Test 02 – Networks

Arthur J. Redfern
arthur.redfern@utdallas.edu
Oct 30, 2019

---

0 Instructions

- There are 41 numbered questions with indicated point values that sum to 100
- Write all of your answers clearly on the answer sheet and turn it in
- No reference materials are allowed
- No help from others is allowed
- Correct answers in red

1 Test

Design [43 points]

1. [6 points] Circle true or false for each of the following statements
   - True / False You design a network to accomplish a goal
   - True / False Under some mild conditions xNNs are universal function approximators
   - True / False A function always exists that maps an arbitrary input to arbitrary output
   - True / False In designing networks you need to think about how the network combines and transforms input data to extract information at the network output
   - True / False The complexity of the function required to map natural images to a 1000 classes is the same as the complexity of the function required to map random images to a 1000 arbitrarily defined classes
   - True / False Error calculation summarizes the accuracy of the function approximation with a scalar

2. [2 points] CNNs exploit spatial structure to extract information from input images.

Consider an ImageNet image classification network that maps 3 x 256 x 256 input images to a 1 x 1000 vector where each vector element corresponds to 1 of the 1000 classes.

3. [2 points] How many levels of / 2 down sampling are typically included in the encoder (tail and body) portion of the network?

5

4. [2 points] List the layers in a reasonable tail design which maps the 3 x 256 x 256 input image to a 64 x 64 x 64 feature map. Use the notation (output channels x input channels x filter rows x filter cols / stride) for CNN style 2D conv layers and (pool region rows x pool region cols / stride) for pooling layers.

64 x 3 x 7 x 7 / 2 CNN style 2D conv followed by 3 x 3 / 2 max pool
Other options with 2 levels of down sampling are possible

5. [4 points] List 4 different classes of network building block configurations used in the body.

Sequential, parallel, dense and residual

6. [2 points] Assume that the output of the encoder is a 2048 x 8 x 8 feature map. List the layers that comprise a modern classification decoder (head) design.

Global average pool with 1 x 2048 output followed by
Fully connected layer with 2048 x 1000 weight matrix and 1 x 1000 bias and no nonlinearity followed by
Soft max (for training) or arg max (for testing)

7. [2 points] If the last trainable layer in the decoder is a M feature x 1000 output class weight matrix, what range M is characteristic of top performing ImageNet classification networks?

$M \geq 1000$

Let's say that the network is a ResNet design and trained with 3 x 256 x 256 input images.

8. [2 points] Circle yes or no

Yes / No Is the ResNet style encoder (tail and body) mathematically compatible with a 3 x 512 x 1024 input image?

Yes / No Is the modern decoder (head) design mathematically compatible with a 3 x 512 x 1024 input image?

Consider an encoder with the following structure:

64 x 3 x 7 x 7 / 2	CNN style 2D conv
3 x 3 / 2	max pool
64 x 64 x 1 x 1 / 1	CNN style 2D conv
64 x 64 x 3 x 3 / 1	CNN style 2D conv
256 x 64 x 1 x 1 / 1	CNN style 2D conv
128 x 256 x 1 x 1 / 2	CNN style 2D conv
128 x 128 x 3 x 3 / 1	CNN style 2D conv
512 x 128 x 1 x 1 / 1	CNN style 2D conv

9. [4 points] Calculate the receptive field size at the output of the above encoder.
   1, 1 + 0 = 1, 1 + 2 = 3, 3*2 – 1 + 0 = 5, 5 + 0 = 5, 5 + 2 = 7, 7 + 0 = 7, 7*2 – 1 + 2 = 15, 15*2 – 1 + 6 = 35
   Answer = 35

10. [2 points] RNNs exploit sequential structure to extract information from input data.

