Due Date: Dec 8th (23:00 ET), 2023

Question 1 (5-5-5-5)**.** (**Autoregressive Models**)

One way to enforce autoregressive conditioning is via masking the weight parameters. [1] Consider a two-hidden-layer convolutional neural network without kernel flipping, with kernel size 3×3 and padding size 1 on each border (so that an input feature map of size 5×5 is convolved into a 5×5 output). Define mask of type A and mask of type B as

$$(\boldsymbol{M}^A)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j < 2 \\ 0 & \text{elsewhere} \end{cases} \qquad (\boldsymbol{M}^B)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

where the index starts from 1. Masking is achieved by multiplying the kernel with the binary mask (elementwise). Specify the receptive field of the output pixel that corresponds to the third row and the third column (index 33 of Figure 1 (Left)) in each of the following 4 cases:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 1 – (Left) 5×5 convolutional feature map. (Right) Template answer.

1. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
2. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for the second layer.
3. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
4. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer.

Your answer should look like Figure 1 (Right).

1. An example of this is the use of masking in the Transformer architecture.

Answer 1. We have:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 2 – Mask A

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 3 – Mask B

(a) If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

(b) If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for the second layer:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

(c) If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for the second layer:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

(d) If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

Question 2 (5-5)**.** (**Normalizing Flows**) In this question, we study some properties of normalizing flows. Let $X \sim P_X$ and $U \sim P_U$ be, respectively, the distribution of the data and a base distribution (e.g. an isotropic gaussian). We define a normalizing flow as $F : \mathcal{U} \to \mathcal{X}$ parametrized by $\boldsymbol{\theta}$. Starting with P_U and then applying F will induce a new distribution $P_{F(U)}$ (used to match P_X). Since normalizing flows are invertible, we can also consider the distribution $P_{F^{-1}(X)}$.

However, some flows, like planar flows, are not easily invertible in practice. If we use P_U as the base distribution, we can only sample from the flow but not evaluate the likelihood. Alternatively, if we use P_X as the base distribution, we can evaluate the likelihood, but we will not be able to sample.

2.1 Show that $D_{KL}[P_X||P_{F(U)}] = D_{KL}[P_{F^{-1}(X)}||P_U]$. In other words, the forward KL divergence between the data distribution and its approximation can be expressed as the reverse KL divergence between the base distribution and its approximation.

2.2 Suppose two scenarios: 1) you don't have samples from $p_X(\boldsymbol{x})$, but you can evaluate $p_X(\boldsymbol{x})$, 2) you have samples from $p_X(\boldsymbol{x})$, but you cannot evaluate $p_X(\boldsymbol{x})$. For each scenario, specify if you would use the forward KL divergence $D_{KL}[P_X||P_{F(U)}]$ or the reverse KL divergence $D_{KL}[P_{F(U)}||P_X]$ as the objective to optimize. Justify your answer.

Answer 2. 1. For the forward we have:

$$\begin{aligned} D_{KL}[P_X||P_F(U)] &= \mathbb{E}_{x\sim p_X}\left[\log p_X(x) - \log p_{F(U)}(x)\right] \\ &= \mathbb{E}_{x\sim p_X}\left[\log p_X(x) - \log p_U(F^{-1}(x)) - \log\left|\det \frac{\partial F^{-1}(x)}{\partial x}\right|\right] \end{aligned}$$

We substitute by the change of variables in the expectation

$$\begin{aligned} x &= F(u) \\ \partial x &= \left|\det \frac{\partial F(u)}{\partial u}\right| \end{aligned}$$

Since F is invertible, the density of X under the transformation F is related to the density of U by:

$$p_{F(U)}(x) = p_U(F^{-1}(x))\left|\det \frac{\partial F^{-1}(x)}{\partial x}\right|$$

With:

$$\left|\det \frac{\partial F^{-1}(x)}{\partial x}\right| = \left|\det \frac{\partial F(u)}{\partial u}\right|^{-1}$$

