Computer Vision
Spring 2024

End Semester Solutions

Q1

Assigned TA: Sreenya Chitluri

Sample Answer

1. From the epipolar constraint $x_1^T F x_2 = 0$, we can isolate the equation of the epipolar line in the second image:

$$l_2 = F x_1$$

   This equation represents the line on which the corresponding point x_2 in the second image lies, given the point x_1 in the first image. Thus, given x_1, the point x_2 is constrained to lie on the line l_2, which is the image of the epipolar line through x_1 in the second image.

2. When the second camera undergoes a pure rotation about its center:

   - Epipolar lines in the second image rotate around the epipole of the first camera.
   - The epipole of the second camera moves in a circular path around its original position in the second image.

3. In stereo matching, besides the epipolar constraint, other constraints are used to determine correspondences between stereo image pairs. Here are a few:

   (a) Ordering Constraint: Ensures that correspondences between pixels in the left and right images maintain the same order along epipolar lines.

   (b) Uniqueness Constraint: Implies that each point in one image corresponds to at most one point in the other image.

   (c) Smoothness Constraint: Encourages smoothness in the disparity map, assuming that neighboring pixels typically have similar disparities.

   (d) Global Constraint: Considers the consistency of disparity values across the entire image to reduce errors caused by local ambiguities.

   (e) Photometric Constraint: Relates the intensities of corresponding pixels in the left and right images, assuming that corresponding points have similar intensities.

Rubric

1. Showing the equation of the line carries **1 mark**
2. Each point carries **0.5 marks**
3. Each constraint carries **0.33 marks**

Q2

Assigned TA: Mandyam Brunda

Sample Answer

1. 2a. The self-attention mechanism uses three matrices - query (Q), key (K), and value (V) - to help the system understand and process the relationships between words in a sentence. These three matrices serve distinct purposes:
   (a) Query (Q): This matrix represents the focus word for which the context is being determined. By transforming the word representation using the query matrix, the system generates a query vector that will be used to compare against other words in the sentence.
   (b) Key (K): The key matrix is used to create key vectors for all words in the sentence. These key vectors help the system measure the relevance or similarity between the focus word (using the query vector) and other words in the sentence. A higher similarity score between the query vector and a key vector indicates a stronger relationship between the corresponding words.
   (c) Value (V): The value matrix generates value vectors for all words in the sentence. These vectors hold the contextual information of each word. After calculating the similarity scores using query and key vectors, the system computes a weighted sum of the value vectors. The weights for each value vector are determined by the similarity scores, ensuring that the final contextual representation is influenced more by relevant words.

   Equation required: softmax(QK/sqrt(d))V
2. 2b) For better gradient flow, stability, reduce the effect of dimension all are accepted answers
3. 2c)
   (a) if considered same patch size by assuming the change in input image dimension in both cases, no effect on Transformer other parameters; yes on position encoding (learnable positional encoding)
   (b) if assumed patch size different and have linear layer that transforms this the flattened patch into fixed size then linear layer and postional embedding parameters (learnable positional encoding)
   (c) if considered no linear layer at the first then weights of q,k,v matrices also get affected along with positional encoding dimension

4. 2d)
   (a) CNN over ViT: Better capture local context
   (b) ViT over CNN: Better capture global context with self attention mechanism

Rubric

1. 2a) softmax required - 0.5 marks are deducted if softmax is mentioned Equation with explaining all terms - full marks
2. 2b) stating which parameters are affected - 0.5, which are not effected - 0.5 marks
3. 2c) if assumptions are not stated clearly marks are not given

Q3

Assigned TA:

Sample Answer

Rubric

Q4: Segmentation

- Working principle of MRF used in Classical FG/BG Segmentation Algorithm (e.g. GrabCut). Specifically, discuss the unary and pairwise energy potentials and how they influence the output segmentation mask.
- How is an FCN in Segmentation different from a CNN-based classifier? For simplicity, assume classification on K=20 classes and semantic segmentation has k=21 classes (20 + BG). Elaborate and explain the architectural modifications. Emphasise the output dimension of the 2 models.
- State 2 drawbacks of applying a model trained on 256x256 images to:
  1. 4k x 4k
  2. 32 x 32

Assigned TA: All

Sample Answer

1. MRFs are graphical models (RF with Markov property: values depend only on one neighbour). For a pixel represented as a node in MRF, *whether it belongs to FG/BG also depends on a grid of 4 neighbours.* Hence, MRF helps *capture the spatial coherence of the labels.* The energy function of MRF has two components:
   (a) Unary Potentials - gives pixel belief (likelihood) for fg/bg. (modelled using GMMs in GrabCut).