11. [2 points] What type of nonlinearity is used as a gating mechanism in GRU and LSTM cells?
    Sigmoid

12. [3 points] Circle true or false for each of the following statements
    - True / False When multiple inputs are processed in a batch, all matrix operations in a RNN layer can be converted from vector matrix multiplication to matrix matrix multiplication
    - True / False RNN layers can be stacked such that the sequential outputs of 1 layer become the sequential inputs of the next layer
    - True / False RNNs have to process information in the same sequential order as humans commonly do in order to generate meaningful features from the input sequence

13. [2 points] This type of RNN structure down samples the number of vectors in the input sequence
    Pyramidal

14. [2 points] Circle true or false for each of the following statements
    - True / False Attention maps an input feature matrix and an input query vector / matrix to an output feature vector / matrix
    - True / False Self attention uses the input feature matrix to create a query matrix

Consider the single headed self attention mapping $\mathbf{Y}^T = \mathbf{A}^T \mathbf{X}^T \mathbf{H}$ where $\mathbf{Y}^T$ is an output feature matrix, $\mathbf{A}^T$ is an attention matrix, $\mathbf{X}^T$ is an input feature matrix, $\mathbf{H}$ is a weight matrix and all matrices are N x N. Assume that each row of $\mathbf{X}^T$ and $\mathbf{Y}^T$ corresponds to a feature vector.

15. [2 points] What is the sum of all of the elements in $\mathbf{A}^T$?
    N

16. [2 points] Which matrix allows mixing across different input feature vectors?
    $\mathbf{A}^T$

17. [2 points] Which matrix allows mixing within a single input feature vector?
    H

Training [25 points]

18. [2 points] Circle true or false for each of the following statements
    - True / False The same things that allow xNNs to achieve a high level of accuracy across a wide range of problems – a large number of trainable parameters and universal function approximation – also make xNNs susceptible to overfitting training data.
    - True / False Various types of regularization methods are used to prevent overfitting to training data and improve generalization to testing data.

19. [2 points] List a common method for preventing individual weight values from becoming too large which frequently leads to overfitting.
    Adding a L1 or L2 regularization term to the error

20. [3 points] Circle true or false for each of the following statements
    - True / False The accuracy of the network output is probably not affected by testing data that differs significantly from training data.
    - True / False A challenge in training with synthetic data is labeling.
    - True / False Advances in computer graphs enable the generation of more realistic looking synthetic images.

21. [2 points] List 2 methods for initializing trainable parameters in a xNN.
Random uniform or truncated Gaussian initialization (Glorot / Xavier, He, …), transferring from a pre trained network (transfer learning)

22. [2 points] List the 2 arguably most important / commonly used network components / modifications that have allowed for the training of arbitrarily deep CNNs
Batch normalization, residual connections

23. [3 points] Consider a CNN style 2D convolution layer composed of the following components: convolution, bias and ReLU nonlinearity. List the components that are present during training if batch normalization is included with the layer
Convolution, batch normalization and ReLU (bias is handled by batch normalization)

24. [3 points] Circle true or false for each of the following statements
True / False Residual connections can be used with CNNs
True / False Residual connections can be used with RNNs
True / False Residual connections can be used with self attention based networks

25. [2 points] In old school CNN image classification decoder (head) design with multiple large fully connected layers, this regularization method was used to help prevent overfitting and implicitly learn an ensemble of classifiers.
Dropout

The Adam method for stochastic weight optimization computes **s**, an exponential moving average of the gradient, and **r**, an exponential moving average of the elementwise square of the gradient, both across batches. After correcting for a bias in the calculation, updates to weights are

$$\mathbf{h} \leftarrow \mathbf{h} - \alpha\, \mathbf{s}./(\mathbf{r}^{.1/2} + \delta)$$

with ./ used to indicate elementwise division and .1/2 used to indicate elementwise square root. δ is a small value.

26. [2 points] Circle true or false for each of the following statements
True / False The exponential moving average used to compute **s** effectively creates a momentum effect.
True / False On a per coordination direction basis, dividing by $(\mathbf{r})^{.1/2}$ makes small weight updates smaller and large weight updates larger.

27. [2 points] What learning rate schedule was used to adapt α in the homework code examples I provided.

Linear warmup with cosine decay

28. [2 points] In data synchronous parallel training consider a system with 1 master, 16 worker machines and a batch size of 32 per worker machine. What is the effective batch size per weight update made by the master machine?

