CO542 Neural Networks Reading Group: Questions for papers

- You are expected to answer the questions for your group if you want to get the 2 marks. You can submit questions for other groups y using the google form https://cepdnaclk.github.io/co542-neural-networks-reading-group/#questions to collect the 0.5*2=1.0 mark for both before and after questions individually.
- You are encouraged to answer questions for more than your assigned paper if you want to develop a deeper understanding of NNs.
- If a question is not clear, please ask the student who asked the question.

Responses from google form up to Submitted 08/03/2021, 23:59 (69) have been considered.

1. **Delving deep into rectifiers: Surpassing human-level performance on imagenet classification**
    a. Demonstrate the need for activation functions using an example NN and a dataset. (Hint: work on the simplest case)
    b. Draw a binary classifier NN with two hidden layers. The fully connected layers should have weights and biases. Consider situations where the hidden layer activations are sigmoid, relu and prelu. Derive the gradients of these layers.
    c. What are the different ways of measuring robustness in ML/NN?
    d. E/15/211: What are the two aspects already driven by the rectifier properties in neural networks?
    e. E/15/016: How does increasing the width in deep models improve the accuracy?
    f. E/15/076: How can we observe the improvement of PReLU over ReLU? What are the various models?
    g. E/15/016 (After): What happens when a_i=0.25 changes? (you don't have to do experiments. Try to predict by reading)
2. **ADAM: A Method For Stochastic Optimization**
    a. E/15/119: What are the differences between Adam and AdaMax algorithms?
    b. E/15/208: AdaGrad decays the learning rate very aggressively as the denominator grows. So what would happen due to this and how can we avoid it?
    c. E/15/065: The RMSProp optimization method lacks a bias correction term. What would be the effect due to this?
    d. E/15/202: It is said that this Adam method is well suited for that are large in terms of data and/or parameters. What is the main disadvantage or what makes it not suitable for small scale datasets?
    e. Consider the function f(x)=exp(w_1x_1+w_2x_2+w_3x_3). Initialize w_1=0.4268, w_2=-0.3421, w_3=0.2310. Consider the data points *f(0,0,0,)=0.01 , f(1.5,1.5,2.0)=0.95 ,f(1,1,0.8)=0.9, f(0.5,0.5,-1)=0.2*. Show the ADAM optimizer (Algorithm 1 in the paper) in action for this data. You may pass individual data points instead of batches to the optimizer.
    f. Compare convex and nonconvex optimization problems.

g. E/15/171: What is the unique thing about adam? Is it proposing something new or combining ideas from previous work?
h. E/15/048: Explain bias correction.
i. E/15/243: What is meant by the training cost in the graphs (that were questioned in the QA session)?

3. **Learning representations by back-propagating errors**
    a. E/15/280: What advantages learning procedure gets, preserving weight symmetry around the middle of the input vector?
    b. E/15/173: How do the weight proportions ensure that eight patterns above the midpoint send a unique activation sum to each hidden unit? (@E/15/280 please clarify which figure you are referring to)
    c. E/15/350: The paper states that this method does not converge rapidly as methods that use second derivatives. However, it can be implemented in parallel hardware, therefore, can produce results faster. What is the meaning of this? Why methods that use second derivatives are not possible to implement in parallel hardware and why this method is possible?
    d. Draw a binary classifier NN with two hidden layers. The fully connected layers should have weights and biases. Consider situations where the hidden layer activations are sigmoid, relu and prelu. Derive the gradients of these layers.
    e. What are the methods of finding gradients? Explain the differences, advantages and disadvantages.
    f. Are there non gradient based methods to train a NN?
    g. E/15/373: What are the disadvantages of backprop?
    h. E/15/179: Give examples to static dataset classification tasks completed by backprop.
    i. E/15/211: What are known as excitatory and inhibitory weights?
    j. E/15/211: Draw a flowchart and explain how backprop, SGD and other main papers we discuss connect in the NN tasks.
    k. E/15/092: Are different gradients prioritized (when we don't have a simple feed forward network eg: resnet)
    l. E/15/299: Show with an example datatpoint why the factor 2 weights are required for symmetry finding. Show how some other set of weights will fail during certain situations while factor two weights always succeed.
    m. Derive the equation derivative(sigmoid)=sig(1-sig) [write this proof in latex]

