AVILEN　E資格試験想定問題集 vol.3

Transformer

【問題】
時系列処理においてRNNやCNNを用いずにAttentionのみを用いたモデルとしてTransformerがある。
このTransformerにはSource-target-attentionとSelf-attentionの2種類のAttentionが用いられている。以下の図においてSource-target-attentionを表しているのは(あ)であり、Self-attentionを表しているのは(い)である。
(あ)(い)に入るものとして正しい組み合わせを選べ。
また，図の左側はエンコーダ、右側はデコーダを表している。

Figure 1: The Transformer - model architecture.

【選択肢】
(a)(あ):①　　(い)②、③
(b)(あ):①、②　(い) ③
(c)(あ):②　　(い)①、③
(d)(あ):③、①　(い) ②

【解答】
(a)(あ):①　(い)②、③

【解説】
Source-target-attentionとSelf-attentionの大きな違いはqueryとAttention　Weightを測る対象である、keyとvalueがqueryと同じかどうかである。
図中①にはAttentionへの入力にエンコーダ側の出力とデコーダ側の値が用いられているため、queryに対してkey,valueが異なるSource-target-attentionである。
また②、③はAttention入力がすべて同じ値の為、query,key,valueが同じである、Self-attentionである。

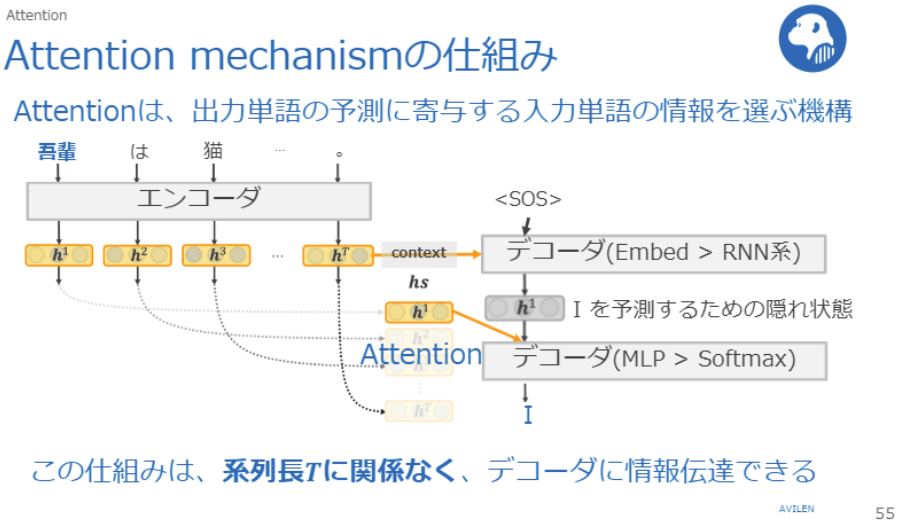
Self-attention, Source-target-attention

【問題】
Attentionは従来の手法と比べてどんな利点があるかを述べた以下の文のうち、誤っているものを選べ。

【選択肢】
(a)CNNと比べるとより長い系列を扱うことができる。
(b)CNNよりも勾配消失が起きづらい。
(c)RNNではできない並列計算が可能である。
(d)Seq2Seqのエンコーダの出力を系列長に関係なくデコーダに伝達することができる。

【解答】
(b)CNNよりも勾配消失が起きづらい。

【解説】
CNNはカーネルサイズよりも大きい系列の依存関係を学習することができないが、Attentionはすべてのkey,value要素に対してqueryのAttention　Weightを計算することから、CNNよりも長い系列に対して依存関係を学習することができる。
しかしこれによって勾配消失問題が避けられるわけではないため誤り。
勾配消失を避ける手法としてqueryとkeyの行列積の結果に対し、次元数の平方根で割ることによってスケーリングを導入したScaled Dot-Product Attentionがある。
またRNNの処理には一つ前の系列の中間層を用いる影響から、一つ前の系列の計算が終わるまで次の系列の計算は行うことができず、処理時間が系列長に応じて伸びてしまうが、Attentionはすべてのkey,value要素に対してqueryのAttention Weightを行列積で一度に求めることができるため、並列計算が可能である。
他にも下図のようにシンプルなSeq2Seqでは捨てられていたエンコーダの出力に対して、Attentionを用いれば系列長に関係なくデコーダ側に情報伝達をすることができる。



Attentionの理論式

【問題】
以下の式の中からAttentionの計算式として正しいものを選べ。また式中のQはquery、Kはkey、Vはvalueを表し、演算子'・'は行列積を表す。

【選択肢】
(a)$Softmax(Q \cdot K^T \cdot V)$
(b)$Softmax(Q \cdot K^T) \cdot V$
(c)$Softmax(Q \cdot K \cdot V)$
(d)$Softmax(Q \cdot K) \cdot V$

【解答】

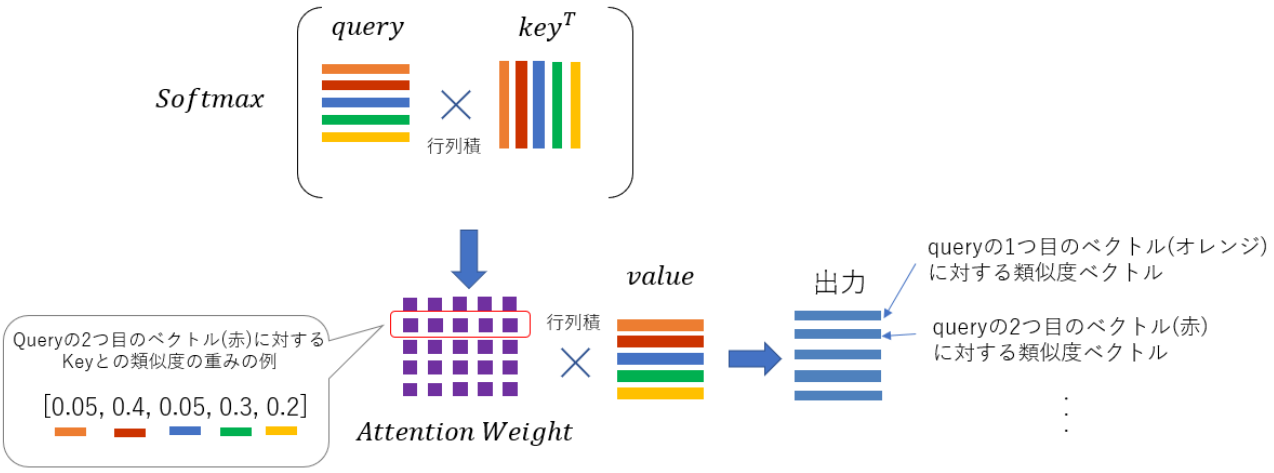
(b)$Softmax(Q \cdot K^T) \cdot V$

【解説】

Attentionは一種の辞書オブジェクトとみなすことができ、queryと辞書(key,value)との類似度(重要度)の高いベクトルを重みづけ和で得るものである。

以下の図(Self-attention)のように各queryとkeyとのベクトルの類似度を行列積によって計算する。この時、keyは転置することに注意する。この計算結果にSoftmax関数を通すことによって、あるqueryのベクトルに対してどのkeyがどれだけ似ているかを範囲[0,1]の総和1で出力する。これは各ベクトルを類似度によって重みづけしており、Attention Weightと呼ぶ。