16*32 = 512

Implementation [32 points]

29. [4 points] Circle true or false for each of the following statements

True / False xNNs are typically specified as hardware agnostic graphs
True / False xNN performance is a measure of how long it takes a network to run
True / False xNN performance is independent of the hardware on which the xNN runs
True / False Appropriately sizing the xNN for the hardware on which it runs is an important component of efficiently utilizing the hardware

30. [3 points] How does quantization affect memory, communication and computation

Memory scales linearly with the number of bits
Communication scales linearly with the number of bits
Addition and comparison operations scale linearly with the number of bits, multiplication operations scale to the square of the number of bits

31. [3 points] Circle true or false for each of the following statements

True / False xNN graph lowering includes hardware agnostic, aware and specific components
True / False xNN graph lowering includes domain agnostic and specific components
True / False xNN graph lowering includes code generation

32. [2 points] Moore's law states that the number of transistors on an integrated circuit doubles every ~ 2 years at a constant cost.

Consider inputs

L = transistor feature size

V = voltage

and approximate semiconductor device physics values

C = capacitance per transistor ($\propto$ L)

D = area density ($\propto 1/L^2$)

E = energy per transistor use ($\propto CV^2$)

f = frequency ($\propto 1/L$)

P = power per area ($\propto$ DEf)

where $\propto$ means "proportional to".

33. [4 points] How does the power per area P change if the transistor feature size L shrinks by 1/2 (i.e., 2 generations of process scaling) but the voltage V stays the same?

C → C/2
D → 4D
E → E/2
f → 2f
P → 4P (so power per area increases by 4x)

Consider a DSA that's designed for xNNs and includes local memory, control, communication and computation.

34. [2 points] Assume for a CNN that all the filter coefficients for all the layers will never fit on the device at the same time. How much local memory would you include in the DSA to optimize performance?

A sufficient amount to keep feature maps fully on device

35. [2 points] This type of buffering strategy allows external to internal communication, internal to compute communication and compute to execute in parallel.

Ping pong

36. [1 points] Circle true or false for each of the following statements

True / False External memory to local memory bandwidth is typically much larger than internal memory to compute bandwidth

37. [2 points] List 2 computational primitives that are useful for enabling high performance xNN implementations

Matrix multiplication, sort

Consider CNN style 2D convolution (just the convolution, no bias or nonlinearity) with the following:

Input feature maps (N_i x L_r x L_c) of size	16 x 256 x 512
Row pad = col pad of size	2 (i.e., output size == input size)
Filter (N_o x N_i x F_r x F_c)	32 x 16 x 3 x 3
Output feature maps (N_o x M_r x M_c) of size	32 x 256 x 512

38. [2 points] If this operation is lowered to matrix matrix multiplication, how many MACs are required using standard inner product based matrix multiplication (no need to multiply out factors)?

Using BLAS M, N and K notation

$M_{problem}$	$= N_o$	$= 32$
$N_{problem}$	$= L_r * L_c$	$= 256*512$
$K_{problem}$	$= N_i * F_r * F_c$	$= 16*3*3$
$MAC_{problem}$	$= M_{problem} * N_{problem} * K_{problem}$	$= 32*256*512*16*3*3$

39. [2 points] If this operation is implemented on a hardware matrix multiplication primitive only capable of 32 x 32 block matrix multiplication (i.e., even if the matrix is smaller it still does the full 32 x 32 matrix multiplication), how many MACs does the hardware compute (no need to multiply out factors)?

Using BLAS M, N and K notation

M_{hw}	$= ceil(M_{problem}, 32)$	$= M_{problem}$
N_{hw}	$= ceil(N_{problem}, 32)$	$= N_{problem}$
K_{hw}	$= ceil(K_{problem}, 32)$	$= 16*3*3 + 16 = 16*10$
MAC_{hw}	$= M_{hw} * N_{hw} * K_{wn}$	$= 32*256*512*16*10$

40. [2 points] Divide the theoretical MACs by the hardware MACs to determine the compute efficiency for this specific operations.

Efficiency = $MAC_{problem}$ / MAC_{hw} = (3*3)/10 = 0.90 or 90 %

41. [3 points] To reduce the cost of a calculation, which of the following are generally used strategies? Circle true or false for each of the following statements

True / False	Exploiting that different operations have different costs and doing less of the higher cost operation at the expense of doing more of the lower cost operation
True / False	Creating intermediate terms that can be re used
True / False	Recursively applying the intermediate term re use strategy