4. **Training Very Deep Networks**
    a. What are the well-known issues of very deep neural networks? Explain them.
    b. The authors claim that they derived motivation from LSTM architecture. Draw images, write equations and clearly explain these similarities.
    c. What other methods are used by researchers to train very deep neural networks?
    d. E/15/211: There are problems facing Very deep network training. So how can we overcome those by using Long Short Term Memory (LSTM) recurrent networks? (Interesting question. TBH I don't know the answer)
    e. E/15/076: What are some better optimizers for dealing with the difficulties of training deep networks?

f. E/15/281: What are the most suitable initialization strategies in deep neural networks?
g. E/15/243: What is the purpose of using an exponentially decaying learning rate at section 3.1? Are there any performance increasing by using that?
h. E/15/271: How the network dynamically adjust the routing based on the current input (in the section 4.1)?
i. E/15/315: Why the bias vector for the transform gate initialized with a negative values like -1,-3?
j. E/15/315 E/15/181: What are Highway networks?
k. E/15/181: Very deep network training still faces problems. It hard to investigate the benefits of very deep network for variety problems. To overcome thus what was they propose.

5. Gradient-based learning applied to document recognition

a. Shed some light on to the debate about learning and handcrafted algorithms.
b. What are the two distinct forms of data augmentation used in image recognition neural networks? describe them. [question moved here because 11 is full. Unfortunately, can't give marks because this is your own paper. Please ask another question.]
c. E/15/373: Stochastic Gradient algorithm is a popular minimization procedure used to find the local minima, how does it achieve it faster than gradient descent or second-order methods on large training sets with redundant samples?
d. Image classification requires properties like scale/rotation/translation-invariant detection. How do CNNs achieve these?
e. Calculate the number of trainable parameters for LeNet-5.
f. Your presentation talks about "Second-degree polynomial classifier : Input - 40-dimensional feature vector : A linear classifier with 821 inputs". How does these numbers come? Calculate and show 821.
g. What is "boosting"?
h. This paper talks about "sub-sampling". What actually happens in this step? What are different types of sub-sampling? (It will be easier to demonstrate this using an example matrix.)
i. E/15/081: Show how Viterbi algorithm works (pseudocode and steps) for a small graph problem.
j. E/15/351: What are the major advantages of GTN?
k. E/15/202: Are there any papers using LeNet on cursive writing?
l. E/15/233, E/15/119 asked questions and they were cleared out in the QA session satisfactorily.

6. Long short-term memory

a. E/15/171: How Long short-term memory differs from Adaptive sequence chunkers?
b. E/15/260: What are the assumptions or requirements for the Naive approach?
c. Draw an LSTM cell (with 3 time steps). Name different sections of the diagram with their intended uses.

d. What is a statistically significant result? (explain for any experiment. Not necessarily NNs)
e. The results section of page 12 talks about 3*10 trials. What is the use of 3 and 10? What do those multiple trials prove? (Hint: 3 proves one thing and 10 proves something else).
f. What are some challenges in sequence tasks? (e.g. delay problem)
g. Explain which part of the LSTM block handles the exploding/vanishing information flow for long sequences (using the diagram you are using)?
h. Compare the real-world impact of LSTM in comparison to CNNs. Which had more impact? Why?
i. Explain the terms RNN, LSTM, and GRU.
j. Explain newer ideas in comparison to LSTM.
    i. Bi-directional LSTM
    ii. Time CNN
k. Questions below this need not be answered in the booklet
l. E/13/087: We know that LSTM is the further developed architecture to RNN architecture. How are recurrent cells in RNN and memory cells in LSTM different from each other? What additional mechanism does a memory cell have compared to a recurrent cell?
m. E/15/142: In the Experiment 3, it is described as LSTM does not encounter fundamental problems if noise and signals are mixed on the same input line. Why does that happen, can you elaborate? Also, what would happen to the noise if Kalman filter training algorithm was introduced into the network?
n. E/15/243: How is Constant error approach differ from Exponentially Decaying Error (In section 3)
o. E/15/271: Why Conventional BPTT use both local and global error flow?
p. E/15/362: Is initially biasing the input gate inj towards zero, the only effective way of solving drift problems at the beginning of learning? If not what are the alternative methods?
q. E/15/179: What is referred to as backpropagation through time in simple RNN?
r. E/15/238: Are there other methods for long-term temporal dependencies learning apart from the one used in this paper?
s. E/15/330: What is “long” and “short” in LSTM?
t. E/15/330: Is it always better to using stacked LSTM than single LSTM?