各queryにkeyの重みづけを行ったAttention Weightとvalueとの行列積を計算することによって、queryと辞書(key,value)との類似度の高いベクトルを重みづけ和で得ることができる。



識別モデル

【問題】
識別関数と識別モデルに関する説明として誤っているものを選べ。

【選択肢】
(a)識別関数を用いることでデータ空間上に超平面を描くことができ、データの分類が直感的に理解できる。
(b)識別モデルは入力データxに対するクラスCの条件付き確率P(C|x)をモデル化する。
(c)ニューラルネットワークは識別モデルに、サポートベクトルマシンやロジスティック回帰は識別関数にあたる。
(d)識別関数は識別モデルよりも少ないデータ数でも問題無くクラス分類できるというメリットを持つ。

【解答】
(c)ニューラルネットワークは識別モデルに、サポートベクトルマシンやロジスティック回帰は識別関数にあたる。

【解説】
ロジスティック回帰もニューラルネットワークと同様に入力に対して条件付き確率を出力するので、識別モデルである。よって上の説明は誤り。
識別関数は入力されたデータに対して属するクラスを出力する関数の事を指す。二次元データならば直線、三次元データならば平面、の様にデータ空間上に超平面を構築することでデータの分類が行われる。
識別モデルは入力データに対して直接クラスを出力するのではなく、各クラスに属する確率をモデル化したものを指す。
識別関数と識別モデルを比べると、識別関数は分布の推定を伴わないので識別モデルよりもデータ数が少しで済むというメリットがある。
識別モデルの例にはニューラルネットワーク、識別関数の例としてはサポートベクトルマシンなどが挙げられる。
また、分類問題を例として考えた時、そのデータ空間を超平面によって分離することが出来るならばそのデータは線形分離可能であり、その様なデータをついて識別を行うモデルを線形モデルと呼ぶ。逆に線形分離不可能であればその識別を行うモデルを非線形モデルと呼ぶ。
先程挙げた例について考えると、サポートベクトルマシンとロジスティック回帰は線形モデル、ニューラルネットワークは非線形モデルである。

【参考】
https://www.hellocybernetics.tech/entry/2017/06/08/010513

生成モデル

【問題】
生成モデルについての説明として誤っているものを選べ。

【選択肢】
(a)生成モデルの代表的な例として、VAE、GANなどが挙げられる。
(b)識別モデルと比べると数学的に複雑であり、多くのデータが必要とされるというデメリットが存在する。
(c)特定のクラスに属する様な人工的なデータを作成することができたり、異常検知のタスクに応用することができる。
(d)以下の様に表される式をモデル化する。(C_k:クラス、x:入力データ)

$$\frac{p(C_k|\boldsymbol{x})p(C_k)}{p(\boldsymbol{x})}$$

【解答】
(d)以下の様に表される式をモデル化する。(C_k:クラス、x:入力データ)

$$\frac{p(C_k|\boldsymbol{x})p(C_k)}{p(\boldsymbol{x})}$$

【解説】
生成モデルでは、入力データxに対するクラスC_kの条件付き確率P(C_k|x)にベイズの定理を適用した以下の式をモデル化するので、誤り。

$$p(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})}$$

生成モデルは、P(C_k|x)を事後確率と捉えてベイズ的にモデル化する。クラスに対する入力データの条件付き確率分布p(x|C_k)を求めることによって、データの生成が可能であるということが最大の特徴。しかし、識別モデルと比較すると当然モデル化する際に学習するパラメータの量も増えるため、必要な学習データも多くなる。
生成モデルの例としては、
AE(オートエンコーダ)において潜在変数を確率分布に落とし込むことで生成画像に連続性とランダム性を持たせたVAE、
GeneratorとDiscriminatorの二つのネットワークを用意してお互いに競合させて学習を進めることでより良い性能のネットワークを作り、高品質な画像生成を行うGANなどが挙げられる。生成モデルは異常検知に応用することもでき、例えばGANの派生モデルであるAnoGANは異常検知タスクに有用とされている。

【参考】
https://www.hellocybernetics.tech/entry/2017/06/08/010513

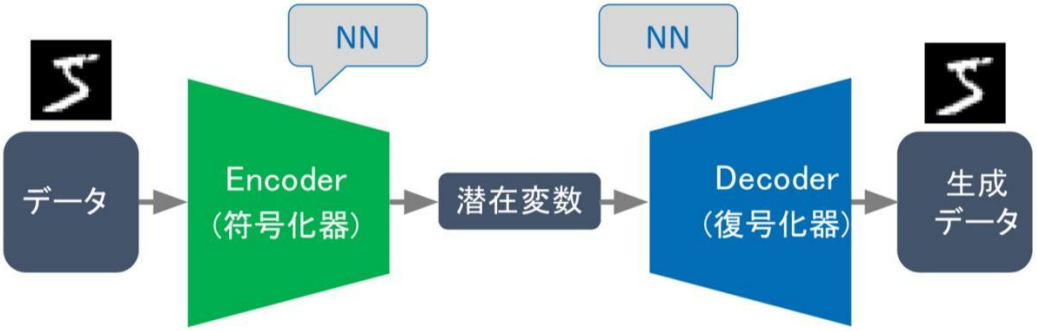
次元削減手法①

【問題】
重要な情報を保ったままデータを低次元に圧縮する次元削減に用いられる手法として正しいものを選べ。

【選択肢】
(a)非線形SVM
(b)Auto Encoder
(c)Adaboost
(d)k近傍法

【解答】
(b)Auto Encoder

【解説】
AEは、ニューラルネットワークを用いた次元削減の手法として知られている。以下の図はAEのアーキテクチャを表したものであり、潜在変数が低次元に圧縮した入力データの情報を保持している。



以下、各選択肢の手法を簡単に説明する。
非線形SVM：非線形データを超平面で分割できるような空間に写像した後に、SVMと同じ方法でクラス分類を行う手法。
Adaboost：いくつかの弱学習器を直列につなげて学習を行うブースティングアルゴリズムの一種。ある弱学習器で学習して分類した結果、誤った分類をしたものに対する重みを増やして次の弱学習器ではその誤分類したデータを優先的に分類する様に学習を行う手法。
k-近傍法：未知のデータの分類をする際、そのデータから見て距離の近いk個のデータの属しているクラスを見て、多数決の様にそのk個の中で属しているクラスが多いクラスに分類する手法。

次元削減手法②

【問題】
PCAは機械学習において代表的な次元削減手法の一つとして知られている。PCAに関する説明として誤っているものを選べ。

【選択肢】
(a)主成分は与えられたデータを特異値分解することによって求めることが可能である。
(b)与えられたデータにおける分散共分散行列の固有値分解によって得られるi番目に大きい固有値に対する固有ベクトルv_iが第i主成分にあたる。
(c)n次元のデータをk次元に変換したい際は分散が小さい方から主成分をk個選んで次元削減を行う。(n>k)
(d)与えられたデータが二次元データであるとき、第1主成分と第2主成分は直交する。

【解答】
(c)n次元のデータをk次元に変換したい際は分散が小さい方から主成分をk個選んで次元削減を行う。(n>k)