As a result:

$$\begin{aligned} D_{KL}[P_X||P_F(U)] &= \mathbb{E}_{u\sim p_{F^{-1}}(x)}\left[\log p_X(F(u)) - \log p_U(u) - \log\left|\det \frac{\partial F(u)}{\partial u}\right|^{-1}\right] \\ &= \mathbb{E}_{u\sim p_{F^{-1}}(x)}\left[\log p_X(F(u)) - \log p_U(u) + \log\left|\det \frac{\partial F(u)}{\partial u}\right|\right] \\ &= \int_u p_{F^{-1}(x)}(u)\log\left(\frac{p_{F^{-1}(x)}(u)}{p_u(u)}\right)du \\ &= D_{KL}[P_{F^{-1}(X)}||P_U] \end{aligned}$$

Log propreties used:

$$log(a^{-1}) = -log(a)$$

2. For scenario 1 we would use the forward $D_{KL}[P_X||P_F(U)]$ since we need to calculate the probability density $P_X(x)$ and the divergence is expressed as an expectation under the distribution P_X, which is doable since we have $P_X(x)$.

   For scenario 2 we would use the reverse $D_{KL}[P_{F^{-1}(X)}||P_U]$ since we need the the model distribution $P_F(U)$ and it can be estimated using samples from P_X. We cannot use the forward KL divergence here as we need to know $P_X(x)$.

Question 3 (3-8-3-14)**.** (**Variational Autoencoders**)

1. Let $p_x^*(.)$ be the true data distribution and $p_x(.;\theta)$ be the model distribution parametrized over θ, a natural criterion to define if $p_x(.;\theta)$ is accurately portraying $p_x^*(.)$ is the *Maximum Likelihood Estimation* (MLE). Sometimes, knowledge about the data can lead us to adopt a model with hidden intermediate variable z to approximate the data distribution, where only the joint distribution $p_{x,z}(.,.,\theta)$ are explicitly defined. For such models, we need to calculate the marginal likelihood $p_x(.) = \int_z p_{x,z}(.,z,\theta)dz$, however, this proves to be difficult. Why ?
   (a) We do not know about $p_(x|z)$ and thus cannot calculate the integral.
   (b) Integration over the hidden variable z can prove to be intractable due to the complexity of $p_(x|z)$ and the curse of dimensionality.
   (c) We don't know and cannot assume what z looks like (i.e. what kind of distribution) and thus cannot calculate the integral.
   (d) The integral over the hidden variable z is intractable because it does not follow a standard distribution like Gaussian or Bernoulli.
2. To avoid the above problem, we can try to avoid $p_x(.)$ and instead aim to establish a lower bound function of it. This involves rewriting the log of the marginal likelihood $\log p_x(.) = \log \int_z p_{x,z}(.,z,\theta)dz$ as a combination of a KL divergence and an *Evidence Lower Bound* (ELBO). This process is facilitated by the introduction of an approximate posterior $q(z|x)$ which approximates the unknown true posterior $p(z|x)$. The choice of q is arbitrary, but we often choose it from simpler classes of distributions such as the Gaussian for practical reasons. Your task is to derive the ELBO function in two ways:
   (a) By decomposing the marginal likelihood as the combination of a KL-divergence between variational and true posteriors over z ($D_{KL}(q(z|x)||p(z|x))$) and the ELBO.
   (b) By using the Jensen Inequality.
3. What is the significance of the above result ? Select all that apply.
   (a) $p_x(.)$ has a lower bound which is the ELBO.
   (b) Maximizing the ELBO is equivalent to minimizing the distributional difference between the approximation $q(z|x)$ and the true (but intractable) $p(z|x)$.
   (c) The ELBO offers a theoretical bound but is not useful in practice for training models with latent variables.
   (d) The choice of q affects the tightness of the lower bound.
4. This question is about importance weighted autoencoder. When training a variational autoencoder, the standard training objective is to maximize the evidence lower bound (ELBO). Here we consider another lower bound, called the Importance Weighted Lower Bound (IWLB), a tighter bound than ELBO, defined as

$$\mathcal{L}_k = \mathbf{E}_{z_{1:k}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{1}{k}\sum_{j=1}^{k}\frac{p(\boldsymbol{x},z_j)}{q(z_j \mid \boldsymbol{x})}\right]$$

   for an observed variable $\boldsymbol{x}$ and a latent variable $\boldsymbol{z}$, k being the number of importance samples. The model we are considering has joint that factorizes as $p(\boldsymbol{z},\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$, $\boldsymbol{x}$ and $\boldsymbol{z}$ being the observed and latent variables, respectively. In the following questions, one needs to make use

of the Jensen's inequality:

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)]$$

for a convex function f.

