آزمون پایانترم یادگیری عمیق **دانشکده برق دانشگاه صنعتی شریف** **دیماه ۱۴۰۲**

❖ **سوال یک (۲۵) :**

الف) (۱۵) $z_{sioux} = \underset{z}{argmin} \|G(z;\Phi) - x_{sioux}\|_2^2$

ب) (۱۰) $\theta_{Enc}^* = \underset{\theta_{Enc}}{argmin} \sum_n \|G(Enc(x_n;\theta_{Enc});\Phi) - x_n\|_2^2$

❖ **سوال دو (۶): تمامی روش‌هایی augmentation که محتوی تصویر را عوض نکند، موارد غیر مجاز مانند: Blurring، Noise، و ....**

❖ **سوال سوم (۸):**

غلط	**دلیل اصلی استفاده از PE جبران عدم استفاده از اطلاعات توالی در معماری Transformerها است.**

❖ **سوال چهارم: (۸ نمره)**

غلط	**خیر در هر دو استفاده می شود. امکان استفاده از ویژگی‌های مختلف از Qها را فراهم می‌کند که امکان استخراج ویژگی متنوع از مکان‌های مختلف در هر دو بخش را فراهم می‌کند.**

❖ **سوال پنجم: (۸ نمره)**

دلیل	**مکانیسم توجه، سه گانه قابل یادگیری Q/K/V**

❖ **سوال ششم: (۸ نمره)**

صحیح	**در بخش Encoder نیازی به MSA نداریم، زیرا کل رشته را با هم داریم.**

❖ **سوال هفتم: (۹ نمره)**

✗	متوسط هر متغیر مخفی، $\mathbb{E}(Z_i)$، از توزیع گوسی پیروی می‌کند: **در VAE توزیع هر متغیر مخفی بر اساس متوسط و واریانس آن تخمین زده می‌شوند که هر دو عدد یقینی هستند**
✓	در تخمین $q(Z)$، استقلال بین ابعاد وجود دارد: **بله این فرض جهت ساده کردن فرم ماتریس کوواریانس وجود دارد.**
✗	متوسط هر متغیر مخفی، $\mathbb{E}(Z_i)$، صفر است.: **چنین فرضی وجود ندارد، هم متوسط و هم واریانس در هر بعد تخمین زده می‌شود.**

❖ **سوال هشتم: (۱۵ نمره)**

✓	در گام‌های ابتدایی آموزش شبکه GAN مقدار $D(G(z))$ نزدیک به صفر است. **به دلیل اینکه آموزش D سریعتر و بهتر از G صورت می‌گیرد و در ابتدا شبکه مولد ضعیف عمل می‌کند، بنابر این در اوایل امکان تشخیص داده Fake وجود دارد.**
✓	در عمل از تابع هزینه $-\frac{1}{m}\sum_{i=1}^{m} log\left(D\left(G\left(z^{(i)}\right)\right)\right)$ به جای $\frac{1}{m}\sum_{i=1}^{m} log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)$ استفاده می‌شود. **برای جلوگیری از صفر شدن گرادیان در ابتدای آموزش از تابع هزینه اول استفاده می‌شود. در حالت $D(G(z))\sim 0$ که اوایل آموزش است، مقدار گرادیان بزرگ است و آموزش بهتر رخ می‌دهد.**
✗	توقف آموزش شبکه زمانی صورت می‌گیرد که به دفعات متوالی $D\left(G\left(z^{(i)}\right)\right)$ برابر با یک باشد.، **نقطه بهینه زمانی است که برای داده Real و Fake هر دو برابر با ۰/۵ باشد.**

❖ **مسئله نهم: (۳۰ نمره)**

Forget Gate for memory reset	$\boldsymbol{f}^t$
Input Gate for input weighting	$\boldsymbol{i}^t$
output Gate for output weighting	$\boldsymbol{o}^t$
Input Block (prepare input for memory cell)	$\boldsymbol{z}^t$
Memory cell	$\boldsymbol{c}^t$
output	$\boldsymbol{h}^t$

$$\frac{\partial J}{\partial \boldsymbol{W}_g} = \sum_{t=1}^{T} \frac{\partial J}{\partial \boldsymbol{h}_t}\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{W}_g} = \sum_{t=1}^{T} \frac{\partial J}{\partial \boldsymbol{h}_t}\left(\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{c}^t}\left(\frac{\partial \boldsymbol{c}^t}{\partial \boldsymbol{W}_g} + \frac{\partial \boldsymbol{c}^t}{\partial \boldsymbol{c}^{t-1}}\frac{\partial \boldsymbol{c}^{t-1}}{\partial \boldsymbol{W}_g}\right) + \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}}\frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{W}_g}\right)$$

$\frac{\partial \boldsymbol{c}^t}{\partial \boldsymbol{c}^{t-1}} = \boldsymbol{f}^t$ عامل کنترل کننده است

❖ **مسئله دهم: (۳۵ نمره)**

الف) (۱۶+۸) $\mathbb{E}\{\boldsymbol{q}^T\boldsymbol{k}\} = \boldsymbol{\mu}^T\boldsymbol{\mu}, \quad var\{\boldsymbol{q}^T\boldsymbol{k}\} = 2\sigma^2\boldsymbol{\mu}^T\boldsymbol{\mu} + d\sigma^4$

ب) **(۶ نمره)** ترم واریانس وابستگی مستقیم دارد، باعث افزایش واریانس (مشکل همگرایی) و افزایش ابعاد می شود.