【解説】
主成分分析における次元削減は分散が大きい主成分から順番に選んで行うので誤り。
主成分分析はデータの次元削減や可視化などを目的とした教師なし学習の手法である。
以下に主成分分析による次元削減の手順とその説明を示す。
①　射影したデータの分散が最大になるような軸を探すことで、情報を多く持つ合成変数である主成分を見つけ出す。詳しい説明は省略するが、これは与えられたデータの分散共分散行列の固有値問題を解くことに帰結する。
②　①で求めた固有値はそれぞれの主成分の分散にあたり、i番目に大きい固有値に対する固有ベクトルv_i が第i主成分にあたる。なお、分散共分散行列は対称行列なので、その固有ベクトルで表される主成分同士は直交する。
③固有値の大きい方からk個の固有ベクトルを選び、元データとk個の固有ベクトルからなる行列の積を取ることで情報を保ったままk次元に次元圧縮することができる。
また主成分分析と特異値分解の間には関連があり、与えられたデータを特異値分解したとき、第i主成分はi番目に大きい特異値に対応する右特異ベクトルv_iと正負の符号を除いて一致する。

オートエンコーダ

【問題】
オートエンコーダに関する説明として誤っているものを選べ。

【選択肢】
(a)出力は以下の式で表され、自分自身への回帰問題を考える際はデコーダにおける活性化関数には恒等写像が用いられることが多い。
(x:入力、y:出力、f1:エンコーダにおける活性化関数、f2:デコーダにおける活性化関数、w1,b1:エンコーダにおける重みとバイアス、w2,b2:デコーダにおける重みとバイアス)

$$y = f_2 f_1(W_2(W_1 x + b_1) + b_2)$$

(b)異常検知やクラスタリングを目的とした機械学習手法の一つとして用いられる。
(c)制限ボルツマンマシンなどと同様にニューラルネットワークを用いた次元削減手法として用いられている。
(d)潜在変数の次元が入力データの次元より小さくなるように構成されたオートエンコーダを不完備なオートエンコーダと呼ぶ。

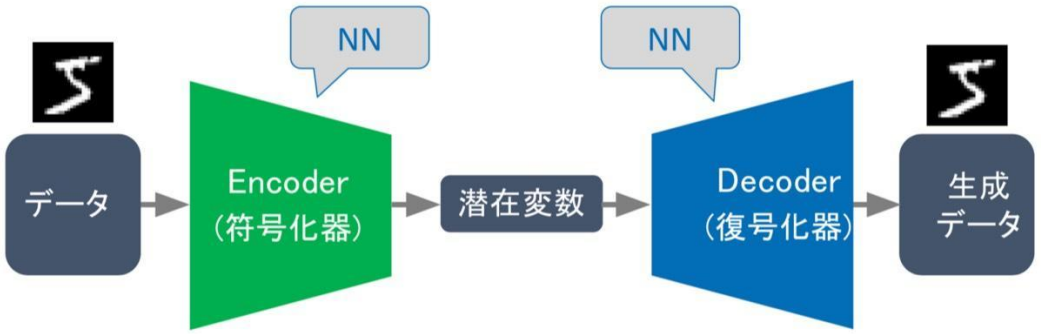
【解答】
(a)出力は以下の式で表され、自分自身への回帰問題を考える際はデコーダにおける活性化関数には恒等写像が用いられることが多い。
(x:入力、y:出力、f1:エンコーダにおける活性化関数、f2:デコーダにおける活性化関数、w1,b1:エンコーダにおける重みとバイアス、w2,b2:デコーダにおける重みとバイアス)

$$y = f_2 f_1(W_2(W_1 x + b_1) + b_2)$$

【解説】
オートエンコーダにおける正しい出力の式は以下の通りであるため、誤り。

$$y = f_2(W_2 f_1(W_1 x + b_1) + b_2)$$

オートエンコーダは、エンコーダとデコーダをニューラルネットワークとして入力と同じものを出力するように学習を行うモデルである。構造は以下の画像の通り。



入力をエンコーダに通すと潜在変数が出力され、潜在変数を入力としてデコーダから出力が得られる。自分自身への回帰問題を考える際はf2には恒等関数を用いることが多い。
利用例としてクラスタリング、異常検知、次元削減などに用いられる。
また、潜在変数の次元が入力データの次元より小さくなるように構成されたオートエンコーダを不完備なオートエンコーダと呼ぶ。逆に潜在変数の次元が入力データの次元より大きくなるように構成されたオートエンコーダは過完備なオートエンコーダと呼ばれる。

Reconstruction Error

【問題】
VAEはAEの派生モデルであり、ガウス分布に潜在変数を押し込むことで生成データに連続性とランダム性を持たせたものである。VAEにおいて最小化する損失関数は、Reconstruction　ErrorとKLダイバージェンスによって表される正則化項で表現される。その理論式として正しいものを選べ。(X:訓練データ、z:ノイズ、q(z|X):エンコーダ、p(X|z):デコーダ、p(z):デコーダの事前分布)

【選択肢】
(a)
$$D_{\mathrm{KL}}\left[p(z)\middle\|q(z|X)\right] - \mathbb{E}_{q(z|X)}\left[p(X|z)\right]$$
(b)
$$D_{\mathrm{KL}}\left[p(z)\middle\|q(z|X)\right] - \mathbb{E}_{q(z|X)}\left[\log p(X|z)\right]$$
(c)
$$D_{\mathrm{KL}}\left[q(z|X)\middle\|p(z)\right] - \mathbb{E}_{q(z|X)}\left[\log p(X|z)\right]$$
(d)
$$D_{\mathrm{KL}}\left[q(z|X)\middle\|p(z)\right] - \mathbb{E}_{q(z|X)}\left[p(X|z)\right]$$
【解答】
(c)
$$D_{\mathrm{KL}}\left[q(z|X)\middle\|p(z)\right] - \mathbb{E}_{q(z|X)}\left[\log p(X|z)\right]$$

【解説】
VAEにおける正しい損失関数を選ぶ問題。
第一項は Reconstruction Error と呼ばれ、入力の元データと生成データの一致度を意味する。この値が大きければ出力が入力を上手く再現出来ていると言える。
なお、この式は対数尤度log p(x)の最大化から導出される。
第二項はKLダイバージェンスを用いた正則化項である。VAEにおける損失関数計算ではデコーダの事前分布p(z)(つまり潜在空間の分布)はN(0,I)として、エンコーダq(z|X)はN(μ,Σ)として仮定されるので、エンコーダq(z|x)=N(μ,Σ)がp(z)=N(0,I)とどれだけ近いかを表している。(μ,Σ:エンコーダの出力、I:単位行列)
第二項が小さくなればこの正則化項によってN(μ,Σ)はN(0,I)に近づき、μとΣによって生成される潜在変数zが標準正規分布に近づく。
なおKLダイバージェンスには非対称性があり、ある確率分布p,q(p≠q)を考えた時、
$$D_{\mathrm{KL}}\left[q(x)\middle\|p(x)\right] \neq D_{\mathrm{KL}}\left[p(x)\middle\|q(x)\right]$$
という関係が成り立つ。

【参考】
https://tips-memo.com/vae-pytorch
https://nzw0301.github.io/assets/pdf/vae.pdf
https://deepblue-ts.co.jp/image-generation/variational-autoencoder-part1/

Reparametorization Trick

【問題】
VAEでは潜在変数zをガウス分布からランダム生成すると、誤差逆伝播計算が不可能であるという問題を解決するためにReparametorization　Trickと呼ばれる手法が用いられている。この手法を用いたVAEにおける潜在変数zを表した式としてふさわしいものを選べ。(N:ガウス分布、μ:エンコーダにおける出力(潜在変数生成の際にガウス分布における平均として扱われるパラメータ)、σ^2:エンコーダにおける出力(潜在変数生成の際にガウス分布における分散として扱われるパラメータ)、I:単位行列)