(a) Show that IWLB is a lower bound on the log likelihood $\log p(\boldsymbol{x})$.
(b) Given a special case where $k = 2$, prove that $\mathcal{L}_2$ is a tighter bound than the ELBO (with $k = 1$).

Answer 3. 1. (b) Integration over the hidden variable z can prove to be intractable due to the complexity of $p_(x|z)$ and the curse of dimensionality.

2. (a) We have:

$$\begin{aligned} D_{KL}(q(z|x)||p(z|x)) &= \mathbb{E}_{q(z|x)}\left[\log q(z|x) - \log p(z|x)\right] \\ &= \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(z|x)}\right] \end{aligned}$$

The expectation over $q(z|x)$ has no effect on $p(x)$ because:

$$\log p(x) = \log p(x) \int_z q(z|x)dz = \int_z q(z|x) \log p(x)dz = \mathbb{E}_{q(z|x)}[\log p(x)]$$

We introduce the distribution $q(z|x)$ and apply Baye's rule $p(x) = \frac{p(x,z)}{p(z|x)}$:

$$\begin{aligned} \log p(x) &= \mathbb{E}_{q(z|x)}\left[\log p(x)\right] \\ &= \mathbb{E}_{q(z|x)}\left[log \frac{p(x,z)}{p(z|x)}\right] \\ &= \mathbb{E}_{q(z|x)}\left[log \frac{p(x,z)}{p(z|x)}\frac{q(z|x)}{q(z|x)}\right] \\ &= \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)q(z|x)}{q(z|x)p(z|x)}\right] \\ &= \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] + \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(z|x)}\right] \\ &= \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] + D_{KL}(q(z|x)||p(z|x)) \end{aligned}$$

So:

$$ELBO = \log p(x) - D_{KL}(q(z|x)||p(z|x))$$

(b) The Jansen inequality applies to a concave function (such as log) with the inequality reversed. So $\log \mathbb{E}[X] \geq \mathbb{E}[\log X]$:

$$\begin{aligned} \log p(x) &= \log \int_z p(x,z)dz \\ &= \log \int_z p(x,z)\frac{q(z|x)}{q(z|x)}dz \\ &= \log \mathbb{E}_{q(z|x)}\left[\frac{p(x,z)}{q(z|x)}\right] \\ &\geq \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right] \quad \text{We apply Jansen's inequality} \end{aligned}$$

$ELBO = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z)}{q(z|x)}\right]$ so $\log p(x) \geq ELBO$

3. (a), (b), (d)
4. (a) We need to show that $\log p(x) \geq \mathcal{L}_k$. We have:

$$\log p(x) = \log \mathbb{E}_{q(z|x)} \left[\frac{p(x,z)}{q(z|x)} \right]$$

If we consider k independant samples from $q(z|x)$:

$$\log p(x) = \log \mathbb{E}_{z_{1:k} \sim q(\boldsymbol{z}|\boldsymbol{x})} \left[\frac{1}{k} \sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right]$$

With Jansen's inequality, we can say that:

$$\begin{aligned} \log p(x) &\geq \mathbb{E}_{z_{1:k} \sim q(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{1}{k} \sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right] \\ &\geq \mathcal{L}_k \end{aligned}$$

As a result, IWLB is a lower bound on the log likelihood $\log p(x)$

(b) Inspired by paper: IMPORTANCE WEIGHTED AUTOENCODERS

$$\mathcal{L}_1 = \mathbb{E}_{z \sim q(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, z_1)}{q(z_1 \mid \boldsymbol{x})} \right] = ELBO$$

$$\mathcal{L}_2 = \mathbb{E}_{z_{1:2} \sim q(z|x)} \left[\log \frac{1}{2} \sum_{j=1}^{2} \frac{p(x,z_j)}{q(z_j|x)} \right]$$

We need to show that $\mathcal{L}_2 \geq \mathcal{L}_1$.