$var\{\boldsymbol{q}^T\boldsymbol{k}\} = 2\sigma^2\boldsymbol{\mu}^T\boldsymbol{\mu} + d\sigma^4$

ج) **(۵ نمره)** باید اثر d را در واریانس از میان برد. $softmax\left(\frac{\boldsymbol{q}^T\boldsymbol{k}}{\sqrt{d}}\right)$

❖ **مسئله یازدهم: (۱۵ نمره)**

شکل ۱	شکل ۲	شکل ۳
تابع هدف (۵ نمره): الف **این تابع هدف فقط برابری ورودی و خروجی را به حداکثر می‌رساند و قیدی برای توزیع فضای مخفی قائل نمی‌شود.**	**تابع هدف (۵ نمره): ب** **این تابع هدف تلاش دارد که ضمن نزدیک بودن فضای مخفی به گوسی متوسط صفر و ماتریس کوواریانس یک، داده‌های ورودی و خروجی شبیه هم بشوند.**	**تابع هدف (۵ نمره): ج** **این تابع هدف فقط تلاش دارد که فضای مخفی نزدیک گوسی متوسط صفر و ماتریس کوواریانس یک بشود و نظمی برای داده‌ها قابل اعمال نیست.**

❖ **سوال دوازدهم: (۱۶ نمره)**

✗	یادگیری خود نظارتی (Self-Supervised Learning) همواره نیاز به داده برچسب گذاری شده را به طور کامل حذف می‌کند. غلط: به صورت معمول، با Pretext task آموزش اولیه شبکه رو تعداد بالای داده بدون برچسب انجام میشود و سپس با تعدادی برچسب محدود، وظیفه مورد نظر آموزش داده می‌شود.
✗	در یادگیری خودنظارتی، هدف بیشینه کردن دقت شبکه در انجام pretext task تعریف شده است. غلط: به صورت کلی دقت شبکه در *pretext task* بدست آمده مهم نیست، هدف اصلی بدست آوردن بهترین دقت در *downstream task* است.
✓	در روش‌های مبتنی بر contrastive learning در یادگیری خودنظارتی، batch size بزرگ (بیش از ۱۰۰۰) برای همگرایی به جواب مطلوب الزامی **نیست**. درست: batch size با اندازه بزرگ برای روش پیشنهاد شده در مقاله SimCLR بسیار مهم است به خاطر اینکه نمونه‌های منفی نیز از batch ورودی ساخته میشوند، ولی در مقالاتی مانند MoCo که نمونه های منفی را در حافظه نگه میدارد، میتوان با batch size با اندازه محدود (۲۵۶) نیز به جواب مطلوب رسید، در واقع نکته مهم تعداد بالای نمونه‌های منفی است و نه batch size ورودی
✗	دو تغییری که باعث افزایش دقت روش‌های مبتنی بر contrastive learning برای یادگیری خودنظارتی شد، استفاده از data augmentation های قوی و linear project head است. غلط: هرچند که استفاده از Linear project head دقت را بهبود میبخشد، ولی بهترین دقت با استفاده از non-linear projection head برای حل مسائل downstream task بدست میاید.

❖ **سوال سیزدهم: (+۱۰ نمره)** اثرگذارترین دانشمند علوم داده بر روی تحولات هوش مصنوعی در دهه اخیر کدام است (**یک** گزینه)

Geoffrey Hinton	Yann LeCun	Fei-Fei Li	Yoshua Bengio	Ian Goodfellow	Andrew Ng

❖ **مسئله چهاردهم: (۴۵ نمره)**

الف) **(۱۵ نمره)** $F = 4 - \int \left\{ \frac{p}{D} + \frac{q}{1-D} \right\} dx$

پ) **(۱۰ نمره)** $D = \frac{\sqrt{p}}{\sqrt{p}+\sqrt{q}}$

ت) **(۱۵ نمره)** $F = 2 - 2\int \sqrt{pq}dx = \int \left(\sqrt{p} - \sqrt{q}\right)^2 dx$ معیار مربعات فاصله بین دو توزیع است.

ث) **(۵ نمره)** $D = \frac{\sqrt{p}}{\sqrt{p}+p} = 0.5$