    :params: alpha - The significance level to be used. This defaults to 5% (indicating a 95% Confidence Interval).
```

```
    :params: alternative - Indicates whether the test is one-sided or two-sided. The default is two-sided.
```

-----  
Returns: a Pandas Series with the calculated t-statistics.

'''

```
diff = a.mean() - b.mean()
res = ss.ttest_ind(a, b, equal_var=False)
means = ws.CompareMeans(ws.DescrStatsW(a), ws.DescrStatsW(b))
confint = means.tconfint_diff(alpha=alpha, alternative=alternative, usevar='unequal')
degfree = means.dof_satt()
index = ['DegFreedom', 'DifferenceInMeans', 'T-Statistic', 'PValue', 'Low95CI', 'High95CI']
return pd.Series([degfree, diff, res[0], res[1], confint[0], confint[1]], index = index)
```

Thank you so much !

[← Reply](#)



[Dave DeBarr](#)

Apr 15, 2021

Why are you calling multiple functions?

Are you using weighted values for the vectors?

If I wanted a simple test of the means for two independent samples, I would vote for Welch's t-test. Here's a simple-ish implementation ...

```
import numpy as np
from scipy.stats import t
def welchs_t_test(vector1, vector2, confidence_level = 0.95):
    percentiles = [ (1.0 - confidence_level) / 2.0, 1.0 - (1.0 - confidence_level) / 2.0 ]
    means = [ np.mean(vector1), np.mean(vector2) ]
    difference_of_means = means[0] - means[1]
    vars = [ np.var(vector1, ddof = 1), np.var(vector2, ddof = 1) ]
    lens = [ len(vector1), len(vector2) ]
    var = vars[0] / lens[0] + vars[1] / lens[1]
    stderr = np.sqrt(var)
    test_statistic = difference_of_means / stderr
    df = (var**2) / (((vars[0] / lens[0])**2) / (lens[0] - 1) + ((vars[1] / lens[1])**2) / (lens[1] - 1))
    bounds = difference_of_means + [ stderr * t.ppf(percentiles[0], df = df), stderr * t.ppf(percentiles[1], df = df) ]
    pvalue = 2 * t.cdf(- np.abs(test_statistic), df = df)
    return [ pvalue, test_statistic, difference_of_means, bounds, df ]
```

```
x1 = [0]*105 + [1]*9895
```

```
x0 = [0]*169 + [1]*9831
```

```
welchs_t_test(x1, x0)
```

[← Reply](#)



[Dave DeBarr](#)

Apr 15, 2021

Reminder: if we can use the same test set to evaluate both models, I would vote for dependent sample testing instead: so we get a better estimate of uncertainty about the difference in means [this can either increase or decrease the variance, depending on the behavior of the models]

[← Reply](#)




[Abhijit Majumdar](#)

Apr 17, 2021

Thank you so much Dave, this helped :-) !

[← Reply](#)




- 

[Benjamin Johnson](#)  
Apr 16, 2021

⋮

I don't have access to the second Kaggle competition. I included my Kaggle username (Valetta) in the homework, as well as in reply to Chenwei's request through the announcements.

↩ [Reply](#) 👍
- 


[Dave DeBarr](#)  
Apr 16, 2021

⋮

Did you navigate to accept the invitation, using the link from the assignment?

Please navigate to the following URL to accept the invitation for this Kaggle task:


<https://www.kaggle.com/t/90efb52a00be40849b3ccde825e6a4c9>

↩ [Reply](#) 👍
- 

[Benjamin Johnson](#)  
Apr 19, 2021

⋮

Thank you; it works now; that must have been what I'd missed.

↩ [Reply](#) 👍
- 

[Carl Gaither](#)  
Apr 17, 2021


⋮

Hi Dave,

My VM is used up - I didn't shutdown after the first assignment, but I know to do so now. Can I get more time on my VM?

Thanks,

Carl

↩ [Reply](#) 👍
- 


[Dave DeBarr](#)  
Apr 17, 2021

⋮

Hi Carl,

I sent email to Rudy. Hopefully, I'll hear from him soon.


Dave

↩ [Reply](#) 👍
- 

[Carl Gaither](#)  
Apr 18, 2021

⋮

Thanks!

↩ [Reply](#) 👍
- 

[Dave DeBarr](#)  
Apr 19, 2021

⋮

Our system administrator (Rudy) said he would go through the list of student VMs and add hours if the hours have been depleted.

Please remember to turn off your VM after you're done using it; but please use it.

If you're out of hours, please contact Rudy by sending an email to "sadm\_rudy514@uw.edu".

← [Reply](#) 👍



[Roman Battisti](#)

Apr 18, 2021

Hi Dave,

Thanks for posting this. I've worked through all of the setup, but one step is giving me the following error:

nvidia-smi

NVIDIA-SMI has failed because it couldn't communicate with the NVIDIA driver. Make sure that the latest NVIDIA driver is installed and running.

I've checked that my drivers are up to date and even downloaded the most recent driver for Linux, but I still get an error. I'm running a GeForce RTX 2080 Ti, but I didn't see any issues online running this particular GPU. It may not be important whether the smi command can interface with the GPU, but any thoughts would be helpful. Thanks for all the additional time you're putting into helping us.

Edited by [Roman Battisti](#) on Apr 18 at 5:21pm

← [Reply](#) 👍



[Dave DeBarr](#)

Apr 18, 2021

Hi Roman,

First, copy any files you'd like to save before the next step [just in case], then ...

Reboot; and try nvidia-smi at the command prompt once it's done rebooting.

I suspect it just needs a reboot.

← [Reply](#) 👍



[Roman Battisti](#)

Apr 18, 2021

Hi Dave,

I tried the reboot with no success. I'm running ubuntu through Windows 10, should I be using something else, like VirtualBox?

On another note, I've noticed several people have been having long run times. In the code you sent, I noticed we are doubling up on our convolutional layers (lines 77-84).

