Test 03 – Applications

Arthur J. Redfern
arthur.redfern@utdallas.edu
May 01, 2019

---

0 Instructions

Logistics

- Name: _____
- UT Dallas ID: _____

Instructions

- There are 30 numbered questions with indicated point values that sum to 100
- Write all of your answers clearly on this test and turn it in
- No reference materials are allowed
- No help from others is allowed
- Correct answers in red

1 Test

Vision [34 points]

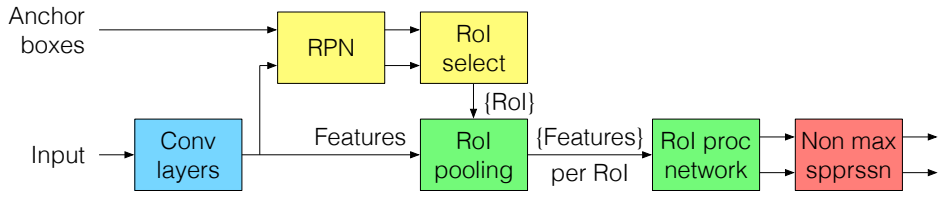
1. [6 points] A key challenge in vision network design is to create features that are both strong (good for classification) and spatially well localized. List 3 methods for creating strong spatially well localized features:

3 of the below 4 possibilities
Skip connections
Spatial pyramid pooling

Atrous convolution
Feature pyramids

The Faster R-CNN approach to multiple object detection includes:

- Input image and pre determined anchor boxes
- Convolutional layers to map from image to features
- A region proposal network with region of interest selection
- Region of interest pooling to create fixed size feature maps for each selected region of interest
- A region of interest processing network that processes fixed size feature maps from each region of interest
- Non maximal suppression



2. [4 points] List 4 possible CNN layer building block types that could be used for the conv layers:

Sequential building blocks (VGG 16 or 19, MobileNet V1, …)
Parallel building blocks (GoogLeNet, Inception V2, …)
Densely connected building blocks (DenseNet 121, …)
Residual building blocks (ResNet 50, MobileNet V2, …)

3. [4 points] List 2 key considerations you as a designer would balance with respect to the choice of the conv layers in a practical real time embedded system implementation:

Performance on available hardware
Accuracy in the context of the full network

4. [4 points] What are the 2 outputs of the region proposal network?

Classification of {object, no object} for each anchor box
Regression to refine the anchor box bounding box coordinates

5. [2 points] Circle true or false for each of the following statements

True / False The region proposal network acts like an attention mechanism

6. [4 points] What are the 2 outputs of the region of interest processing network?
   Classification of {class 0, class 1, class 2, …} for each region of interest
   Regression to refine the bounding box coordinates for each region of interest

7. [4 points] Circle true or false for each of the following statements
   True / False Non maximal suppression removes or combines overlapping predictions of the same object
   True / False 1 point in a precision recall curve is generated for each choice of confidence threshold (the value over which predictions are kept, under which predictions are discarded)

8. [2 points] Mask R-CNN for object based image segmentation adds a 3rd output to the region of interest processing network. What does this 3rd output do?
   Classifies each pixel in the fixed size feature map as {part of object, not part of object}, effectively creating a segmentation mask

9. [4 points] Circle true or false for each of the following statements
   True / False Finding the same point in 2 spatially separated images of the same scene, a key component of stereo depth estimation, can be formulated as a classification problem
   True / False Finding the same point in 2 images of the same scene separated in time, a key component of motion estimation, can be formulated as a classification problem

Speech [33 points]

Common pre processing for speech includes segmenting the speech waveform into overlapping vectors and transforming each of the vectors (commonly with a DFT followed by additional dimensionality reduction related processing).

10. [4 points] Circle true or false for each of the following statements
   True / False If a non invertible MFCC transformation is used for the dimensionality reduction, then no information from the original speech waveform is lost
   True / False It's possible to use a RNN for processing the pre processed speech vectors
   True / False Stacking the pre processed speech vectors shoulder to shoulder in a feature map allows a CNN to be used for processing the pre processed

speech signal

True / False Stacking the pre processed speech vectors shoulder to shoulder in a matrix allows a self attention based network to be used for processing the pre processed speech signal

Consider a pre trained speaker identification network that maps speech signals to feature vectors and has a database of feature vectors for many speakers

11. [8 points] Circle true or false for each of the following statements

True / False A reasonable strategy for speaker identification is to compute a feature vector for a speaker, compare the cosine of the angle between the speaker's feature vector and all of the feature vectors in the database, find the best match (smallest angle), and declare that to be the speaker if the match is above a threshold (angle is below a threshold)

True / False A challenge of having a database of many speakers and shorter feature vector lengths is that feature vectors get closer together, the distance between different speakers decreases in feature space and it generally becomes more difficult to distinguish different speakers

True / False Speaker identification works equally well in noisy and quiet environments as noise does not affect xNN accuracy

True / False In order for this strategy of speaker identification to work without requiring a specific phrase, some portion of the feature vector needs to be invariant to the specific phrase

