Self-supervised Learning [AI-835] || Mid-Term Exam

26 Sep 2023 || 2 pm – 3:30 pm

Course Instructor: V. Ramasubramanian || 9880942033

TA: Dhanya Eledath || 8105753702

Instructions

- This exam is a MCQ Exam for 90 minutes.
- The question paper has **20 questions** each carrying 1 mark – totaling 20 marks.
- Only ONE answer is correct for each question. No Negative-Marking.
- **Good Luck!!...**

1. ImageNet dataset can be represented as a 'source' dataset $D_s = \left\{x_i^{(s)},\ y_i^{(s)}\right\}_{i=1}^{M}$, where $x_i^{(s)}$ is the feature vector of an image (or simply the image itself, e.g. image of a 'cat') and $y_i^{(s)}$ is the corresponding human annotated ground truth label (e.g. the word 'cat') of the image $x_i^{(s)}$. Which of the following is correct with regard to Self-supervised Learning (SSL) algorithms that typically use ImageNet for pretraining to generate Foundation Models? → **CO1**

   a. $y_i^{(s)}$ is used in SSL algorithms' pretext tasks

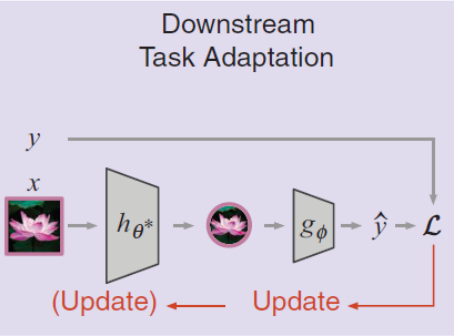
   b. $y_i^{(s)}$ is available, but not used in SSL algorithm's pretext tasks

      → **CORRECT ANSWER**

   c. $y_i^{(s)}$ is generated automatically by SSL algorithm's pretext task

   d. None of the above

2. In the adjoining figure, h_{θ^*} is the 'frozen' Foundation Model (or pretrained model) obtained in a pre-training stage of an SSL algorithm. $g_\emptyset$ is a classifier 'linear-probe' or 'linear-head' in a downstream adaptation setting with downstream 'task' supervised data $D_t = \left\{x_i^{(t)},\ y_i^{(t)}\right\}_{i=1}^{N}$ of N samples, each of which is indicated as $\{(x, y)\}$ in the figure. Which of the following best represents the learning of the 'linear probe' or 'linear head'? → **CO1**



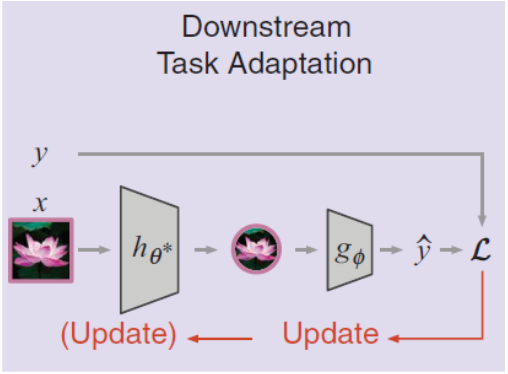
   a. $\emptyset^* = \arg\min_{\theta,\emptyset} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(g_\emptyset\left(h_\theta\left(x_i^{(t)}\right)\right), y_i^{(t)}\right)$

   b. $\emptyset^* = \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(g_\emptyset\left(h_\theta\left(x_i^{(t)}\right)\right), y_i^{(t)}\right)$

   c. $\emptyset^* = \arg\min_{\emptyset} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(g_\emptyset\left(h_{\theta^*}\left(x_i^{(t)}\right)\right), y_i^{(t)}\right)$ → **CORRECT ANSWER**

   d. $\emptyset^* = \arg\min_{\emptyset} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(g_\emptyset\left(h_{\theta^*}\left(x_i^{(t)}\right)\right)\right)$

3. In the adjoining figure, h_{θ^*} is the Foundation Model (or pretrained model) obtained in a pre-training stage of an SSL algorithm. $g_{\emptyset}$ is a classifier in a downstream adaptation setting with downstream 'task' supervised data $D_t = \left\{x_i^{(t)},\ y_i^{(t)}\right\}_{i=1}^{N}$ of N samples, each of which is indicated as $\{(x, y)\}$ in the figure. Which of the following best represents the learning of the entire downstream model under 'fine-tuning', with the understanding that h_θ is initialized with θ^* of the Foundation Model → **CO1**



a. $\theta^\dagger, \emptyset^\dagger = \arg\min_{\theta,\emptyset} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(g_\emptyset\left(h_\theta\left(x_i^{(t)}\right)\right), y_i^{(t)}\right)$