~~7. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion~~

a. E/15/373: What are the advantages of denoising autoencoders when compared to ordinary autoencoders?

8. Generative adversarial nets

a. E/15/279: How Generative adversarial nets different from adversarial examples?
b. E/15/325: What is the difference between input samples in a generator and a discriminator network?
c. What is a Markov chain? What is the equation for the markov property? Identify all the terms.

d. E/15/077: Why adversarial networks better than methods based on Markov chains when considering distributions?
e. Explain all the terms in equation (1).
    i. When you are explaining 1-D(G(z)) you should explain every term as given below.
        1. What is z?
        2. What is G?
        3. What is G(z)?
        4. What is D?
        5. What is D(G(z))?
        6. Why do we need 1-D(G(z))?
    ii. You have to explain the full equation. Not just this term.
f. A basic generator can output only a certain class of images. What more advanced types of generators. Give examples you can use that idea for.
g. Discuss the global issues coming from the wrongful use of GANs.
h. Explain the pros and cons of using a GAN to generate data for a limited dataset issue.
i. Explain the words Infimum and supremum using
    i. Words
    ii. Set notation
    iii. An example
j. What is a min-max game? How similar/different is it to GAN?
k. E/15/362: If we use any other gradient-based learning rule other than momentum, how would the obtaining results vary?
l. E/15/154: What is the role of noise and noise sampling?
m. E/15/081: Explain in steps and equations how the discriminator is useful for the training of the generator in GAN? (question corrected)
n. E/15/238: How does the discriminator being too good, impacts the generator training to fail?

9. Compression of deep convolutional neural networks for fast and low power mobile applications

a. E/15/208: When can we use fine-tuning in deep convolutional neural networks? (i.e, how should be the datasets)
b. E/15/209: When is it trivial to compress the whole convolutional neural network?
c. E/15/209: What is the reason to minimize the reconstruction error of linear kernel tensors in Tucker decomposition?
d. E/15/351: How the one-shot whole network compression scheme reduces the power consumption in the main memory?
e. Prove the complexity equations by calculating the number of operations.
f. E/15/246 Explain caffeinated convolution.
g. E/15/350 Will PCA work? (please explain this question)
h. E/15/209: Explain the ideas of stride and padding with diagrams.
i. Explain different places NN are deployed (server, edge, fog). What are the power/network/performance considerations?

j. How is the rank selection done?

~~10. Pytorch: An imperative style, high-performance deep learning library~~

11. ImageNet Classification with Deep Convolutional Neural Networks

a. What are the famous datasets/competitions for image classification?
b. Explain the words like training/validation/testing sets.
c. What are the metrics used in image classification tasks?
d. What are other image-related tasks (except for classification)?
e. Compare this NN to the newer image classification networks. What are the differences you see between this and newer ones?
f. E/15/065: What is the purpose of using dropout layers in convolutional neural networks?
g. E/15/073: Are there any advantages when directly applying a pooling layer after each convolutional layer that is used in this paper rather than stacking multiple convolutional layers?
h. E/15/211: Pooling layers are used in CNN's architecture. What is the condition to occur overlapping pooling?
i. Calculate the correct number of trainable parameters in this NN
j. Explain the words like translational, rotational invariance. (there are around ⅘ such terms)
k. E/15/073 Why is RELU better than sigmoid?
l. E/15/208 What are the reasons for CNNs to outperform DNN for image tasks?
m. E/15/369 How does this fine-tuning happen? Explain the transfer learning procedure.
n. E/15/316 Pros and cons of doing data augmentation on the CPU versus GPU.

~~12. Simple Online And Realtime Tracking With A Deep Association Metric.~~

a.

13. Attention is all you need

a. E/15/279: Why do we need positional encoding?
b. E/15/280: What is the purpose of using an encoder-decoder structure in neural sequence transduction models?
c. You can find the recording of the class https://cepdnaclk.github.io/co542-neural-networks-reading-group/#schedule here. Watch the explanation again, watch some more youtube videos or tutorials to understand the concept, ask for clarifications and record the presentation again with a simple explanation on attention mechanism.
d. E/15/325: How are attention weights calculated? (there is a trainable part. Explain the cost functions used in this training).
e. E/15/348: What are the main differences between beam search and greedy search?
f. E/15/369: Why softmax? [explain the properties of softmax output].
g. E/15/299: How do Q and K come into play? Explain with an example.