Let $I \subseteq \{1, \ldots, k\}$ with $|I| = m$ be a uniformly distributed subset of distinct indices from $\{1, \ldots, k\}$. We will use the identity:

$$\mathbb{E}_{z_{1:m} \sim q(z|x)} \left[\frac{1}{m} \sum_{j=1}^{m} \frac{p(x,z_j)}{q(z_j|x)} \right] = \frac{1}{k} \sum_{i=1}^{k} \frac{p(x,z_i)}{q(z_i|x)}$$

We can get:

$$\mathbb{E}_{z_1 \sim q(z|x)} \left[\frac{p(x,z_1)}{q(z_1|x)} \right] = \frac{1}{2} \sum_{i=1}^{2} \frac{p(x,z_i)}{q(z_i|x)}$$

As a result:

$$
\begin{aligned}
\mathcal{L}_2 &= \mathbb{E}_{z_1:2\sim q(z|x)}\left[\log \frac{1}{2}\sum_{j=1}^{2}\frac{p(x,z_j)}{q(z_j|x)}\right] \\
&= \mathbb{E}_{z_{1:2}\sim q(z|x)}\left[\log \mathbb{E}_{z_1\sim q(z|x)}\left[\frac{p(x,z_1)}{q(z_1|x)}\right]\right] \\
&\geq \mathbb{E}_{z_{1:2}\sim q(z|x)}\left[\mathbb{E}_{z_1\sim q(z|x)}\left[\log \frac{p(x,z_1)}{q(z_1|x)}\right]\right] \quad \text{(Jensen inequality for concave function log)} \\
&\geq \mathbb{E}_{z_1\sim q(z|x)}\left[\log \frac{p(x,z_1)}{q(z_1|x)}\right] \quad (4) \\
&\geq \mathcal{L}_1
\end{aligned}
$$

(4) : Remove the nested expectation $\rightarrow z_{1:2}$ and z_1 are drawn from the same distribution $q(z|x)$, the outer expectation over $z_{1:2}$ becomes redundant since we're only interested in z_1.

Therefore, $\mathcal{L}_2 \geq ELBO$ and $\mathcal{L}_2$ is a tighter bound than the ELBO.

Question 4 (2-2-2-3-3-10). (**Generative Adversarial Networks**)

1. Consider a Generative Adversarial Network (GAN) which successfully produces images of apples. Which of the following propositions is false ?
   (a) The generator aims to learn the distribution of apple images.
   (b) The discriminator can be used to classify images as apple vs. non-apple.
   (c) After training the GAN, the discriminator loss eventually reaches a constant value.
   (d) The generator can produce unseen images of apples.
2. Which of the following cost functions is the non-saturating cost function for the generator in GANs (G is the generator and D is the discriminator) ? Note that the cost function will be minimized w.r.t the generator parameters during training.
   (a) $J^{(G)} = \frac{1}{m}\sum_{i=1}^{m} \log(1 - D(G(z^{(i)})))$
   (b) $J^{(G)} = -\frac{1}{m}\sum_{i=1}^{m} \log(D(G(z^{(i)})))$
   (c) $J^{(G)} = \frac{1}{m}\sum_{i=1}^{m} \log(1 - G(D(z^{(i)})))$
   (d) $J^{(G)} = -\frac{1}{m}\sum_{i=1}^{m} \log(G(D(z^{(i)})))$
3. After training a neural network, you observe a large gap between the training accuracy (100%) and the test accuracy (42%). Which of the following methods is commonly used to reduce this gap ?
   (a) Generative Adversarial Networks
   (b) Dropout
   (c) Sigmoid activation
   (d) RMSprop optim
4. Given the two options of (A) saturating cost and (B) non-saturating cost, which cost function would you choose to train a GAN ? Explain your reasoning. (1-2 sentences)
5. You are training a standard GAN, and at the end of the first epoch you take note of the values of the generator and discriminator losses. At the end of epoch 100, the values of the loss functions are approximately the same as they were at the end of the first epoch. Why are the quality of generated images at epoch 1 and epoch 100 not necessarily similar ? (1-2 sentences)
6. Let p_0 and p_1 be two probability distributions with densities f_0 and f_1 (respectively). We want to explore what we can do with a trained GAN discriminator. A trained discriminator is thought to be one which is "close" to the optimal one:

