Test 03 – Applications

Arthur J. Redfern
arthur.redfern@utdallas.edu

Dec 04, 2019

0 Instructions

- There are 26 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 [28 points]

1. [4 points] A key challenge in vision network design is to create features that are both

    A. Strong (good for classification)

    B. Spatially well localized

2. [4 points] What 2 types of feature maps does the skip connection approach to pixel classification (semantic segmentation) combine?

    A. Shallow weak features that are spatially well localized

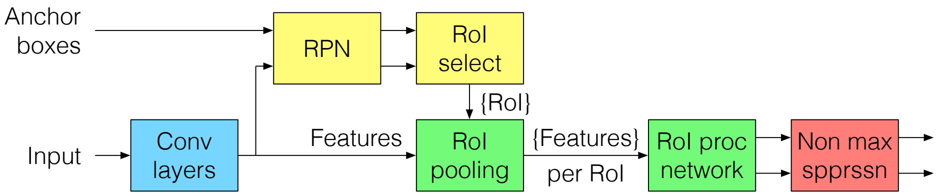
    B. Deep strong features that are spatially poorly localized

3. [4 points] Instead of increasing the receptive field size through pooling, what does the atrous convolution approach to pixel classification do to the filter?

    Up samples the filter

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

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



4. [4 points] What are the 2 outputs of the region proposal network?
   A. Classification of {object, no object} for each anchor box
   B. Regression to refine the anchor box bounding box coordinates

5. [4 points] What are the 2 outputs of the RoI processing network?
   A. Classification of {class 0, class 1, class 2, …} for each RoI
   B. Regression to refine the bounding box coordinates for each RoI

6. [4 points] Mask R-CNN for object based image segmentation adds a 3rd output to the RoI 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

7. [4 points] Say a skip connection based pixel classification (semantic segmentation) network and a Mask R-CNN network are both used to label all the pixels in an image. What additional info (if any) does the Mask R-CNN network provide that the skip connection network does not?

   Which object pixels belongs to

Language [30 points]

8. [2 points] What type of learning do the xNN based embedding methods and xNN based language modeling methods we discussed all use for network training?

Self supervised learning (will also accept unsupervised learning)

9. [2 points] What type of vectors are typically used for character embeddings?

1 hot

10. [2 points] What do individual word embedding methods rely on for assigning dense vectors?

The distributional hypothesis or words used in the same context tend to have similar meanings

11. [2 points] All dense individual word embeddings can be written as the multiplication of 1 hot row vector (representing the word) and this

A matrix where each row corresponds to the dense vector embedding

12. [6 points] CBOW and skip gram based Word2Vec individual word embeddings are trained with a how many layer neural network? After training, which layer does the embedding (assume layers are numbered 1, 2, 3, …)? Are there any nonlinearities between the first and last linear layers?

A. 2 layer neural network
B. Layer 1 does the embedding
C. No nonlinearities between the 2 layers

13. [4 points] BERT based sentence embeddings are trained using 2 different tasks. What are the 2 tasks?

A. Masked word prediction
B. {Next sentence, not next sentence} prediction

14. [4 points] Count based N gram language models are based on the chain rule of probability. What is the result of 1 step of the rule applied to $P(w_{n-1}, w_{n-2}, \ldots, w_1, w_0)$?

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

15. [4 points] Let E be a matrix that maps from 1 hot words w_t^T to dense vectors x_t^T and P be a matrix that maps from dense vectors y_t^T to 1 hot word estimates $w_t^{hat,T}$. Write out an equation for a 1 layer standard RNN based language model that connects x_t^T to y_t^T.

$y_t^T = f(x_t^T H + y_{t-1}^T G + v^T)$

Consider an encoder – attention – decoder style language translation network. Assume that the encoder uses a matrix E_1 to map from 1 hot words w_t in language 1 to dense vectors x_t and a multilayer transformer to map from dense vectors x_t to strong features y_t, t = 0, …, T – 1.

16. [2 points] What does the attention mechanism allow for the decoder?

    A weighted average, typically a function of the decoder state, of all of the input features to be used for generating output words in language 2

17. [2 points] Consider a sequential decoder where the output of the final RNN layer $z_{t'}$ is projected via matrix P_2 and a softmax to a pmf of potential words $p_{t'}$, t' = 0, …, T' – 1, in language 2. A method for improving translation results is to use the few most likely words from $p_{t'}$ in subsequent decoder steps along with an external language 2 model using this search technique

    Beam search

Speech [22 points]

18. [6 points] Speech pre processing typically converts the time domain speech waveform to a sequence of vectors with ~ frequency domain elements. List 3 different types of network structures (but not variations of the same network structure) that can be used to generate strong features from this type of pre processed speech.

    A. RNNs
    B. CNNs
    C. Self attention

19. [4 points] Consider a command recognition network with an encoder that maps from pre processed speech vectors to strong features. The decoder would typically map from strong features to this?

    A vector where each element corresponds to a command; depending on the system, maybe 1 additional element corresponding to not any of the known commands

20. [4 points] Consider a speaker identification network where the output of the encoder is strong feature vector x^T and the decoder is x^T H followed by arg max. How would you use imprinting to add a new speaker?

    Have the new speaker say multiple phrases and average together the resulting x^T to create a vector x_{new}^T

    Normalize and append x_{new} as a new column to H

21. [4 points] The RNN Transducer method of speech to text transduction includes 2 separate networks that are combined via a joint network. What is the purpose of each of these 2 nets?

A. An acoustic model network that maps from speech to a pmf of phonemes, graphemes or word pieces trained using the CTC error function to handle speech to label alignment

B. A language model network that maps from previous phonemes, graphemes or word pieces to a pmf of the next phoneme, grapheme or word piece

22. [4 points] Let {a, b, ..., z, ^} be the graphemes predicted by a speech to text transduction network that uses the CTC error function and decoding method with ^ being the special symbol. What word is decoded from the following network output: k, ^, i, i, i, t, t, t, ^, t, e, e, ^, n, ^, n

Step 1: k, ^, i, t, ^, t, e, ^, n, ^, n
Step 2: k, i, t, t, e, n, n
Answer: kittenn

Games [20 points]

23. [4 points] List 4 components of a Markov decision process that are used in reinforcement learning.

A. S is the set of all valid states, $\mu: \rightarrow S$ is the initial state function

B. A is the set of all valid actions

C. $R: S \times A \rightarrow \mathbb{R}$ is the reward function that maps states and actions to a reward

D. $P: S \times A \rightarrow S$ is the state transition function that maps states and actions to new states

24. [4 points] Consider a game that starts at discrete time t and continues forever. Write an equation for the infinite horizon discounted total reward where r_t is the reward at time t and γ is the discount factor.

$R_{t:\infty}(\tau, \gamma) = \Sigma_{i=t:\infty} \gamma^{i-t} r_i$, $\gamma \in (0, 1)$

25. [6 points] A deep Q network for playing Atari typically maps from this input to this output.

A. Input = a ~ 4 x height x width tensor of 4 video frames with some normalization

B. Output = a vector ~ 18 Q values, 1 for each joystick / button combination

26. [6 points] An AlphaZero style approach to playing Go or chess uses this type of search and these 2 types of networks which serve what purposes?

A. Monte Carlo tree search

B. With a policy network to bias the search (implicitly narrowing the breadth of the tree)

C. With a value network to predict outcomes (implicitly reducing the depth of the tree)