```
filter_count = 64
```

```
for block in range(convolution_block_count):
```

```
    model.add(layers.Conv2D(filters = filter_count, kernel_size = (3, 3), activation = "relu", padding = "same", kernel_regularizer = regularizer))
```

```
    model.add(layers.Conv2D(filters = filter_count, kernel_size = (3, 3), activation = "relu", padding = "same", kernel_regularizer = regularizer))
```

```
    model.add(layers.MaxPooling2D(pool_size = (2, 2)))
```

```
    filter_count += 64
```

if (convolution\_block\_count > 0):

model.add(layers.Flatten())

I don't recall adding multiple layers like this successively (normally I've seen one layer added per iteration of the loop). Is this intentional? I think this would add to the back propagation runtime, since the convolutional layers are much deeper than intended. Though it's been a while since I last looked at the structure of a convolutional model, so I may be misremembering.

Edit: On second thought, if the second convolution layer wasn't there, there would be a max pooling layer after each convolution layer, which I don't think I've seen before (though I've definitely not seen a wide variety of architectures). So it's probably me just misremembering...

Edited by [Roman Battisti](#) on Apr 18 at 9:15pm

← [Reply](#) 👍

○



[Dave DeBarr](#)

Apr 18, 2021

Your running Ubuntu on Windows 10 using hypervisor? [I can only say I have not tested that configuration.]

It took me 1.5 hours to get through the homework using the Azure Lab Services VM.

Reminder: we're currently treating the convolution blocks as black boxes; but yes, this was intentional. Think of it as being similar to the first convolution block of the Visual Geometry Group's 16 layer model (VGG16):

<https://github.com/tensorflow/tensorflow/blob/v2.4.1/tensorflow/python/keras/applications/vgg16.py#L142-L148> ; or the convolutional block in Example 10.3 of Deep Learning Illustrated.

← [Reply](#) 👍

○



[Roman Battisti](#)

Apr 18, 2021

I'm using the Ubuntu app available through Microsoft. I confess I haven't used Ubuntu before this, so I'm guessing at which application I should be using with my current desktop setup.

A while ago and to test myself a bit, I did a small recreation of (I think) VGG16 using a tutorial. I did it mostly to work through the architecture and learn how skip connections are written in Keras. It's been a while, so I must be misremembering what I did there. Thank you for the clarification.

I ran the code on the VM and it took a similar amount of time.

← [Reply](#) 👍

○



[Michael Young](#)

Apr 18, 2021

FYI -- The lab took me about 4.5 hours to run on the azure virtual machine. I also had to rename the downloaded file.

← [Reply](#) 👍

○



[Dave DeBarr](#)

Apr 18, 2021

That doesn't sound right :\\

Can you post the console output, from the time you start the session until the completion of Trial #1?

It's not terribly important we're leveraging the GPU for this particular assignment, but it'll become more important as the quarter goes on.

← [Reply](#) 👍

 [Dave DeBarr](#)  
Apr 18, 2021

Thanks for the heads-up on the filename! The course number changed.


[← Reply](#) 

 [Michael Young](#)  
Apr 19, 2021

I closed the VM. Do you know if the output is saved in a /var/log or somewhere?

From my memory, I can only remember that trial 6 ish took about 1.5 hours.

[← Reply](#) 

 [Dave DeBarr](#)  
Apr 19, 2021


I don't think it saves the console output unless you ask it to save the console output.


Trial 6-ish should have only been 2 epochs. For it to take 1.5 hours, something must have gone wrong. Please let me know if you encounter trouble in the future.

[← Reply](#) 

 [Rajesh Arora](#)  
Apr 18, 2021

Th lab assignment was taking a longtime for me as well initially. However, I followed revised instructions that Dave posted above and it took about 45 mins finish with the changes.

[← Reply](#) 


 [Michael Young](#)  
Apr 19, 2021