【選択肢】
(a)

$$z = \mu + \sigma^2 * \epsilon \quad (\epsilon \sim \mathcal{N}(0, I))$$

(b)

$$z = \mu + \sigma * \epsilon \quad (\epsilon \sim \mathcal{N}(0, I))$$

(c)

$$z = \mu + \sigma * \epsilon \quad (\epsilon \sim \mathcal{N}(\mu, \sigma^2))$$

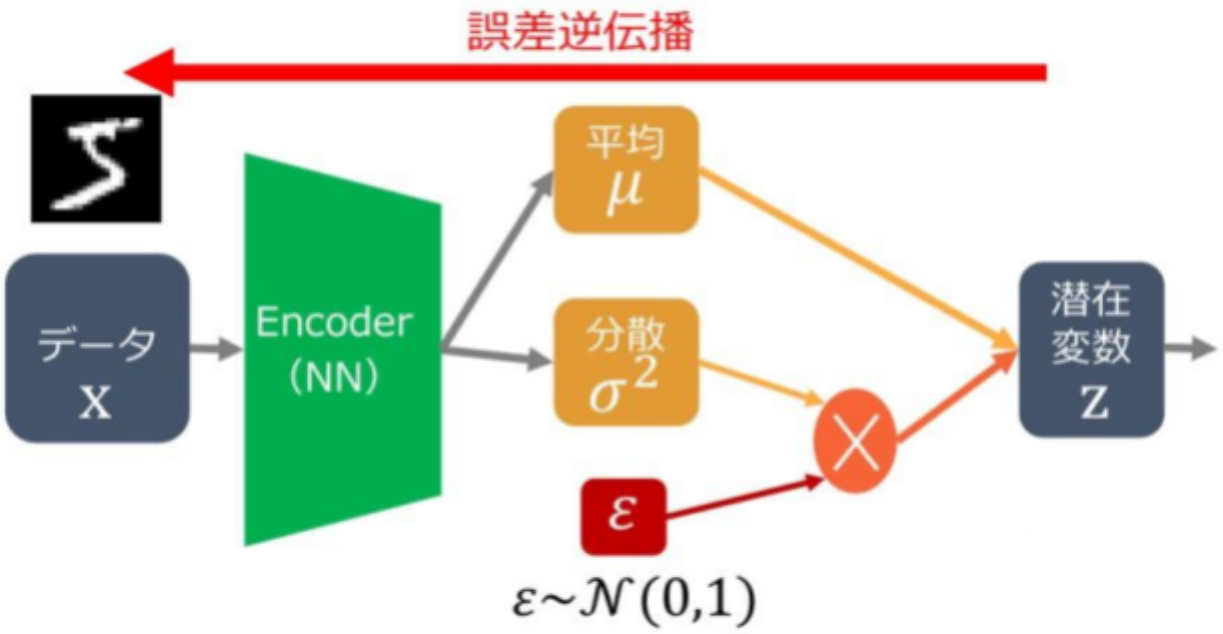
(d)

$$z = \mu + \sigma^2 * \epsilon \quad (\epsilon \sim \mathcal{N}(\mu, \sigma^2))$$

【解答】
(b)

$$z = \mu + \sigma * \epsilon \quad (\epsilon \sim \mathcal{N}(0, I))$$

【解説】
ガウシアンノイズε~N(0,I)を用いて解答の式のようにzを計算する。このように計算することでガウス分布からサンプリングした値を得ることができ、εを記憶しておけば誤差逆伝播計算も可能となる。Reparametorization Trickを用いたVAEの構造は以下の様な図で表される。



GANの特徴

【問題】
画像生成分野において非常に有用とされているモデルの一つであるGAN(敵対的生成ネットワーク)に関する説明として正しいものを選べ。(z:ノイズ、x:訓練データ、G(z):Generatorにおける出力、D(x):Discriminatorにおける出力、p_data:訓練データの分布、p_z:ノイズの分布)

【選択肢】
(a)学習が上手く行われて理想的な性能を獲得した場合、G(z)の出力はxの分布に近付き、D(x)の出力は1に近づく。
(b)Generatorは、Discriminatorが訓練データと生成データを上手く判別できないようなデータを生成するように学習を行う。Generatorにおける目的関数は以下の式で表される。

$$\min_G \mathbb{E}_{z \sim p_z(x)}[log(1 - D(G(z)))]$$

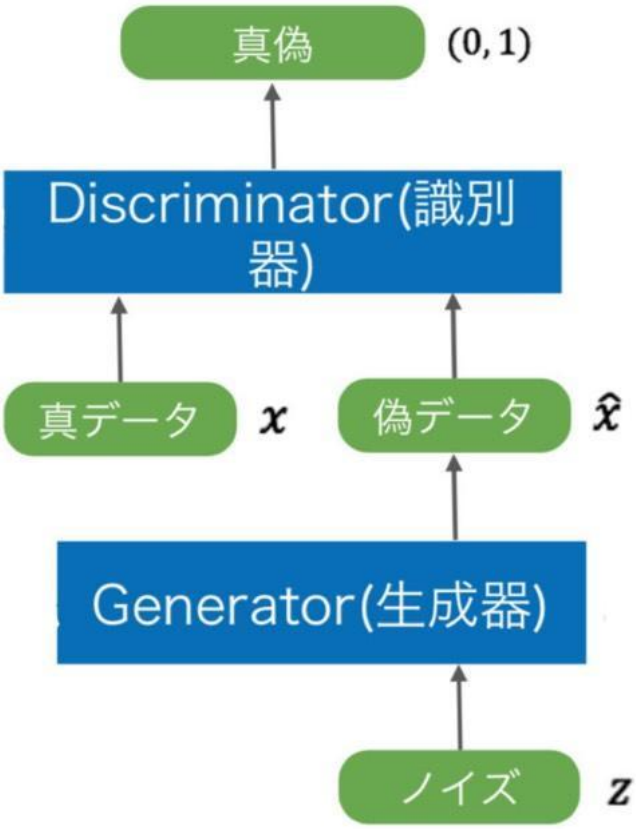
(c)Unrolled GANはGANの学習が困難である要因の一つであるmode collapseの対策として有用とされている。
(d)Discriminatorは訓練データと生成データの判別が出来るように学習を行う。Discriminatorにおける目的関数は以下の式で表される。

$$\min_D \mathbb{E}_{x \sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z \sim p_z(x)}[log(1 - D(G(z)))]$$

【解答】
(b)GeneratorはDiscriminatorが訓練データと生成データを上手く判別できないようなデータを生成するように学習が行われる。Generatorにおける目的関数は以下の式で表される。

$$\min_G \mathbb{E}_{z \sim p_z(x)}[log(1 - D(G(z)))]$$

【解説】
GANのアーキテクチャは以下の図の通りである。



Generatorはノイズから画像データを生成し、Discriminatorは入力が真のデータ(訓練データ)である確率を出力する。
Generatorは訓練データと同じような分布を出力出来るように学習を行うため、D(G(z))の値が1になる様に学習が進む。よって解答の式は正しい式である。
各選択肢の誤っている理由については以下の通り。
「学習が上手く行われて理想的な性能を獲得した場合、G(z)の出力はxの分布に近付き、D(x)の出力は1に近付く。」
理想的な性能ならば、Discriminatorは訓練データと生成データの判別が出来なくなっているはずであるため、D(x)の出力は0.5に近づく。
「Discriminatorは訓練データと生成データの判別が出来るように学習を行う。Discriminatorにおける目的関数は以下の式で表される。」
正しい式は以下の通りである。選択肢では最小となるDが最適とされているが、最大化するDが最適であるため、誤り。GeneratorとDiscriminatorの目指す性能と式の意味を合わせて考えると理解しやすい。

$$\max_D \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(x)}[log(1-D(G(z)))]$$

