Advanced Topics in Deep Learning

CS 672

Final Exam

Duration: (At least) 3 hours **Marks: 160**

Instructions:

1. This exam is open note, open book, and open internet.
2. NO communication/ consultation among yourselves during the exam is allowed.
3. You are NOT allowed to leave your specified seat during the exam for more than 5 minutes at a stretch.
4. NO one is allowed to leave the exam hall before the completion of 3 hours.
5. Parts of a single question MUST be together in the answer sheet.

1. Loss Functions and ArcFace:

A) Suppose you have trained an ArcFace model `(M)` on `'C'` number of different identities. Now, you have a face image `(I)` with its claimed identity `(id)`. Explain how one can verify this claim using model `(M)`. Describe the verification process by referring to the following figure (Figure 2 in the given paper):

ArcFace_CVPR_2019.pdf

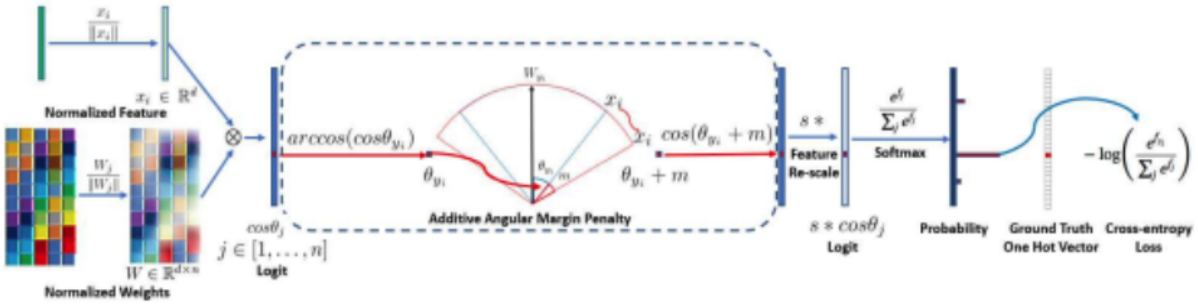
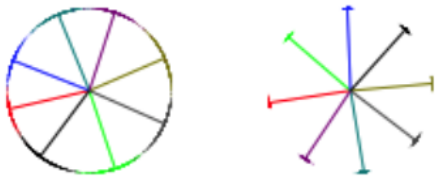


Figure 2. Training a DCNN for face recognition supervised by the ArcFace loss. Based on the feature x_i and weight W normalisation, we get the $\cos\theta_j$ (logit) for each class as $W_j^T x_i$. We calculate the $arccos\theta_{y_i}$ and get the angle between the feature x_i and the ground truth weight W_{y_i}. In fact, W_j provides a kind of centre for each class. Then, we add an angular margin penalty m on the target (ground truth) angle θ_{y_i}. After that, we calculate $\cos(\theta_{y_i} + m)$ and multiply all logits by the feature scale s. The logits then go through the softmax function and contribute to the cross entropy loss.

B) Explain the loss function defined in Eq.1 and Eq. 2 using Figure 3 of the above-mentioned paper (the following figure). Draw a similar plot for the hyper-sphere loss function given in Eq. 2?



(a) Softmax (b) ArcFace

Figure 3. Toy examples under the softmax and ArcFace loss on 8 identities with 2D features. Dots indicate samples and lines refer to the centre direction of each identity. Based on the feature normalisation, all face features are pushed to the arc space with a fixed radius. The geodesic distance gap between closest classes becomes evident as the additive angular margin penalty is incorporated.

C) Explain how the intra and inter loss functions, defined in Eq. 5 and Eq. 6 (given below) are decreasing the *"angle/arc between the sample and the ground truth centre"* and increasing the "*angle/arc between different centres"*, respectively. State the significance of the normalization factor in Eq. 6.

Intra-Loss is designed to improve the intra-class compactness by decreasing the angle/arc between the sample and the ground truth centre.

$$L_5 = L_2 + \frac{1}{\pi N}\sum_{i=1}^{N}\theta_{y_i}. \quad (5)$$

Inter-Loss targets at enhancing inter-class discrepancy by increasing the angle/arc between different centres.

$$L_6 = L_2 - \frac{1}{\pi N\,(n-1)}\sum_{i=1}^{N}\sum_{j=1, j\neq y_i}^{n}\arccos(W_{y_i}^{T}W_j). \quad (6)$$

[**Marks:** 9+(3*4)+9 = **30**]

2. Inversion for Generation:

A.

i) The following figure describes a method of GAN inversion. Explain the concept of GAN inversion concisely.