→ **CORRECT ANSWER**

b. $\emptyset^\dagger = \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(g_\emptyset\left(h_\theta\left(x_i^{(t)}\right)\right), y_i^{(t)}\right)$

c. $\emptyset^* = \arg\min_{\emptyset} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(g_\emptyset\left(h_{\theta^*}\left(x_i^{(t)}\right)\right), y_i^{(t)}\right)$

d. $\theta^\dagger, \emptyset^\dagger = \arg\min_{\theta,\emptyset} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(g_\emptyset\left(h_{\theta^*}\left(x_i^{(t)}\right)\right)\right)$

4. Consider a highly domain mismatched setting where the pre-training unsupervised data is ImageNet and the downstream supervised data is a small-size medical-imaging domain data (e.g. CT, MRI, Radiology scan images). Under such a condition, which of the following best reflects the overall SSL performance? → **CO2**

    a. SSL can be made to work in an effective manner without any downstream adaptation (learning only the linear-head or even fine-tuning the entire network)

    b. SSL is ineffective even with downstream adaptation of a linear-head

    c. SSL is effective only when pretrained with abundant medical imaging 'unsupervised' data instead of ImageNet data → **CORRECT ANSWER**

    d. None of the above

5. Consider a highly domain mismatched setting where the pre-training unsupervised data is ImageNet and the downstream supervised data is a small-size medical-imaging domain data (e.g. CT, MRI, Radiology scan images). Under such a condition, which of the following yields the best performance? → **CO2**

    a. Transferring lower layers of the Representation Learning function (model) h_{θ^*}

    → **CORRECT ANSWER**

    b. Transferring all the layers of the Representation Learning function (model) h_{θ^*}

    c. Transferring the top layers of the Representation Learning function (model) h_{θ^*}

    d. None of the above

6. Consider an SSL setting where 'dense labels' are available for the target task, i.e., the downstream supervised data $D_t = \left\{x_i^{(t)},\ y_i^{(t)}\right\}_{i=1}^{N}$ is large in size (i.e., large N). Which of the following best represents the overall performance. → **CO2**

    a. SSL is very helpful (with a large performance difference over the 'only' supervised counterpart)

    b. Directly training a supervised model on D_t gives the best performance

    → **CORRECT ANSWER**

    c. Nothing can be inferred

    d. None of the above

7. Consider a highly domain 'matched' setting where the pre-training unsupervised data is ImageNet and the downstream supervised data is 'naturally occurring everyday objects' image data whose classes are not present in ImageNet. Under such a condition, which of the following yields the best performance? → **CO2**

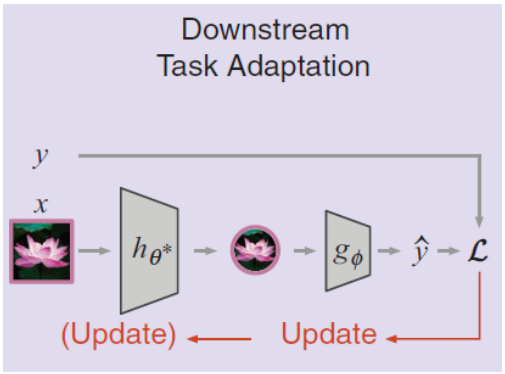
    a. SSL is most effective under this condition, with a downstream adaptation (learning the linear-probe or fine-tuning the entire model)
    → **CORRECT ANSWER**