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(\boldsymbol{x}))].$$

   (a) For the first part of this problem, derive an expression we can use to estimate the Jensen-Shannon divergence (JSD) between p_0 and p_1 using a trained discriminator. We remind that the definition of JSD is $\text{JSD}(p_0, p_1) = \frac{1}{2}\big(KL(p_0\|\mu) + KL(p_1\|\mu)\big)$, where $\mu = \frac{1}{2}(p_0 + p_1)$.
   (b) For the second part, we want to demonstrate that a optimal GAN Discriminator (i.e. one which is able to distinguish between examples from p_0 and p_1 with minimal NLL loss) can be used to express the probability density of a datapoint $\boldsymbol{x}$ under f_1, $f_1(\boldsymbol{x})$ in terms of $f_0(\boldsymbol{x})$ [2]. Assume f_0 and f_1 have the same support. Show that $f_1(\boldsymbol{x})$ can be estimated by $f_0(\boldsymbol{x})D(\boldsymbol{x})/(1 - D(\boldsymbol{x}))$ by establishing the identity $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1 - D^*(\boldsymbol{x}))$.

2. You might need to use the "functional derivative" to solve this problem. See "19.4.2 Calculus of Variations" of the Deep Learning book or "Appendix D Calculus of Variations" of Bishop's Pattern Recognition and Machine Learning for more information.

Hint: Find the closed form solution for D^*.

Answer 4. 1. (c)

2. (b)

3. (d)

4. We would chose non-saturating since we want to avoid the generator's gradient saturating, meaning avoiding it becomes so small, it slows or even stops the learning of the gradient.

5. The loss doesn't reflect the quality of the images. The generator could be producing increasingly higher quality images, but if the discriminator is also improving at a similar rhythm, the loss values may not significantly change, even though the quality of the generated images is improving through the epochs.

6. (a) We take the equation of D* from the slide 6 of the presentation about GANs

$$D^*(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_1(\boldsymbol{x}) + f_0(\boldsymbol{x})}$$

The D_{KL} terms of our JSD can be refactored as:

$$\begin{aligned} KL(p_0 \| \mu) &= \mathbb{E}_{\boldsymbol{x}\sim p_0}\left[\log f_0(\boldsymbol{x}) - \log \frac{1}{2}(f_0(\boldsymbol{x}) + f_1(\boldsymbol{x}))\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_0}\left[\log 2\frac{f_0(\boldsymbol{x})}{f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})}\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_0}\left[\log 2(1 - D^*(\boldsymbol{x}))\right] \end{aligned}$$

$$\begin{aligned} KL(p_1 \| \mu) &= \mathbb{E}_{\boldsymbol{x}\sim p_1}\left[\log f_1(\boldsymbol{x}) - \log \frac{1}{2}(f_0(\boldsymbol{x}) + f_1(\boldsymbol{x}))\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_1}\left[\log 2\frac{f_1(\boldsymbol{x})}{f_1(\boldsymbol{x}) + f_0(\boldsymbol{x})}\right] \\ &= \mathbb{E}_{\boldsymbol{x}\sim p_1}\left[\log 2D^*(\boldsymbol{x})\right] \end{aligned}$$

As a result, we can use the following equation to estimate JSD:

$$JSD(p_0, p_1) = \frac{1}{2}\left(\mathbb{E}_{\boldsymbol{x}\sim p_0}\left[\log 2(1 - D^*(\boldsymbol{x}))\right] + \mathbb{E}_{\boldsymbol{x}\sim p_1}\left[\log 2D^*(\boldsymbol{x})\right]\right)$$

(b) We have $D^*(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_1(\boldsymbol{x})+f_0(\boldsymbol{x})} \Rightarrow f_1(\boldsymbol{x}) = D^*(\boldsymbol{x})(f_1(\boldsymbol{x}) + f_0(\boldsymbol{x}))$:

$$\begin{aligned} f_1(\boldsymbol{x}) - D^*(\boldsymbol{x})f_1(\boldsymbol{x}) &= D^*(\boldsymbol{x})f_0(\boldsymbol{x}) \\ f_1(\boldsymbol{x})(1 - D^*(\boldsymbol{x})) &= D^*(\boldsymbol{x})f_0(\boldsymbol{x}) \\ f_1(\boldsymbol{x}) &= \frac{D^*(\boldsymbol{x})f_0(\boldsymbol{x})}{(1 - D^*(\boldsymbol{x}))} \end{aligned}$$

In conclusion, the probability density $f_1(\boldsymbol{x})$ under the optimal discriminator can be estimated as $\frac{D^*(\boldsymbol{x})f_0(\boldsymbol{x})}{(1-D^*(\boldsymbol{x}))}$.