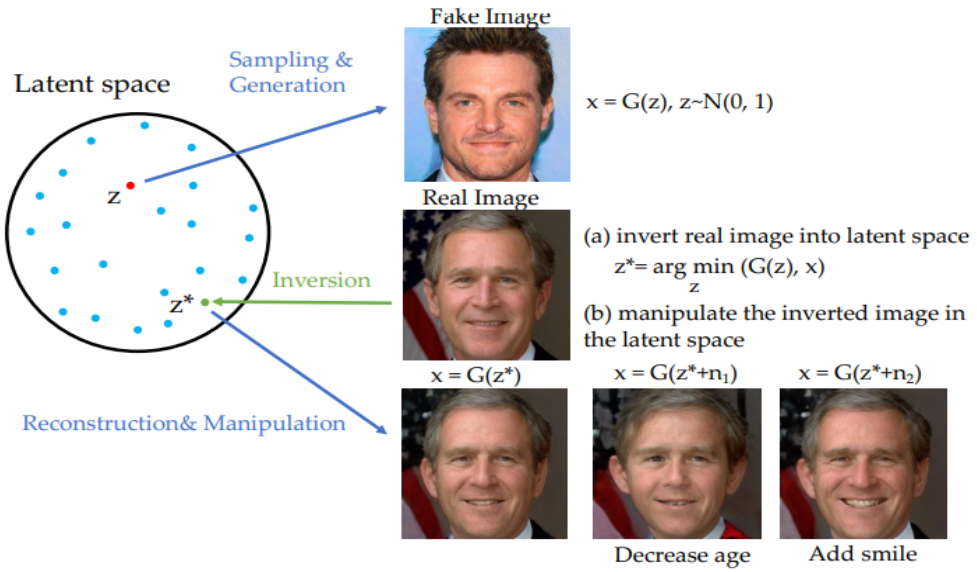


Fig. 1. Illustration of GAN inversion. Different from the conventional sampling and generation process using trained generator G, GAN inversion maps a given real image x to the latent space and obtains the latent code $\mathbf{z}^*$. The reconstructed image x^* is then obtained by $x^* = G(\mathbf{z}^*)$. By varying the latent code $\mathbf{z}^*$ in different interpretable directions *e.g.*, $\mathbf{z}^* + \mathbf{n_1}$ and $\mathbf{z}^* + \mathbf{n_2}$ where $\mathbf{n_1}$ and $\mathbf{n_2}$ model the age and smile in the latent space respectively, we can edit the corresponding attribute of the real image.

ii) The survey GANInversion_Survey.pdf discussed three types of GAN inversion methods. Explain those methods in your words with proper notations and diagrams by referring to the following figure from the paper.