~~14. Image style transfer using convolutional neural networks~~

15. Vqa: Visual question answering

a. What are some simpler image-related/language-related tasks that have to be solved before solving VQA?
b. Explain the 3 main AI tasks that should combine to build a VQA system. What are the NN architectures used in these 3 main fields?
c. Explain what are NP completeness, Turing completeness, and AI completeness. Where does VQA fall into?
d. There is a set of questions where (1) the NN doesn't need the image or a caption to answer (2) the NN doesn't need the image to answer.
    i. Pick an image, write a caption.
    ii. Give 3 questions where you
        1. Don't need anything
        2. Need s the caption only
        3. Needs both the image and caption.
    iii. Do not use the images form this paper.
e. This paper uses COCO dataset. Give some description about this dataset. What is there, how many images, what categories, what resolution? What tasks can be done on COCO?
f. Assume a VQA system with one of the following capabilities
    i. Semantic segmentation
    ii. Instance segmentation
    iii. Panoptic segmentation
    iv. Pick an image, assume that the VQA has only one such capability. Write one question it will answer properly and one question it will fail at.
g. The paper states the algorithm uses a 1030 dimensional representation using the top 1000 words and bag of words technique. Consider an array **ar** with 5000 words / 1500 unique words. Write the pseudocode for generating this representation. (if pseudocodes are boring, write the python code)
h. Draw the Q+I+C NN architecture mentioned in this paper. You can draw only the first and last layers for VGGnet.
i. What are these abstract images? What is learned from them? Explain.
j. @Imesh, assume that someone else is presenting this paper. What are the questions you will ask? Write one of them here and answer it. [Please don't say "I am not going to ask any question". We know you will]
k. E/15/281: Can we take an open-ended task as a subset of a multiple-choice task ?
l. E/15/154: What are the constraints that reduces the accuracy? Does there any special characteristics have in COCO dataset to overcome those limitations? Or can we use any image dataset other than COCO?
m. E/15/315: What are vision+language tasks?
n. E/15/181: what is multi modal knowledge?

16. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning

a. E/15/287: What is the difference between sample-based variational inference and stochastic variational inference?

b. E/15/287: How is the suitable dropout probability determined? Is there a specific way or is it a trial and error process?
c. E/15/123: How the Model uncertainty can be quantified by looking at the entropy or variation ratios of the model prediction?
d. E/15/316: When adapting model uncertainty in reinforcement learning with a dropout Q-network, will the likelihood of exploitation increase with the episodes?
e. E/15/173: Why do we need a Bernoulli distribution? Derive a Gaussian distribution starting from a Bernoulli distribution.
f. E/15/187: How does dropout prevent overfitting?
g. E/15/048: How is training and testing different in RL compared to supervised learning?
h. E/15/123: Is there any specific reasons why the particular datasets were used? What is special about them?
i. What is the probability of 99% of the neurons dropping out?
j. What are the two different types of uncertainty?

17. <u>Datasets</u>

a. Make a table with the dataset, data type, tasks, and papers using this dataset. (for all the datasets used in papers 1-16).
b. Image datasets have images of different sizes. Plot graphs for different image-based datasets using the following parameters as axes. Note: If there are X parameters, you will have to draw ${}^{X}C_{2}$ graphs.
    i. Year released.
    ii. Resolution (in megapixels)
    iii. Datapoints in the dataset.
    iv. No of classes.
c. Consider an imbalanced dataset. What are the challenges in training an NN on this dataset? How to overcome those challenges?
d. Name a few open challenges that cannot be solved because of the absence of a dataset.
e. Some machine learning tasks are done using synthetic datasets due to the absence of real datasets. What are the pros and cons of this approach?
f. Explain the concept of unimodal, bimodal, and multi-modal data. Give one or two examples.
g. What type of datasets demands a meta-learning approach? Explain with examples.
h. List at least 5 datasets specific to Sri Lankan context with links.
i. Creating a dataset (for tasks like face detection or medical diagnostics) introduces a privacy issue. How is this handled?
j. Consider the data being collected at the University of Peradeniya, faculty of engineering. Tabulate this data with the following columns
    i. Data
    ii. Form (electronic or not)
    iii. Amount of datapoints
    iv. What can you do with this data?