   (b) Pairwise Potential - accounts for similarity between labels and encourages spatial coherence.

2. After the FCN, an Avg Pool and Linear Classifier (Softmax) is present in the classifier, while a decoder block is present in a segmentation model. Replace the average pool and linear classifier by upsampling and classifier on each pixel; Vector of size $K = 20$ for the classifier. Output dim: $B \times 20$, and tensor of size $H \times W \times K = 21$ for segmentation. Output dim: $B \times H \times W \times 21$

3. 1 - too big an image doesn't allow a big enough receptive field; 2 - too much pooling on the small image.

Rubric

1. [1] For MRF working principle. [.5] for Explaining 2 potentials. [.5] for correct influence.

2. [.5] for FCN Difference. [1] for architectural modifications. [.5] for correct output dimensions. *Marks given for both /w Batch & /wo Batch*

3. [.5] for each correct drawbacks.

Q5: Object Detection [5]

Assigned TA: Mohd Hozaifa Khan

Sample Answer

1. 3 operations per box. Then, for a complete feature, we combine multiple filters by adding them with a -ve or +ve sign. So, the total number of operations is how the filter is designed. Pick one filter, and on combining, explain what happens.

Rubric

1. [2] Explanation for one filter [1]. 4 corner points. sum = d - (c + b) + a; # of operations [1 = **[.5]** for one filter + **[0.5]** for # operations combining filters]

2. [1] HOG and SIFT

   - Similarity: both are based on image gradients. [.5]
   - Difference: SIFT is a local descriptor, and HOG is a global part descriptor. HOG will describe the whole image; SIFT will describe points. [.5]

3. [2] RCNN → Fast RCNN → Faster RCNN.

   - Fast RCNN: 1 CNN evaluation instead of many for each proposal. RoI Pooling. [.5]. Diagram [.5]
   - Faster RCNN has RPN and proposals from the same network [.5]. Diagram [.5].

Q6: CLIP

Assigned TA: Mohd Hozaifa Khan

Sample Answer

Loss used in CLIP is InfoNCE (Noise-Contrastive Estimation), a contrastive loss used for self-supervised learning where negative samples are treated as noise.
Given a set of $X = x_1, x_2, ..., x_N$ of N random samples containing one positive sample (x_{pos}) from $p(x|c)$ and $N-1$ negative samples from the 'proposal' distribution $p(x)$, we optimize:

$$L_N = -\mathbb{E}_X \left[\log \frac{f(x,c)}{\sum_{x_j \in \mathbf{X}} f(x,c)} \right] \tag{1}$$

Here, $f(x,c)$ estimates the density ratio, which is:

$$f(x,c) = \exp\left(x \cdot c^T\right) \propto \frac{p(x|c)}{p(x)} \tag{2}$$

Note that the probability of detecting a positive sample correctly is $p(C = pos|X,c)$:

$$p(C = pos|X,c) = \frac{f(x,c)}{\sum_{x_j \in \mathbf{X}} f(x,c)}$$

At a closer look, $p(C = pos|X,c)$ seems like an output of softmax. If we substitute $p(C = pos|X,c)$ in 1, it becomes a Cross Entropy Loss over softmax outputs.
We consider 2 scenarios here: (1) $X =$ represents the encoding of Image, and C represents the encoding of text. (2) Vice versa. We then take a mean of both.

Explanation : It's a contrastive loss where we optimize the mutual information, $\mathcal{I}$, instead of distance (e.g., triplet or margin loss). We maximize $\mathcal{I}$ between positive pairs (x,c) and minimize $\mathcal{I}$ between negative pairs (x,c). Mutual information is approximated using f as mentioned above, and it also represents scoring functions. We compute $\mathcal{I}$ using the exponential of cosine similarity.

Rubric

- [1] InfoNCE Eqn [0.5] + Explanations [0.5]. *Marks have been deducted if the notations used are not explained. While both InfoNCE and Avg Cross Entropy (CE) have been considered, it is necessary to clarify the* CONTRASTIVE-NESS *in the context of CE Loss. No marks for margin or triplet loss*
- [1] Theoretically, *yes* [0.5]. It would require logits to be $-\infty$. So, practically, it cannot go to 0. [0.5]. *Mentioning mathematical properties like* $\log 1 = 0$ *won't be given marks. You've to give a logical interpretation.*
- [1] Yes, big batches are better because you have more negatives. Leads to better performance. [1] *No marks for* ***Yes*** *without correct explanations*

Q7

Assigned TA:

Sample Answer

Rubric

Q8

Assigned TA:

Sample Answer

Rubric