    b. SSL is not effective even with a downstream adaptation

    c. SSL is effective only without a downstream adaptation

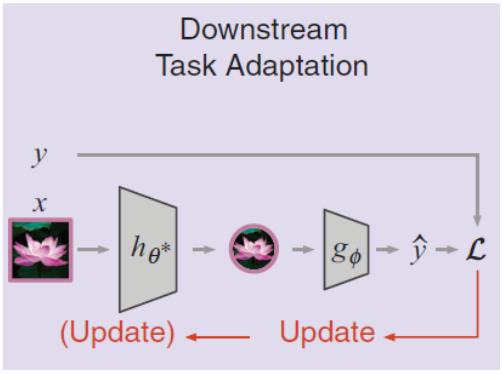
    d. None of the above

8. In the adjoining figure, h_{θ^*} is the Foundation Model (or pre-trained model) obtained in a pre-training stage of an SSL algorithm. $g_{\emptyset}$ is a classifier in a downstream adaptation setting with downstream 'task' supervised data $D_t = \left\{x_i^{(t)},\ y_i^{(t)}\right\}_{i=1}^{N}$ of N samples, each of which is indicated as $\{(x, y)\}$ in the figure. Which of the following best represents the condition for an "SSL design choice" in favor of Linear Probing (h_{θ^*} is frozen and the linear-head $g_{\emptyset}$ is optimized on D_t). → **CO2**



   a. When the source and target domains mismatch severely
   b. When N is large and there is a problem of overfitting $g_{\emptyset}$
   c. When performance comparisons need to be made in a fair benchmarking of a wide variety of h_{θ^*}, e.g., as may be obtained from different pretext task and pretraining methods → **CORRECT ANSWER**
   d. None of the above

9. In the adjoining figure, h_{θ^*} is the Foundation Model (or pre-trained model) obtained in a pre-training stage of an SSL algorithm. $g_{\emptyset}$ is a classifier in a downstream adaptation setting with downstream 'task' supervised data $D_t = \left\{x_i^{(t)},\ y_i^{(t)}\right\}_{i=1}^{N}$ of N samples, each of which is indicated as $\{(x, y)\}$ in the figure. Which of the following best represents the condition for an "SSL design choice" in favor of Fine-tuning (h_θ is not frozen, but initialized with θ^* and optimized on D_t, together with the classifier $g_{\emptyset}$). → **CO2**



   a. When the source and target domains mismatch severely

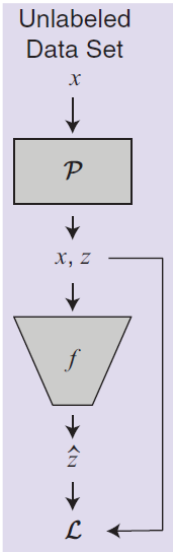
      → **CORRECT ANSWER**

   b. When N is small

   c. No design choice can be made or is possible in favor of Fine-tuning

   d. None of the above

10. The adjoining figure shows a generic SSL "pre-training" pipeline, with unlabeled data x, pretext task process $\mathcal{P}$ which generates a pseudo-label z and a pretraining function (model) f which yields a prediction $\hat{z}$ as close as possible to z, and whose parameters θ are learnt by minimizing the loss $\mathcal{L}(z, \hat{z})$. Let $\mathcal{L}(z, \hat{z})$ be alternatively represented as $l_{pre}(x; \theta)$ to capture its dependence on x and θ. Let p_{pre} be the distribution of the raw data x, and $\hat{p}_{pre}$ (where, $\hat{p}_{pre} \sim p_{pre}$) is the empirical distribution over a large number of independent samples from p_{pre}. Which of the following best represents the learning of the foundation model's parameters $\hat{\theta}_{FM}$? → **CO2**



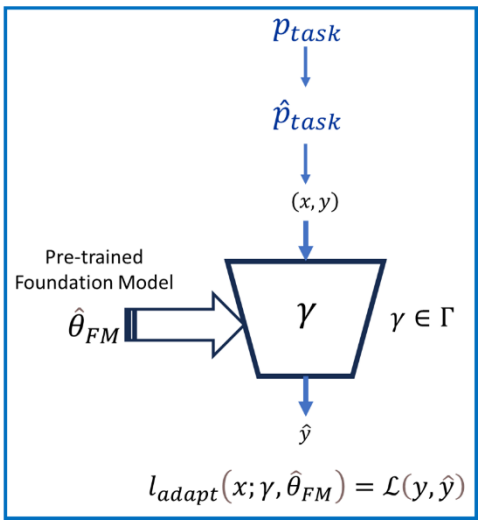
a. $\hat{\theta}_{FM} = \arg\max_{\theta \in \Theta} \mathbb{E}_{x \sim \hat{p}_{pre}} \left[l_{pre}(x; \theta) \right]$

b. $\hat{\theta}_{FM} = \arg\min_{\theta \in \Theta} \mathbb{E}_{x \sim p_{pre}} \left[l_{pre}(x; \theta) \right]$

c. $\hat{\theta}_{FM} = \arg\max_{\theta \in \Theta} \mathbb{E}_{x \sim p_{pre}} \left[l_{pre}(x; \theta) \right]$

d. $\hat{\theta}_{FM} = \arg\min_{\theta \in \Theta} \mathbb{E}_{x \sim \hat{p}_{pre}} \left[l_{pre}(x; \theta) \right]$ → **CORRECT ANSWER**