v. Privacy concerns in using them

18. **Loss functions**
    a. What is optimization? (in mathematics)
    b. Map the mathematical concept of optimization to NNs by using a table. One column should have the term form mathematics and the other column should have the term denoting that idea in NN.
    c. What is the maximum likelihood?
    d. Make a table with the loss function, equation, tasks that it is used to, and papers using this loss function (for all papers 1-16).
    e. What properties are important for a mathematical function to be used as a loss function?
    f. Explain what KL divergence is. Why is it not considered as a distance? (explain with an example and proof).
    g. What is regularization?
    h. How can applying a loss function over the weight values act as a regularization step?
    i. Explain the situations where the loss function value of very different for the training and testing sets.
    j. Consider a situation where we need an NN to classify images. Assume that the performance metric we need to improve is the classification accuracy. Why don't we optimize the NN to minimize the negative value of classification accuracy as the loss value?
    k. E/15/142: In the paper, Stacked Denoising Autoencoders by Pascal Vincent at el, could you please explain about the association of loss function in equation 4 for binary x values?

Questions that don't get marks (ask another question)

1. **Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.**
2. Adam
    a. E/15/092: Will the decaying learning rate make it locally optimized? *Reason: where does it say adam has a decaying learning rate?*
3.
4. **Training Very Deep Networks**
    a. E/15/208: Are there any other approaches to train very deep networks other than what is introduced(highway network) in the paper? *Reason: I have already asked this question.*
    b. E/15/280: In the 4th paper (Training very deep network), as a solution to the bottleneck caused by stacking several non-linear transformations LSTM recurrent networks introduced. What approaches LSTM takes to overcome this problem? *Reason: Enough questions*
    c. E/15/315: Why the dimensionality of input and output should be the same? *Reason: Where is this said? Please elaborate to get marks*

d. E/15/315: how does the backpropagation work in deep neural networks? *Reason: what?*
e.
f.
g.
h.
i.
j.
k.

5. **Gradient-based learning applied to document recognition**
    a. E/15/208: *Reason: Your question was already here. Please ask a new one.*
    b. E/15/092: *Reason: tSNE was not even invented back when this paper was published. Please ask a new one.*
    c. E/15/363: *Reason: It was more of a comment than a question. Please ask a new one.*
6. **LSTM**
    a. E/15/081: Why did they used similar number (256) of training and test strings in Experiment 1?Is there any specific reason? *Reason: This is just a random question about a single experiment. Not a valid question.*
7.
8. **GAN**
    a. E/15/233: It is mentioned that they can train both, generative model and the discriminative model using only the highly successful backpropagation and dropout algorithms. What are the parameters used to recommend that these two algorithms are highly successful? *Reason: This course exists because these algorithms are highly successful*
9. **Compression of deep convolutional neural networks for fast and low power mobile applications**
    a. E/15/362:What kind of drawbacks we could get by selecting the rank by considering data-driven one-shot decision via empirical Bayes with automatic relevance determination (ARD) prior? Are those recoverable by any other existing method? *Reason: We have finished this paper weeks ago. Ask questions from remaining papers.*
10.
11.
12. **ImageNet Classification with Deep Convolutional Neural Networks**
    a. E/15/139: How can we determine number of parameters that can be fit into a model without overfitting? How does the number of data points affects this? *Reason: We have concluded this paper. Too late to ask questions.*
    b. E/15/139: What is the use of stride on the input layer and how do we determine the optimal stride size? *Reason: We have concluded this paper. Too late to ask questions.*
13. **Attention**

a. E/15/363: A question about a hyperparameter. *Reason: If we start giving marks for questions like "why are there N layers or M neurons" we will not have any insightful questions coming*

14.

15. VQA

a. E/15/271:What are the difficulties of the VQA dataset for the MS COCO images using several baselines and novel methods? *Reason: What does this even mean? Please resubmit.*
b. E/15/315: What is natural language processing?
c. E/15/181: What is meant by text- based Q&A

16.

17.

18. Loss Functions

a. E/15/315: What are common loss functions using in neural networks? *Reasons: Please :-)*