Question 5 (5-5-5-5)**.** (**Self-Supervised Learning: Paper Review**)

In this question, you are going to write a **one page review** of the A Simple Framework for Contrastive Learning of Visual Representations paper.

Your review should have the following four sections: Summary, Strengths, Weaknesses, and Reflections. For each of these sections, below we provide a set of questions you should ask about the paper as you read it. Then, discuss your thoughts about these questions in your review.

(5.1) **Summary:**

(a) What is this paper about ?
(b) What is the main contribution ?
(c) Describe the main approach and results. Just facts, no opinions yet.

(5.2) **Strengths:**

(a) Is there a new theoretical insight ?
(b) Or a significant empirical advance ? Did they solve a standing open problem ?
(c) Or a good formulation for a new problem ?
(d) Any good practical outcome (code, algorithm, etc) ?
(e) Are the experiments well executed ?
(f) Useful for the community in general ?

(5.3) **Weaknesses:**

(a) What can be done better ?
(b) Any missing baselines ? Missing datasets ?
(c) Any odd design choices in the algorithm not explained well ? Quality of writing ?
(d) Is there sufficient novelty in what they propose ? Minor variation of previous work ?
(e) Why should anyone care ? Is the problem interesting and significant ?

(5.4) **Reflections:**

(a) How does this relate to other concepts you have seen in the class ?
(b) What are the next research directions in this line of work ?
(c) What (directly or indirectly related) new ideas did this paper give you ? What would you be curious to try ?

This question is subjective and so we will accept a variety of answers. You are expected to analyze the paper and offer your own perspective and ideas, beyond what the paper itself discusses.

Answer 5. Summary:
The paper introduces SimCLR, a new framework for contrastive learning of visual representations, aiming to improve the effectiveness of learning image representations without relying on labeled data. This framework significantly enhances the performance of unsupervised learning in visual tasks by uniquely utilizing various tools to maximize agreement between differently augmented views of the same data example. Specifically, SimCLR employs creative data augmentations to generate new views, uses larger batch sizes, and a MLP at the end of the network for representation learning (improving >10% with it). Importantly, SimCLR achieves state-of-the-art performance on different well-known datasets, markedly outperforming by 7% existing methods in specific metrics, which underscores the framework's effectiveness in enhancing image recognition tasks.

Strength:
The paper marks a significant advance in unsupervised learning, it builds upon existing knowledge with innovative approaches rather than introducing new theoretical insights. The experiments are comprehensive and meticulously detailed, including extensive information about each experiment's processes. Section B of the appendix, offering additional experimental results, further solidifies the thoroughness of the research. The inclusion of SimCLR's pseudo-algorithm and an in-depth explanation of the processes provides the community with valuable tools for experimentation, especially beneficial in domains where labeled data is scarce. The paper effectively advances the methodology within unsupervised learning rather than posing a new problem, giving us reflections on the future potential of unsupervised versus supervised learning.

Weakness:
The method's reliance on large batch sizes and extended training epochs may limit its practical application in environments with limited resources. Although the paper tests the framework across 12 natural image classification datasets, the inclusion of more diverse datasets, could have further established the framework's robustness/versatility. While the framework is effective, I would say it's more of an incremental advance over previous contrastive learning methods, rather than a completely novel design. Moreover, the paper could benefit from more insiights into specific applications or domains where the unsupervised approach would be most impactful.

Reflection:
This paper's approach to contrastive learning has notable parallels with the principles of Generative Adversarial Networks (GANs), particularly in how each system learns to distinguish between different types of data representations. Additionally, the emphasis on unsupervised learning (especially contrastive) aligns with from Olivier J. Hénaff's presentation we had in class. Future research could integrate DINO's self-supervised learning principles or BERT's context-based learning approach to enhance contrastive learning frameworks. Moreover, applying SimCLR's methodology to text data, similar to BERT's approach, could be a fruitful direction. Combining SimCLR with generative models like GANs or Variational Autoencoders might lead to more sophisticated semi-supervised learning techniques, enhancing the model's ability to differentiate and generate complex data distributions. Finally, it could inspire us to tackle the challenge of video data augmentation, focusing on the development of fast and scalable methods,we could significantly improve the applicability of contrastive learning for video analysis with SimCLR's augmentation techniques.