I read through everything and didn't see what changes to make to reduce the time, can you tell me which post?

[← Reply](#) 

 [Dave DeBarr](#)  
Apr 19, 2021

For what it's worth: I'm guessing Rajesh is using a different GPU. My Titan V GPU completes the fashion homework in less than 20 minutes; but it's a newer GPU than the K80.

[← Reply](#) 

 [Andrew Feldman](#)  
Apr 19, 2021

Hi Dave. Question. Since my free time from work can be spotty, I have done my best this semester to get a head-start on the reading, spending 30min-1hr each day. My goal is to complete a first-pass through all the readings ASAP, and then review them in lockstep with the progression of the course. Would it be possible to share the full list of readings for the course? I extracted the following from the Syllabus:

- DLP Chapters 1 & 2
- DLP Chapters 3 & 4
- DLI Chapters 1 & 10, DLP Chapter 5
- DLI Chapters 2 & 11, DLP Chapter 6
- DLI Chapters 3 & 12, DLP Chapter 8
- DLI Chapters 4 & 13

Additionally, the syllabus mentions "Transformers and Sequence-to-Sequence models" but does not indicate the readings associated.

However, the syllabus does not indicate whether the above is a complete list or just a subset of readings.

So would it be possible to clarify,

1. Is the above a complete list?
2. What is the reading associated with "Transformers and Sequence-to-Sequence models"?
3. More generally, if there will be any research papers or other long-form content (i.e. content besides the textbook) that we will be asked to read in future weeks, would it be possible to share that at the outset? Again, I would not ask except that I am trying to avoid the situation where a particular reading assignment comes out simultaneously with a big work obligation such that I can't find time. This was an issue in the first two courses and I ended up getting pretty behind, so trying to avoid that as it was no fun for myself or for Chenwei :)

Thanks so much for all your support.

Best,

Andy

Edited by [Andrew Feldman](#) on Apr 19 at 4:35am

[← Reply](#) 



[Dave DeBarr](#)

Apr 19, 2021

That is the complete list. It's possible you could check out something like [The Annotated Transformer \(harvard.edu\)](#) ; but 1) it's not light reading and 2) there are some necessary tweaks (like learning the position embeddings). The derivative works will assume you've read the initial work.

I'm sure Transformers will make it into textbooks soon, but it's not quite there yet. That said, I am convinced that Transformers are going to be around for a while. It's providing state of the art performance in many tasks now. We'll talk about why in class.

If you have time, consider checking out [The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. \(jalammar.github.io\)](#)

So to recap: we'll chat about attention, transformers, and sequence-to-sequence models in class :)

[← Reply](#) 



[Michael Young](#)

Apr 19, 2021

Dave, have you been able to get CUDA/tensorflow working with GPUs on Windows 10 successfully?

I attempted to run it on the linux subsystem (ubuntu 18), but was having a few issues (nvidia-smi could not connect error). Would like to try my native OS. Have you tried this recently? I see a few guides and wanted to give it a whirl.

I have 2 Nvidia 1080s, and wanted to see if it could take advantage of both of them.

[← Reply](#) 



[Dave DeBarr](#)

Apr 19, 2021

Yes; I have gotten Nvidia GPUs to work on Windows, but this is not a common configuration. It's much more common for folks to use GPUs on Ubuntu machines. In my opinion, it would be much better to use the Azure VMs for now. To use Windows, you'll need to ...

- \* make sure you install the card(s) correctly
- \* install the driver for the card(s)
- \* install CUDA toolkit 11.0, which is expected by Tensorflow 2.4.1 ([GPU support | TensorFlow](#) )

\* install cuDNN library 8.0, also expected by Tensorflow 2.4.1

See: [https://www.tensorflow.org/install/gpu#windows\\_setup](https://www.tensorflow.org/install/gpu#windows_setup)

← Reply 