11. The adjoining figure shows the pipeline for a downstream task adaptation. p_{task} represents the distribution of the downstream task input data, and $\hat{p}_{task}$ is the empirical distribution of 'few' samples sampled from p_{task} (i.e., the input $x \sim \hat{p}_{task} \sim p_{task}$). Supervised pair (x, y) is given as input to the learnable downstream model $\gamma \in \Gamma$. Foundation Model $\hat{\theta}_{FM}$ is transferred from pre-training stage of SSL, and used as the representation learning function of the downstream task. γ could be the linear-head in a linear-probing setting (with the Foundation Model $\hat{\theta}_{FM}$ frozen) or could be the entire "Foundation Model + Classifier" in a fine-tuning setting. $l_{adapt}(x; \gamma, \hat{\theta}_{FM}) = \mathcal{L}(y, \hat{y})$ is the per sample loss. Which of the following represents the learning of the optimal downstream model denoted as $\gamma_{task}(\hat{\theta}_{FM})$?. → **CO2**



a. $\gamma_{task}(\hat{\theta}_{FM}) = \arg\min_{\gamma \in \Gamma} \mathbb{E}_{x \sim \hat{p}_{task}} \left[l_{adapt}(x; \gamma, \hat{\theta}_{FM}) \right]$

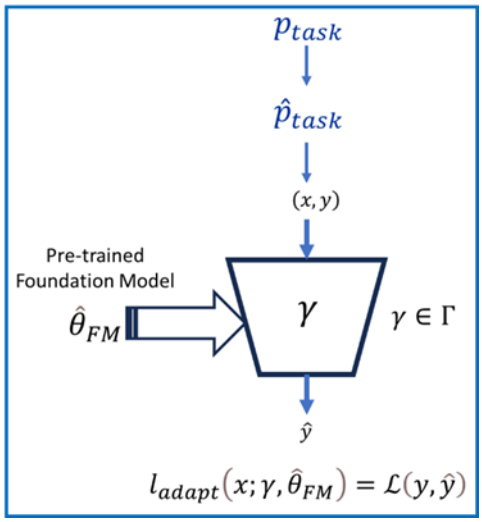
→ **CORRECT ANSWER**

b. $\gamma_{task}(\hat{\theta}_{FM}) = \arg\min_{\gamma \in \Gamma} \mathbb{E}_{x \sim p_{task}} \left[l_{adapt}(x; \gamma, \hat{\theta}_{FM}) \right]$

c. $\gamma_{task}(\hat{\theta}_{FM}) = \arg\max_{\gamma \in \Gamma} \mathbb{E}_{x \sim \hat{p}_{task}} \left[l_{adapt}(x; \gamma, \hat{\theta}_{FM}) \right]$

d. $\gamma_{task}(\hat{\theta}_{FM}) = \arg\max_{\theta \in \Theta} \mathbb{E}_{x \sim \hat{p}_{task}} \left[l_{adapt}(x; \gamma, \hat{\theta}_{FM}) \right]$

12. The adjoining figure shows the pipeline for a downstream task adaptation. p_{task} represents the distribution of the downstream task input data, and $\hat{p}_{task}$ is the empirical distribution of 'few' samples sampled from p_{task} (i.e., the input $x \sim \hat{p}_{task} \sim p_{task}$). Supervised pair (x, y) is given as input to the learnable downstream model $\gamma \in \Gamma$. Foundation Model $\hat{\theta}_{FM}$ is transferred from pre-training stage of SSL, and used as the representation learning function of the downstream task. γ could be the linear-head in a linear-probing setting (with the Foundation Model $\hat{\theta}_{FM}$ frozen) or could be the entire "Foundation Model + Classifier" in a fine-tuning setting. The optimal downstream model is denoted as $\gamma_{task}(\hat{\theta}_{FM})$. $l_{adapt}(x; \gamma, \hat{\theta}_{FM}) = \mathcal{L}(y, \hat{y})$ is the per sample loss. Which of the following represents the effective performance of the adapted downstream model $\gamma_{task}(\hat{\theta}_{FM})$ during inference, i.e., the population loss on unseen test data?. → **CO2**



a. $\mathcal{L}_{adapt}(\gamma_{task}, \hat{\theta}_{FM}) = [l_{adapt}(x; \gamma, \hat{\theta}_{FM})]$