なお、GeneratorとDiscriminatorの二つの目的関数を組み合わせたGAN全体としての目的関数は以下の通りである。
それぞれの項の意味と共に式の意味を考える。
まず、第一項について。D(x)は訓練データを入力とした時にDiscriminatorが訓練データであると判別する確率を表すので、Discriminatorを訓練データと生成データの判別が出来るように学習させたいことを考えるとlogD(x)の値を大きくすれば良いことが分かる。よって最大化するようなDiscriminatorが最適である。
次に第二項について。D(G(z))は生成データ(G(z))を入力とした時にDiscriminatorが訓練データであると判別する確率を表すので、Discriminatorを訓練データと生成データの判別が出来るように学習させたいことを考えるとlog(1-D(G(z)))を大きくすればよく、GeneratorをDiscriminatorが訓練データであると判断してしまうようなデータが生成できるように学習させたいことを考えるとlog(1-D(G(z)))を小さくすればよいことが分かる。よって最大化するようなDiscriminator、最小化するようなGeneratorが最適になる。
以上より、式全体としては最大化するDiscriminator、最小化するGeneratorが最適となる。

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}(x)}[logD(x)]+\mathbb{E}_{z\sim p_z(x)}[log(1-D(G(z)))]$$

「Unrolled GANはGANの学習が困難である要因の一つであるmode collapseの対策として有用とされている。」
mode collapseがGANの学習が困難である要因の一つであることは正しいが、Unrolled GANはmode collapseへの対策として用いられる手法ではないため誤り。Unrolled GANはGeneratorよりもDiscriminatorの方が学習が早く進んでしまう問題への対策としての手法である。

GANの目的関数

【問題】
画像生成分野において非常に有用とされているモデルの一つであるGAN(敵対的生成ネットワーク)において用いられる目的関数として正しいものを選べ。(z:ノイズ、x:訓練データ、G(z):Generatorにおける出力、D(x):Discriminatorにおける出力、p_data:訓練データの分布、p_z:ノイズの分布)

【選択肢】
(a)

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(x)}[\log(1 - D(G(z)))]$$

(b)

$$\max_G \min_D \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(x)}[\log(1 - D(G(z)))]$$

(c)

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}(x)}[\log(1 - D(x))] + \mathbb{E}_{z\sim p_z(x)}[\log(1 - D(G(z)))]$$

(d)

$$\max_G \min_D \mathbb{E}_{x\sim p_{data}(x)}[\log(1 - D(x))] + \mathbb{E}_{z\sim p_z(x)}[\log(1 - D(G(z)))]$$

【解答】
(a)

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(x)}[\log(1 - D(G(z)))]$$

【解説】
GANの正しい目的関数を選ぶ問題。Generatorは、Discriminatorが訓練データと生成データを判別できない様なデータを生成できるように学習を行う。Discriminatorは訓練データと生成データを判別できるように学習を行う。解答の式について、それぞれの項の意味と共に式の意味を考える。
まず、第一項について。D(x)は訓練データを入力とした時にDiscriminatorが訓練データであると判別する確率を表すので、Discriminatorを訓練データと生成データの判別が出来るように学習させたいことを考えるとlogD(x)の値を大きくすれば良いことが分かる。よって最大化するようなDiscriminatorが最適である。
次に第二項について。D(G(z))は生成データ(G(z))を入力とした時にDiscriminatorが訓練データであると判別する確率を表すので、Discriminatorを訓練データと生成データの判別が出来るように学習させたいことを考えるとlog(1-D(G(z)))を大きくすればよく、GeneratorをDiscriminatorが訓練データであると判断してしまうようなデータが生成できるように学習させたいことを考えるとlog(1-D(G(z)))を小さくすればよいことが分かる。よって最大化するようなDiscriminator、最小化するようなGeneratorが最適になる。
以上より、式全体としては最大化するDiscriminator、最小化するGeneratorが最適となる。
この様に意味を式と照らし合わせて考えると解答の式が正しく、他の選択肢ではGeneratorとDiscriminatorが目指す性能を得られないことが分かる。

DCGAN

【問題】
畳み込みニューラルネットワークを用いたGANの代表的なモデルであるDCGANのDiscriminatorにおいて用いられている活性化関数の式としてふさわしいものを選べ。

【選択肢】
(a)

$$f(x) = \begin{cases} x & (x > 0) \\ -x & (x \leq 0) \end{cases}$$

(b)

$$f(x) = \begin{cases} x & (x > 0) \\ \alpha x & (x \leq 0) \end{cases}$$

(αは学習可能パラメータ)

(c)

$$f(x) = \begin{cases} x & (x > 0) \\ e^x - 1 & (x \leq 0) \end{cases}$$

(d)

$$f(x) = \begin{cases} x & (x > 0) \\ 0.01x & (x \leq 0) \end{cases}$$

【解答】
(d)

$$f(x) = \begin{cases} x & (x > 0) \\ 0.01x & (x \leq 0) \end{cases}$$

【解説】
Discriminatorで用いられている活性化関数はLeaky ReLUである。
DCGANは畳み込みニューラルネットワークを用いたGANの派生モデルである。
Generatorでは、転置畳み込み、バッチ正規化、ReLUの三つのフローが繰り返される様な構造をしており(出力層はtanhが用いられている)、
Discriminatorはシンプルな畳み込みニューラルネットワークと同じ様な構造をしているが、活性化関数にLeaky ReLUが用いられていることが特徴として挙げられる。通常のReLUでは入力が0未満の時は出力が0になってしまうが、入力が0未満の時は負の値を出力するLeaky ReLUを使うことによって勾配が0になってしまうことを防ぎ、学習を安定させている。
どちらにもpoolingは用いられていないことも特徴の一つである。
なお、他の選択肢の式は以下の通り。選択肢の式は全てReLU派生の活性化関数である。

$$AVR : f(x) = \begin{cases} x & (x > 0) \\ -x & (x \leq 0) \end{cases}$$

$$ELU : f(x) = \begin{cases} x & (x > 0) \\ e^x - 1 & (x \leq 0) \end{cases}$$

$$PReLU : f(x) = \begin{cases} x & (x > 0) \\ \alpha x & (x \leq 0) \end{cases}$$

(αは学習可能パラメータ)

【参考】
https://blog.negativemind.com/2019/09/07/deep-convolutional-gan/
https://www.slideshare.net/HiroyaKato1/gandcgan-188544721