12. [3 points] Order the following operations from least power (1) to most power (3) as used in a key word spotting for voice wake up application:

2 Voice activity detection

3 Key word (wake word) detection

1 Sound detection

13. [2 points] Circle true or false for each of the following statements

True / False Command recognition can be accomplished by a xNN that maps from an input of pre processed speech vectors to an output vector of length (C + 1) x 1, where C is the number of commands and +1 is an extra class to represent none of the known commands

14. [2 points] List 1 key challenge in working with typical training data for speech to text transduction:

Unknown alignment between the speech waveform and the label text

The network must produce plausible language is also ok, but this really a challenge for network design and the other answer is better

15. [4 points] Consider a CTC network (decoding method) for speech to text transduction. What is the CTC decoded output of the following sequence where a dash "–" is used to indicate the special blank character and an underscore "_" is used to indicate a space:
Y_{greedy} = (i, i, i, _, s, s, p, e, l, l, l, –, l, l, l, l, _, r, r, r, e, e, –, e, l, y, y, _, b, a, a, –, a, a, d, d)
CTC decoded(Y_{greedy}) = i spell reely baad

16. [2 points] Circle true or false for each of the following statements
True / False CTC decoding is a 1 to 1 mapping
True / False If the network predicts graphemes then there are no out of vocabulary words

17. [2 points] The RNN transducer model of speech to text transduction includes a joint network that combines inputs from 2 networks: the 1st is a network that predicts phonemes, graphemes or word pieces from the pre processed speech vectors. What is the input and predicted output of the 2nd network?
The 2nd network is a network that takes as an input previously predicted phonemes, graphemes or word pieces and predicts the next phoneme, grapheme or word piece.

18. [2 points] Assume that the final output of a speech to text transduction network is matrix of grapheme probabilities x output characters. As an example with indexing that starts at 0, the value at (5, 9) would be the probability that "E" is the 10th output character. Also assume that an external language model has been trained on a large amount of text to predict probabilities for the next grapheme given the previous graphemes. What is the typical method used to combine predictions from the speech to text transduction network with predictions from the external language model to create better predictions?
Beam search

19. [4 points] Text to speech systems commonly use a 2 part strategy. What are these 2 parts?
1. Convert text to an intermediate representation (e.g., text normalization followed by an attention based network that converts the normalized text to a mel scale spectrogram)
2. Convert the intermediate representation to audio (e.g., modified WaveNet style audio generation from mel scale spectrograms)

Language [29 points]

20. [6 points] What type of embedding is typically used for:

Characters (assume a small number)?	1 hot
Words?	Dense vector
Sentences?	Dense vector or matrix

21. [2 points] What is the distributional hypothesis?

Words that are used in the same context tend to have the same meaning

22. [4 points] Circle true or false for each of the following statements

True / False All individual dense word embeddings can be performed by matrix multiplication of an embedding matrix with a 1 hot encoded word vector

True / False The skip gram variant of Word2Vec learns an embedding by learning to predict words in a context window around the target word

23. [4 points] Count based N gram language models are based on the chain rule of probability. What is the result of the chain rule of probability applied to $P(w_{n-1}, w_{n-2}, \ldots, w_1, w_0)$?

$P(w_{n-1}, w_{n-2}, \ldots, w_1, w_0) = P(w_{n-1} \mid w_{n-2}, \ldots, w_1, w_0)\, P(w_{n-2}, \ldots, w_1, w_0)$

24. [3 points] Circle true or false for each of the following statements

True / False NNs can be used for language modeling

True / False CNNs can be used for language modeling

True / False RNNs can be used for language modeling

25. [2 points] Circle true or false for each of the following statements

True / False Language translation via word substitution using a bi lingual dictionary works well

26. [2 points] Consider a sequence to sequence language translation system with an encoder RNN variant that creates an output thought vector used to initialize the hidden state of a decoder RNN variant. What is a drawback of this architecture?

The meaning of the full sentence is encoded into a single vector

27. [4 points] At the encoder, attention based language translation networks create a matrix of features via stacking individual feature vectors shoulder to shoulder. The decoder then uses an attention mechanism on the encoded matrix of features to create state dependent feature

vectors used for generating words. What are 2 common types of network building blocks used for the encoder and decoder in this encoder – attention – decoder based language translation network?

RNN variants (RNN, GRU and LSTM building blocks in various configurations)
Self attention variants (multi head self attention building blocks in various configurations)

28. [2 points] What does the attention mechanism allow at the decoder with respect to the alignment between input features from language 1 with output words of language 2?

Attention allows for a non monotonic alignment between output words and input features

Games [0 points]

This section of questions was removed due to weather cancelling class and the makeup lectures being optional.

Art [4 points]

As we also didn't cover art during the course, I'll only ask 2 incredibly difficult "art" related questions

29. [2 point] Art is a nickname for this person's 1st name (hint 1: look at the top of the 1st page; hint 2: a person in the front of this room has this 1st name)

Arthur

30. [2 points] Circle true for each of the following statements

True A person with the nickname Art wishes you the best in your academic journey, subsequent careers and life!