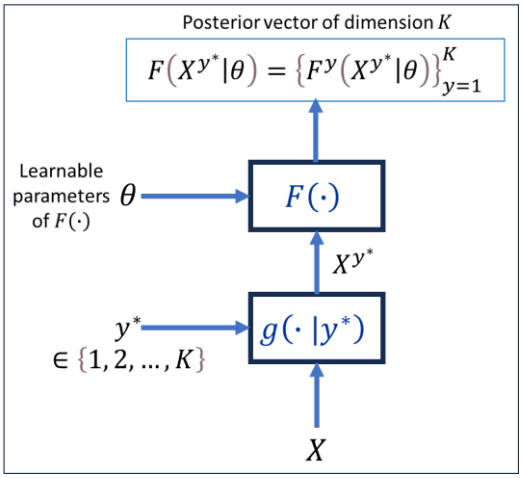
b. $\mathcal{L}_{adapt}(\gamma_{task}, \hat{\theta}_{FM}) = \mathbb{E}_{x \sim \hat{p}_{task}}[l_{adapt}(x; \gamma, \hat{\theta}_{FM})]$

c. $\mathcal{L}_{adapt}(\gamma_{task}, \hat{\theta}_{FM}) = \mathbb{E}_{x \sim p_{task}}[l_{adapt}(x; \gamma, \hat{\theta}_{FM})]$

→ **CORRECT ANSWER**

d. $\mathcal{L}_{adapt}(\gamma_{task}, \hat{\theta}_{FM}) = \mathbb{E}_{x \sim \hat{p}_{task}}[-l_{adapt}(x; \gamma, \hat{\theta}_{FM})]$

13. The adjoining figure shows a pipeline of SSL pre-training based on the pretext task of "Predicting Rotations". X is an input image which is subject to a geometric transform (rotation) $g(\cdot\,|y^*)$ yielding the rotated image $X^{y^*} = g(X|y^*)$, for a given rotation $y^* \in \{1, 2, \dots, K\}$ out of 1-of-K rotations, given in general by $g(\cdot\,|y) = Rot(X, (y-1) \times 90^\circ)$. $F(\cdot)$ is a learnable function (model) with learnable parameters θ and yields a posterior vector (set of probabilities) $F(X^{y^*}|\theta) = \{F^y(X^{y^*}|\theta)\}_{y=1}^{K}$, where $F^y(X^{y^*}|\theta)$ is the posterior probability of rotation class y given input 'rotated' image X^{y^*} and parameter θ. Then, which of the following represents the negative log-likelihood "loss", as derived from cross-entropy loss between the posterior vector $F(X^{y^*}|\theta)$ and the 1-hot encoding vector of the pseudo-label y^*? → **CO3**



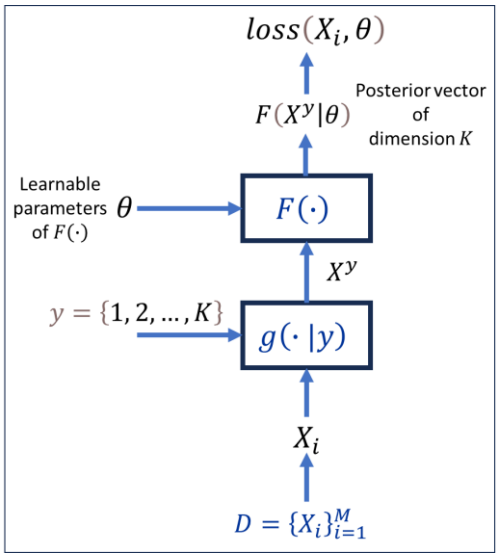
a. $-\log\left(F^{y^*}(g(X|y^*)|\theta)\right)$ → **CORRECT ANSWER**