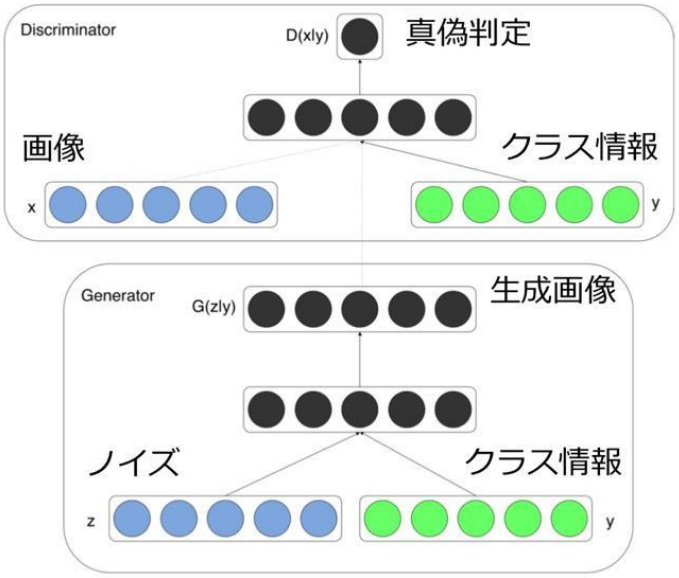
GANの発展モデル

【問題】
GANの一種であるConditional GANやその発展モデルに関する説明として正しいものを選べ。

【選択肢】
(a)Conditional GANの派生モデルであるInfoGANでは、ラベル付けしたデータを用いるのに加えて、生成画像分布に大きな影響を持つ潜在変数の獲得のために潜在変数と画像分布の相互情報量を評価関数に導入している。
(b)Generatorの入力にはノイズベクトルと条件ベクトルを、Discriminatorの入力には画像データと、画像と同じサイズに変換した条件ベクトルを用いる。
(c)Discriminatorは入力データの真偽の判定(訓練データか生成データか)とその画像のクラス分類を行う。
(d)従来のDCGANなどのモデルも生成する画像のクラスを指定することは可能であったが、Conditional GANによってクラス指定を伴った生成画像の質が向上した。

【解答】
(b)Generatorの入力にはノイズベクトルと条件ベクトルを、Discriminatorの入力には画像データと、画像と同じサイズに変換した条件ベクトルを用いる。

【解説】
Conditional GANはデータのクラス情報もGeneratorとDiscriminatorの入力に利用することで、利用生成画像の質とクラス識別性を向上させたモデルである。アーキテクチャは以下の図の通り。
Generator、Discriminatorの両方の入力に条件ベクトルを用いることが大きな特徴。
実装上Discriminatorにおける入力は画像のサイズに合わせる必要があるので、条件ベクトルは画像のデータサイズ(width*height)として入力される。



「Conditional　GANの派生モデルであるInfoGANでは、ラベル付けしたデータを用いるのに加えて、生成画像分布に大きな影響を持つ潜在変数の獲得のために潜在変数と画像分布の相互情報量を評価関数に導入している。」
後半の説明はInfo GANについての正しい説明であるが、Info GANはラベル付けされたデータ無しに生成画像の制御を行ったモデルであるため、この説明は誤り。
「Discriminatorは入力データの真偽の判定(訓練データか生成データか)とその画像のクラス分類を行う。」
この説明はACGANの特徴についての説明であるため、誤り。
「従来のDCGANなどのモデルも生成する画像のクラスを指定することは可能であったが、Conditional GANによってクラス指定を伴った生成画像の質が向上した。」
DCGANは生成画像の質の向上させた手法であり、クラス指定は出来ないため誤り。

WaveNetの結合確率

【問題】
音声合成の分野における深層学習を用いた代表的な手法であるWaveNetでモデル化される結合確率の式を選べ。(x={x_1,...x_T}:開始時刻1から終了時刻Tまでにおける入力波形データ)

【選択肢】
(a)

$$p(\boldsymbol{x}) = \prod_{t=1}^{T} p(x_t|x_1, \ldots, x_{t-1})$$

(b)

$$p(\boldsymbol{x}) = \sum_{t=1}^{T} p(x_t|x_{t-1})$$

(c)

$$p(\boldsymbol{x}) = \prod_{t=1}^{T} p(x_t|x_{t-1})$$

(d)

$$p(\boldsymbol{x}) = \sum_{t=1}^{T} p(x_t|x_1, \ldots, x_{t-1})$$

【解答】
(a)

$$p(\boldsymbol{x}) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$$

【解説】
WaveNetは自身の過去の出力に依存して次の時刻の出力を行う自己回帰モデルの一種である。自己回帰モデルは複雑な確率モデルを単純な確率分布の積に分解でき、複雑な確率分布を学習しやすいという特徴がある。自己回帰モデルにはステップ数が大きくなってしまう問題があったが、WaveNetではCNNを用いることで学習の効率化を行っている。

【参考】
https://xtech.nikkei.com/dm/atcl/mag/15/00144/00023/

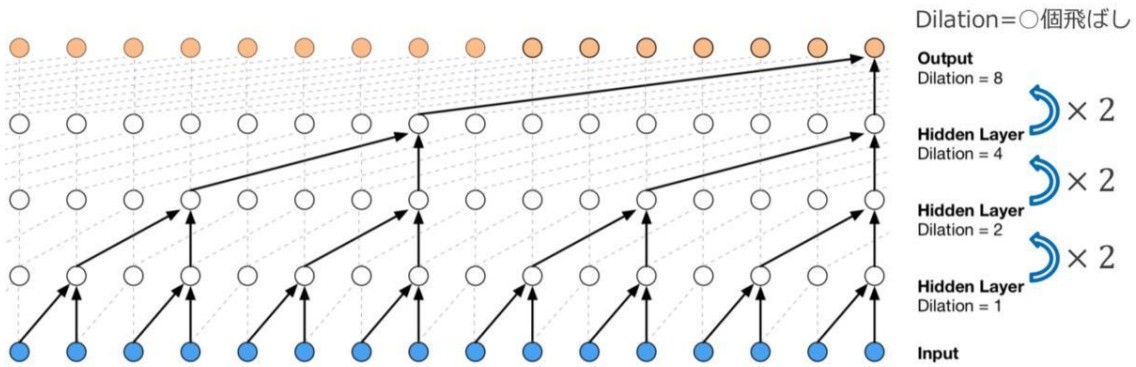
WaveNetの特徴

【問題】
WaveNetに関する説明として誤っているものを選べ。

【選択肢】
(a)単純なRNNと異なり、回帰的な接続が存在しないため、並列計算が行えるので出力を高速に計算することが可能である。
(b)層が深くなるにつれて畳み込み処理を行うノードを離す度合い(Dilation)を小さくするDilated causal convolutionを用いている。層が一つ深くなるにつれてDilationは1/2倍ずつ小さくなっている。
(c)Residual Blockとskip-connectionが勾配消失を避けるために用いられている
(d)Dilated causal convolutionによって、少ないパラメータで広い範囲の情報を処理することが可能となっている。

【解答】
(b)層が深くなるにつれて畳み込み処理を行うノードを離す度合い(Dilation)を小さくするDilated causal convolutionを用いている。層が一つ深くなるにつれてDilationは1/2倍ずつ小さくなっている。

【解説】
Dilated causal convolutionでは層が深くなるにつれてDilationは二倍ずつ大きくなるので誤り。WaveNetにおける最大の特徴はcausal convolutionとdilated convolutionという二つの手法を組み合わせたDilated causal convolutionを用いていることである。causal convolutionとは、通常の畳み込みを音声処理に適応しようとすると本来得られない未来の情報を参照してしまうため考案された過去の情報のみを参照して行われる畳み込み処理である。RNNと異なり、回帰結合を持たないため、並列計算が可能であり学習時間が少ないというメリットも持っている。しかし、通常のcausal convolutionでは、受容野を増やすためにはパラメータ数が膨大になってしまうという問題があった。これを解決するためにカーネルの要素間を0パディングすることで広い受容野を畳みこむ手法であるDilated convolutionをcausal convolutionに導入し、層が深くなるにつれて畳み込み処理を行うノードを離す度合い(Dilation)を大きくしていく以下の図の様なDilated causal convolutionが考案された。Dilationは一層上がるごとに2倍ずつ大きくなる様に設計されている。また、勾配消失問題への対策としてResidual Blockとskip-connectionが用いられていることもWaveNetの特徴の一つである。