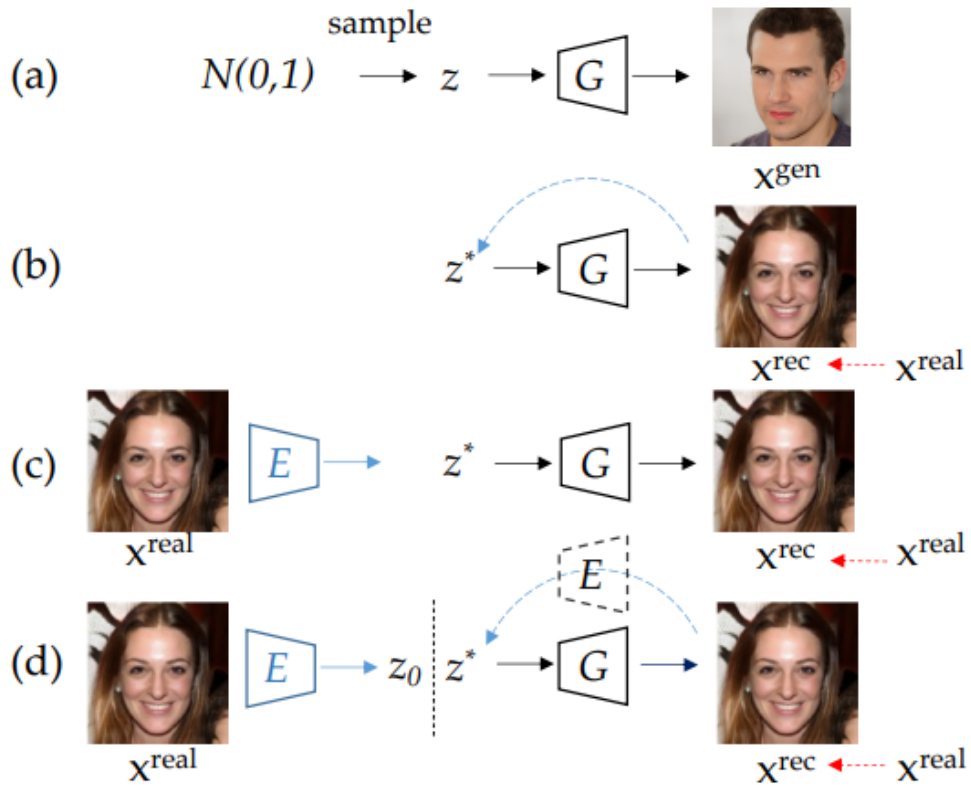
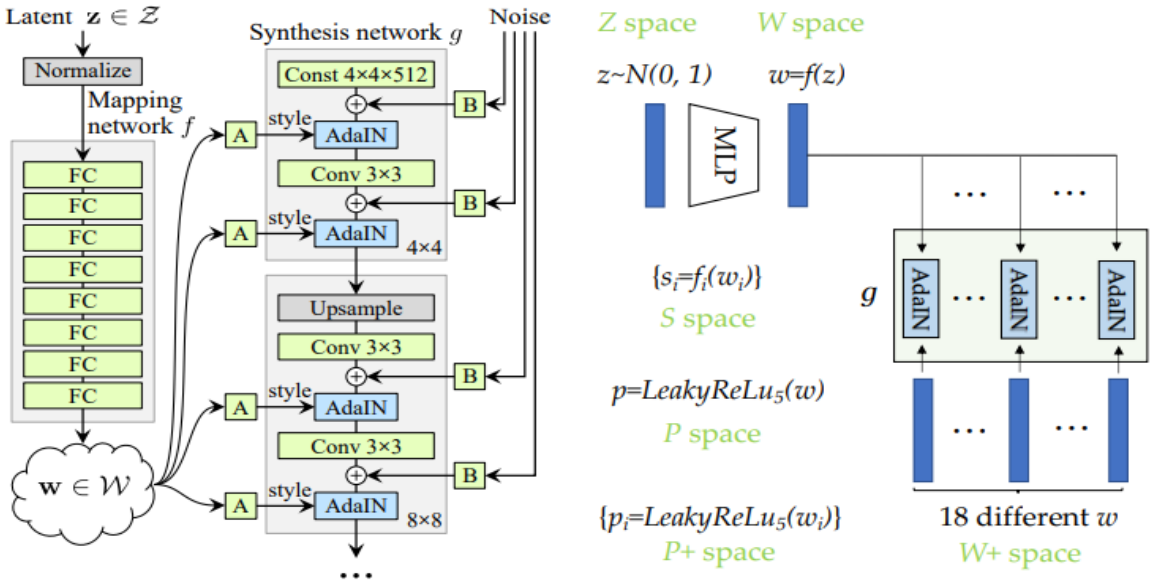


Fig. 3. Illustration of GAN Inversion Methods. (a) Given a well-trained GAN model G, photo-realistic images $\mathbf{x}^{\text{gen}}$ can be generated from randomly sampled latent vectors $\mathbf{z}$. GAN inversion aims to obtain the latent code $\mathbf{z}^*$ for a given image $\mathbf{x}^{\text{real}}$. A **learning-based** inversion method aims to learn an encoder network to map an image into the latent space such that the reconstructed image based on the latent code look as similar to the original one as possible. An **optimization-based** inversion approach directly solves the objective function through back-propagation to find a latent code that minimizes pixel-wise reconstruction loss. A **hybrid** approach first uses an encoder to generate initial latent code and then refines it with an optimization algorithm. Depicted by the dotted E, the well-trained encoder is included in [19] as a regularizer for optimization. Blue blocks represent trainable or iterative modules, and red dashed arrows indicate the supervisions.

iii) Following is the architecture of the style-based generator. Describe its inversion method. Explain what are the P+ space and W+ space.



(a) Style-based generator (b) Latent spaces of StyleGAN

Fig. 2. (a) Architecture of the style-based generator. (b) The latent spaces from which the inversion methods are constructed. The synthesis network g and AdaIN in (b) are the same as in (a).

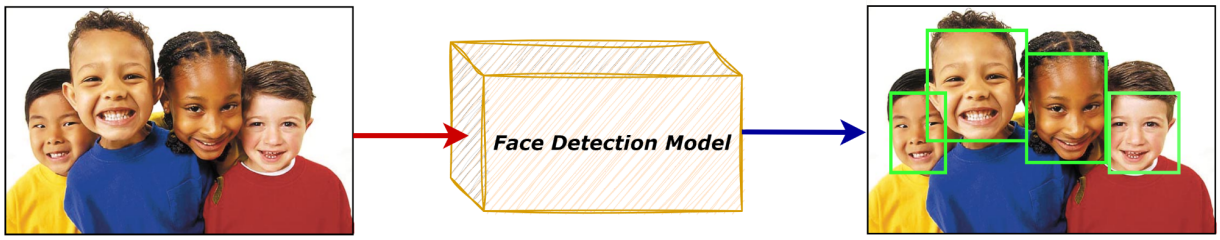
iv) How does GANInversion_Image2StyleGAN.pdf invert GAN (explain the procedure using Algorithm 1)?