b. $\log\left(F^{y}(g(X|y^*)|\theta)\right)$

c. $-\log\left(F^{y}(g(X|y)|\theta)\right)$

d. $F^{y^*}(g(X|y^*)|\theta)$

14. The adjoining figure shows a pipeline of SSL pre-training based on the pretext task of "Predicting Rotations". X is an input image which is subject to a geometric transform (rotation) $g(\cdot\,|y)$ yielding the rotated image $X^y = g(X|y)$, for a rotation $y = 1, \ldots, K$, i.e., each of K rotations, given in general by $g(\cdot\,|y) = Rot(X, (y-1) \times 90^\circ)$. $F(\cdot)$ is a learnable function (typically, a ConvNet model) with learnable parameters θ. The pretraining with this pretext task is done with an unsupervised (unlabeled) dataset $D = \{X_i\}_{i=1}^M$ of M samples (e.g., raw images). The optimal ConvNet model parameters are obtained as



$$\theta^* = \arg\min_\theta \frac{1}{M} \sum_{i=1}^{M} loss(X_i, \theta)$$

Let $F^y(X^y|\theta)$ is the posterior probability value (scalar) at the y^{th} place (of rotation class y) in the posterior vector generated by the ConvNet $F(\cdot)$ for an input X^y. Then which of the following represents the term "$loss(X_i, \theta)$" in the above optimization equation? → **CO3**

a. $loss(X_i, \theta) = \frac{1}{K} \sum_{y=1}^{K} \log\left(F^y(g(X|y)|\theta)\right)$

b. $loss(X_i, \theta) = -\log\left(F^y(g(X_i|y)|\theta)\right)$

c. $loss(X_i, \theta) = -\frac{1}{K} \sum_{y=1}^{K} \log\left(F^y(g(X_i|y)|\theta)\right)$ → **CORRECT ANSWER**

d. $loss(X_i, \theta) = -\frac{1}{K} \sum_{y=1}^{K} \log\left(F^i(g(X_i|y)|\theta)\right)$

15. A and B are two inputs (e.g. raw images) to a Siamese Network $f(\cdot)$ which embeds A as $f(A)$ and B as $f(B)$. (A, B) could be a 'Positive Pair' (both A and B belong to the same class) or a Negative Pair (A and B belong to different classes), specified by $y^* = 1$ and $y^* = 0$ respectively. Training the Siamese Network involves optimizing its parameters to minimize the Contrastive Loss. If m is the margin (to bring the samples further apart up to the margin, under the Hinge Loss for Negative-Pairs), which of this represents the Contrastive Loss $\mathcal{L}(A, B)$? → **CO3**

a. $y^* \|f(A) - f(B)\|^2 + (1 - y^*) \min(0, m^2 - \|f(A) - f(B)\|^2)$

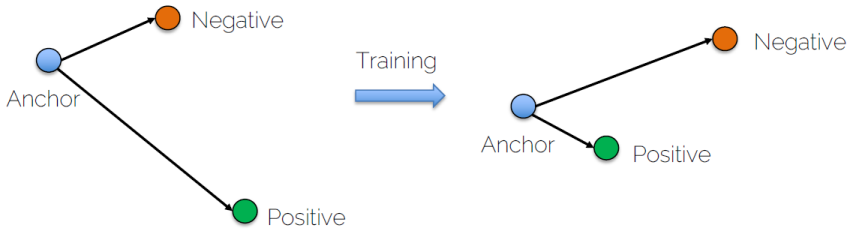
b. $(1 - y^*) \|f(A) - f(B)\|^2 + y^* \max(0, m^2 - \|f(A) - f(B)\|^2)$

c. $y^* \|f(A) + f(B)\|^2 + (1 - y^*) \max(0, m^2 - \|f(A) + f(B)\|^2)$

d. $y^* \|f(A) - f(B)\|^2 + (1 - y^*) \max(0, m^2 - \|f(A) - f(B)\|^2)$

→ **CORRECT ANSWER**

16. A, P and N are inputs (e.g., raw images, termed 'Anchor', 'Positive' and 'Negative' samples) to a Triplet Network $f(\cdot)$ which embeds the input samples as $f(A), f(P)$ and $f(N)$ respectively. (A, P) is a 'Positive Pair' (A and P belong to the same class) and (A, N) is a 'Negative Pair' (A and N belong to different classes). The Triplet Network is trained with a "Triplet Loss" which serves to rearrange the samples (in the left side of the adjoining figure) into the samples in the right side of the figure 'after' training (i.e., after the training, the positive examples will be closer to the anchor while the negative examples will be farther from it). Which of the following represents the Triplet Loss $\mathcal{L}(A, P, N)$ with a positive margin m?
→ **CO3**