【参考】
https://deepsquare.jp/2020/04/wavenet/#:~:text=WaveNet%E3%81%A8%E3%81%AF%E3%80%81%E9%87%8F%E5%AD%90%E5%8C%96,%EF%BC%8F16bit%E3%80%8D%E3%81%AA%E3%81%A9%E3%81%AE%E3%81%93%E3%81%A8%E3%80%82

方策勾配定理

【問題】
強化学習における方策ベースでの代表的な学習手法である方策勾配法において、方策勾配定理を用いて計算される方策勾配∇J(θ)(目的関数の方策パラメータθに関する偏微分)の式として正しいものを選択肢から選べ。(π:方策、J:目的関数、Q:行動価値関数、a:行動、s:状態)

【選択肢】
(a)

$$\nabla_\theta J(\theta) = E_{\pi_\theta}[\nabla_\theta Q^{\pi_\theta}(s,a) log \pi_\theta(a|s)]$$

(b)

$$\nabla_\theta J(\theta) = E_{\pi_\theta}\left[\frac{\nabla_\theta Q^{\pi_\theta}(s,a)}{log \pi_\theta(a|s)}\right]$$

(c)

$$\nabla_\theta J(\theta) = E_{\pi_\theta}\left[\frac{Q^{\pi_\theta}(s,a)}{\nabla_\theta log \pi_\theta(a|s)}\right]$$

(d)

$$\nabla_\theta J(\theta) = E_{\pi_\theta}[Q^{\pi_\theta}(s,a) \nabla_\theta log \pi_\theta(a|s)]$$

【解答】
(d)

$$\nabla_\theta J(\theta) = E_{\pi_\theta}[Q^{\pi_\theta}(s,a)\nabla_\theta log\pi_\theta(a|s)]$$

[解説]
方策勾配法は、方策勾配定理を用いて方策勾配を計算して確率的勾配法に従って方策パラメータθを更新する手法である。Jを最大化するθを求めることで最適方策を求めている。
方策勾配定理の導出は以下のようにされる。

p を状態遷移確率、r を報酬とする。また、方策π_θによる行動と状態の列を
$\tau = (s_0, a_0, \ldots a_t, s_{t+1})$ とすると、

$$p_\theta(\tau) = p(s_0)\prod_{t=0}^{\tau} p(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t)$$

$$\begin{aligned}
\nabla J(\theta) &= \nabla_\theta V(s_0) \\
&= \nabla_\theta \sum_\tau r(s_t, a_t) p_\theta(\tau) \\
&= \sum_\tau r(s_t, a_t) \nabla_\theta p_\theta(\tau) \\
&= \sum_\tau r(s_t, a_t) p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)
\end{aligned}$$

ここで、

$$\begin{aligned}
\nabla_\theta \log p_\theta(\tau) &= \nabla_\theta \log \left(p(s_0)\prod_{t=0}^{\tau} p(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t) \right) \\
&= \nabla_\theta \left(\log p(s_0) + \sum_{t=0}^{T} \log p(s_{t+1}|s_t, a_t) + \log \pi_\theta(a_t|s_t) \right) \\
&= \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)
\end{aligned}$$

であるので、

$$\begin{aligned}
\nabla J(\theta) &= \sum_\tau r(s_t, a_t) p_\theta(\tau) \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \\
&= \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t) Q^\pi(s_t, a_t)]
\end{aligned}$$

この定理を用いることによって、Q値を用いて累積報酬を増加させる方策の勾配を求めることが可能となっている。

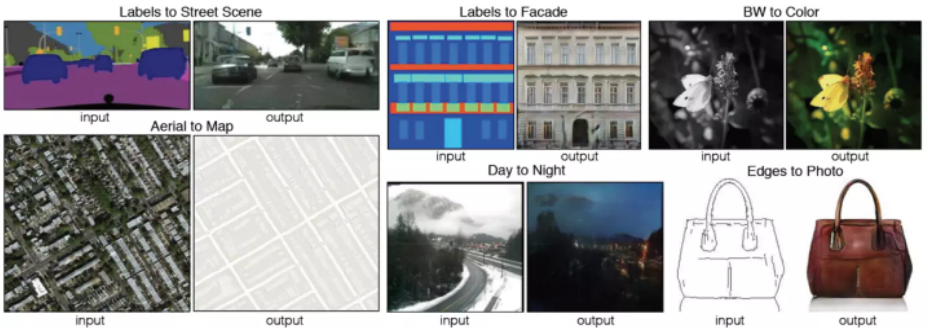
pix2pix

【問題】
敵対的生成ネットワーク(GAN)を用いて入力画像を別の特徴を持った出力画像に変換を行う
pix2pixと呼ばれる手法がある。
このpix2pixの応用タスクとして不適切なものを選べ。

【選択肢】
(a)線画画像の着色
(b)ピクセルごとのラベルデータに対する画像の生成
(c)昼の画像を夜の画像に変換する。
(d)画像内の物体領域の抽出

【解答】
(d)画像内の物体領域の抽出

【解説】
pix2pixはGANを用いた画像変換モデルの基本となるモデルであり、Conditional　GANの一種でもある。
Conditional　GANでは条件ベクトルと画像のペアを学習データとしていたが、pix2pixは条件画像と画像のペアを学習データとし、その対応関係を学習する。
このpix2pixではピクセルごとにラベルづけされたデータ(下図左上など)に対して画像を生成することもできる。下図のLabels　to　Street　Sceneを例に挙げると青色でラベル付けされた領域は車を、紫色でラベル付けされた領域は道路を表しており、ラベルの領域ごとに本物の画像のような物体を描画できていることが分かる。
しかし、物体領域の抽出などの物体検出のタスクはpix2pixでは行うことができないため回答は誤り。
また、pix2pix のGeneratorにはU-Netが採用されている。そのため、エンコーダとデコーダの特徴を結合するスキップ接続を持っており、エンコーダ部分で失われた局所的な空間情報を補完することができる。
他にも、U-Netの複数層にDropoutを設けることで各層にノイズベクトルzを与えている。このノイズzと条件画像を入力として画像を生成している。



DQN

【問題】
DQN(Deep Q network)は強化学習における深層学習を用いた代表的な学習手法である。DQNにおいて学習の安定のために施されている様々な工夫についての説明として正しいものを選べ。

【選択肢】
(a)与えられたデータの時系列的な相関が強いと特定のパターンにのみ過学習してしまう場合があるため、経験した(s,a,R(s,a),s`)の四つの情報をメモリに蓄積し、学習の際にはその情報をランダムにサンプリングする。(s:状態、a:行動、R:報酬、s`:次の状態)
(b)報酬の値を正ならば1、負ならば-1として固定することでQ値の急激に変化できるようにして学習を安定させている。この手法には報酬の大小の情報が重要なタスクを扱う際には望ましい結果を得られないこともあるというデメリットも存在する。
(c)教師ラベルの役割を担うQ Networkの出力が変化すると、Target Q Networkの収束先が定まらなくなってしまう問題を解決したのがFixed Target Q Networkである。
(d)Fixed Target Q Networkを用いたDQNの損失関数は以下の様な式で表される。(R:報酬、Q:行動価値関数、s:状態、a:行動、γ:ハイパーパラメータ、「-」の上付き添字:固定されたパラメータ、「`」の上付き添字:遷移後におけるパラメータ)

$$E\left[\frac{1}{2}\left(R(s,a)+\gamma \underset{a'}{max} Q_{target}(s',a';\theta^{-})-Q(s,a;\theta^{-})\right)^{2}\right]$$