B. Why does the *Edit-friendly Diffusion model* editDDPM.pdf claim to be the inversion of the *Denoising Diffusion Probabilistic Model* DDPM_Ho.pdf? Explain by comparing both algorithms. How one can modify/edit/change the colour of some selected parts of an image?

C. Explain the ArcFace inversion by referring to the following paper:
ArcFace_inversion.pdf

[**Marks:** (5x4)+10+10 = **40**]

3. Face Detection using Reinforcement Learning:



The output of a face detection model is one/ more bounding boxes around the faces, present in the image, as depicted above. An efficient model not only detects all the faces (be these of human or animal) in the image but also predicts the best bounding box coordinates that enclose the face regions only. Suppose you are required to build a face detection model using reinforcement learning (DRL algorithms and framework). Assume that each input image contains only a single face.

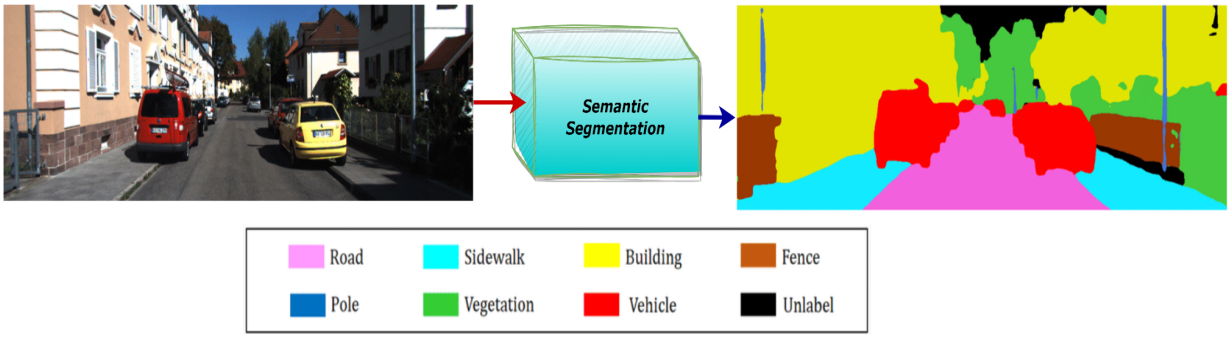
Write an algorithm along with its training methodology with thorough design details of the following entities:

Environment:

i) The state space,
ii) The action space,
iii) The reward,
iv) The agent (which algo and why),
v) Other necessary supporting elements of the algorithm (RB, Episodes, timesteps, data generation, gamma).

[**Marks:** 10+(5*4) = **30**]

4. Segmentation using Diffusion:



The above figure depicts the method of semantic segmentation, which divides an image into several meaningful parts by categorizing each pixel according to some predefined labels. Devise a method for such image segmentation using a probabilistic diffusion model (DDPM_Ho.pdf). Your design should thoroughly describe and justify the followings:

- The forward process and its hyper-parameters
- The training algorithm with its loss function
- The inference/sampling method

[**Marks:** 10+10+10 = **30**]

5. Transformer:

A) Describe the *Probe-sparse self-attention* mechanism in Informer.pdf and *self* and *cross attention* mechanism of Block-Recurrent Transformers.pdf.

B) The above-mentioned papers claim that they improved the time and space complexity of the vanilla transformer Transformer.pdf. Compute the time and space complexity of the `Informer` and `Block-Recurrent Transformer` to find whether their claim is valid or not.

C) Transformers take all the parts/tokens of sequential data together (all at once) as the input and hence produce position invariant representation. But their positional information plays a critical role in any time series prediction and should not be lost from the data. Hence, the input tokens of such models should be accompanied by a suitable positional encoding. Given the following three choices of positional encoding, which will you choose? Compare the following choices and select one with a proper justification.

- **Linear:** Positions are encoded by an increasing sequence of integers, e.g. 1, 2, 3, …
- **Range:** Positions are encoded by the real values within a range, e.g. [0,1]
- **Sinusoidal:** The position of each part is a sinusoidal function (choose a suitable function and explain the answer with respect to this).

D) As we all know Natural Language Processing (NLP) tasks need word embeddings as their input, which is well understood.

i) How is one-hot-encoded word embedding done?

ii) Why do we prefer the *Word2Vec* method over one-hot-encoded word embedding?

[**Marks:** (6+6)+(3+3)+(3*3)+(1+2) = **30**]