a. $\max(0, \|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + m)$

→ **CORRECT ANSWER**

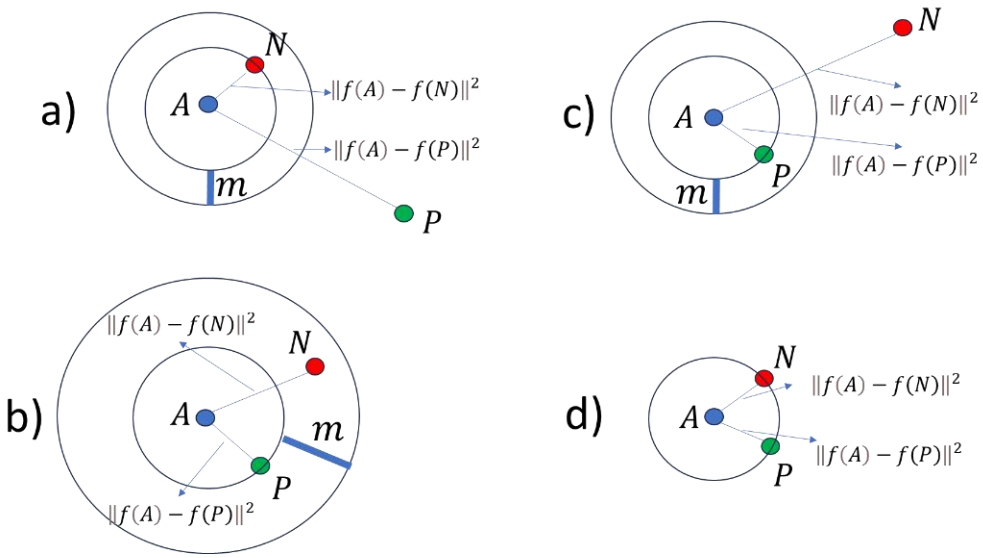
b. $\min(0, \|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + m)$

c. $\max(m, \|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + m)$

d. $\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + m$

17. A, P and N are inputs (e.g., raw images, termed 'Anchor', 'Positive' and 'Negative' samples) to a Triplet Network $f(\cdot)$ which embeds the input samples as $f(A), f(P)$ and $f(N)$ respectively. (A, P) is a 'Positive Pair' (A and P belong to the same class) and (A, N) is a 'Negative Pair' (A and N belong to different classes). The Triplet Network is trained with a "Triplet Loss". Which one of the following 4 figures represents the terminating condition in the training with the Triplet Loss $\mathcal{L}(A, P, N)$ with a positive margin m, i.e., once this configuration is reached, the training can stop?

→ CO3



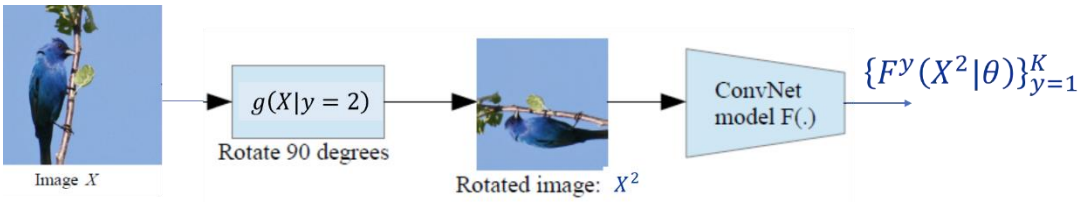
a. Figure (a)

b. Figure (b)

c. Figure (c) **→ CORRECT ANSWER**

d. Figure (d)

18. The adjoining figure shows the pipeline for the pretext task of "Predicting Rotations". Let $F^y(X^y|\theta)$ is the posterior probability value (scalar) at the y^{th} place (of rotation class y) in the posterior vector generated by the ConvNet $F(\cdot)$ for an input X^y. For the figure shown (with rotation of 90°, corresponding to $y = 2$), which of the following correctly represents the position of the peak posterior value, after the ConvNet has been pretrained optimally?. → **CO3**



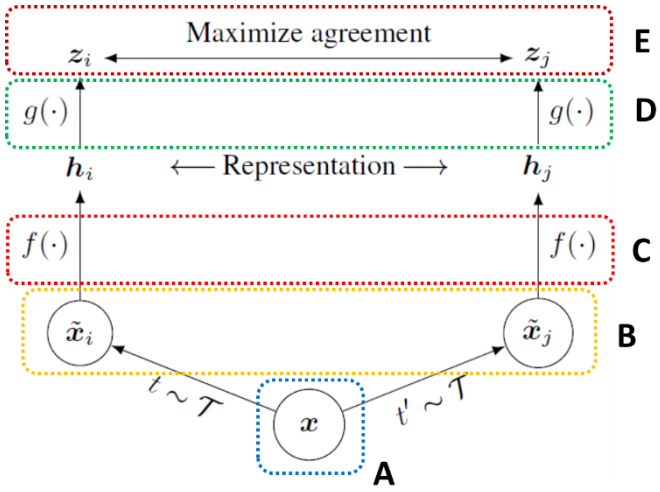
a. 1st component, i.e., $F^1(X^2|\theta)$

b. 2nd component, i.e., $F^2(X^2|\theta)$ → **CORRECT ANSWER**

c. 3rd component, i.e., $F^3(X^2|\theta)$

d. Nothing can be inferred

19. The adjoining figure shows the flow-diagram of SimCLR. Which one of the following annotations best describes the various modules/steps marked A, B, C, D and E in the figure? → **CO3**



a. **A** – embedding obtained from ResNet encoder, **B** – set of embeddings derived from A via stochastic augmentations $t, t' \in \mathcal{T}$ (a set of transformations), **C** – Siamese network (ResNet Encoder), **D** – Projection-Head, typically an MLP, **E** – Contrastive Loss

b. **A** – input raw image, from a randomly sampled mini-batch of size N, **B** – set of $2N$ samples derived from A via stochastic augmentations $t, t' \in \mathcal{T}$ (a set of transformations), **C** – Siamese network (ResNet Encoder), **D** – Projection-Head, typically an MLP, **E** – Contrastive Loss [NT-Xent (Normalized Temperature-Scaled Cross-Entropy) Loss] → **CORRECT ANSWER**

c. **A** – embedding obtained from ResNet encoder, **B** – set of $2N$ embeddings derived from A via stochastic augmentations $t, t' \in \mathcal{T}$ (a set of transformations), **C** – Siamese network (ResNet Encoder), **D** – Sigmoid / RELU non-linear activation, **E** – Triplet Loss

d. **A** – embedding obtained from ResNet encoder, **B** – set of $2N$ embeddings derived from random perturbations of A, **C** – Triplet network, **D** – Sigmoid / RELU non-linear activation, **E** – Triplet Loss

20. The following figure is the pseudo-code of the SimCLR Learning algorithm. Following the notations given in this code, indices $i, j \in \{1, 2, \ldots, 2N\}$ are pointers to the $2N$ final embeddings generated by the Encoder $f(\cdot)$ (ResNet) and Projection Head $g(\cdot)$ pipeline. The blue-arrow in the figure points to $l(i,j)$ which is the loss between any i-th sample and j-th sample in these $2N$ embeddings, where (i,j) could be positive-pairs or negative-pairs in general. Which of the following correctly represents $l(i,j)$ – in the place marked by the red-rectangle? →**CO3**

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, temperature τ, form of $f, g, \mathcal{T}$.
for sampled mini-batch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^{\top} \boldsymbol{z}_j / (\tau \|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i,j)$ **as** [redacted]
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network f

a. $l(i,j) = \log \frac{exp(s_{i,j})}{\sum_{\substack{m=1 \\ m \neq i}}^{2N} exp(s_{i,m})}$

b. $l(i,j) = -\log \frac{exp(s_{i,j})}{\sum_{m=1}^{2N} exp(s_{i,m})}$

c. $l(i,j) = -\log \frac{exp(s_{i,j})}{\sum_{\substack{m=1 \\ m \neq i}}^{2N} exp(s_{i,m})}$ → **CORRECT ANSWER**

d. $l(i,j) = \log \frac{exp(-s_{i,j})}{\sum_{\substack{m=1 \\ m \neq i}}^{N} exp(s_{i,m})}$