【解答】
(a)与えられたデータの時系列的な相関が強いと特定のパターンにのみ過学習してしまう場合があるため、経験した(s,a,R(s,a),s`)の四つの情報をメモリに蓄積し、学習の際にはその情報をランダムにサンプリングを行う。(s:状態、a:行動、R:報酬、s`:次の状態)

【解説】
解答の説明はExperience Replayについての正しい説明である。
「報酬の値を正ならば1、負ならば-1として固定することでQ値の急激に変化できるようにして学習を安定させている。この手法には報酬の大小の情報が重要なタスクを扱う際には望ましい結果を得られないこともあるというデメリットも存在する。」
この説明はReward Clippingについてのものであるが、「Q値の急激な変化」という部分が誤り。Q値の急激な変化を抑えて学習を安定させる手法である。
「教師ラベルの役割を担うQ Networkの出力が変化すると、Target Q Networkの収束先が定まらなくなってしまう問題を解決したのがFixed Target Q Networkである。」
この説明はFixed Target Q Networkについての誤った説明である。教師ラベルの役割を持つのはTarget Q Networkの出力であり、Fixed Target Q NetworkはQ Networkの収束を安定させるためにTarget Q Networkのパラメータを固定する手法である。
「Fixed Target Q Networkを用いたDQNの損失関数は以下の様な式で表される。(R:報酬、Q:行動価値関数、s:状態、a:行動、γ:ハイパーパラメータ、「-」の上付き添字:固定されたパラメータ、「`」の上付き添字:遷移後におけるパラメータ)」
上述の通りFixed Target Q Networkは、Target Q Networkのパラメータを固定する手法であるが、この選択肢の式ではTarget Q NetworkとQ Network両方のパラメータを固定してしまっているため、誤り。正しい式は以下の通りである。

$$E\left[\frac{1}{2}(R(s,a)+\gamma \underset{a'}{max} Q_{target}(s',a';\theta^{-})-Q(s,a;\theta))^2\right]$$

AlphaGo

【問題】
AlphaGoについての説明として誤っているものを選べ。

【選択肢】
(a)モデル内で用いられているニューラルネットワークは全て畳み込みニューラルネットワークである。
(b)教師あり学習と強化学習の両方を用いてモデル全体の学習が行われている。
(c)value networkは現状の盤面における最善手の確率分布を算出し、policy networkは現状の盤面における勝率を算出する。
(d)与えられた状態に対して、モンテカルロ木探索法をベースとしたアルゴリズムに従って行動を選択する。

【解答】
(c)value networkは現状の盤面における最善手の確率分布を算出し、policy networkは現状の盤面における勝率を算出する。

【解説】
value networkは現状の盤面における勝率を算出し、policy networkは現状の盤面における最善手を算出するので、誤り。
Alpha Goは、Rollout policy, SL policy network, RL policy network, value networkの四つのモデルを用いて学習が行われる。シミュレーションを行う際にモンテカルロ木探索法を用いることも特徴。これは与えられた状態において勝率の高い選択を中心的に選択しながら探索出来ていない部分を減らすために用いられている。
Rollout policy:ロジスティック回帰を用いて現状の盤面から最善手を予測。人間のデータで教師あり学習を行う。高速に処理できるモデルを用いることでシミュレーションの実行回数を増やしている。
SL policy network:CNNで現状の盤面から最善手を予測。人間のデータで教師あり学習を行う。
RL policy network:SL policy networkを初期値として、強化学習を方策勾配法で行って自己対戦しながらネットワークの更新を行う。value networkのための教師データを作成する。
Value network:CNNで現状の盤面から勝率を算出。RL policy networkによって作成されたデータで学習を行う。
簡単な学習の流れとしては、人間のプレイデータを利用してSL policy networkを作り、このSL policy networkを初期値としてAlpha Go同士での対戦を行ってRL policy networkの強化学習を行いつつ新しいデータを作成。この作成したデータでvalue networkを作成する。これらの学習を行う際にRollout policyとモンテカルロ木探索法を組み込んでいる。詳しい学習の流れについては複雑なので割愛する。

【参考】
https://www.slideshare.net/suckgeunlee/aialphago
https://qiita.com/gifucom17/items/3096ac60522f8b815a32

表現学習

【問題】
次の選択肢の中から誤った説明であるものを選べ。

【選択肢】
(a)転移学習では、他のタスクで学習済みモデルの重みを初期値として、新しいタスクのために追加された層も含めてモデル全体のパラメータの学習が行われる。
(b)機械学習のタスクにおいて時間経過で目標データの形質が変化することをコンセプトドリフトと呼ぶ。例としてはマルウェア認識などが挙げられる。
(c)ラベル付きの犬の画像を一切与えずに、犬についての情報を持つテキストデータを学習させたモデルを犬の画像認識タスクに適応するようなケースをゼロショット学習と呼ぶ。
(d)マルチタスク学習とは単一のモデルで複数のタスクの学習を行うことである。例として、YOLOなどの物体検出モデルはクラス分類と領域特定の二つのタスクを解いている。

【解答】
(a)転移学習では、他のタスクで学習済みモデルの重みを初期値として、新しいタスクのために追加された層も含めてモデル全体のパラメータの学習が行われる。

【解説】
この説明はファインチューニングについてのものであるので誤り。転移学習では、学習済みモデルの重みはそのまま用いて追加した層の重みのみ学習を行う。
ゼロショット学習では、ラベル付きのデータは全く与えられず、対象物ではなく対象物に関する色や形状などの情報についての特徴ベクトルを学習して、未知のデータの分類を行う。似た手法にワンショット学習が挙げられるが、これは対象物以外の類似物などの情報を大量に与えた上でラベル付きの対象物のデータを一つ与えて学習を行い、その分類などを行うものである。例えばラベル付きの文字の画像を一つだけ与えて学習を行った後に書体の違うその文字の分類も出来るように、それ以外の様々な手書き文字のデータを与えて「手書き文字」についての学習を行う、という様なケースはワンショット学習と呼ばれる。
コンセプトドリフトとマルチタスク学習については選択肢の説明通りである。

量子化

【問題】
モデル圧縮手法の一つである量子化に関する説明として正しいものを選べ。

【選択肢】
(a)浮動小数点で表されるネットワークのパラメータを低bitで表現する手法であり、計算処理を高速にするために用いられている。
(b)学習時に、ある割合でいくつかのノードを無効にする手法のことである。
(c)寄与の小さいパラメータを0にすることで、精度を保ったままパラメータを減らすことが可能である。
(d)既に学習済みのモデルの入出力を用いてより軽量なモデルを学習する手法である。

【解答】
(a)浮動小数点で表されるネットワークのパラメータを低bitで表現する手法であり、計算処理を高速にするために用いられている。

【解説】
量子化は、浮動小数点で表されるネットワークのパラメータを低bitで表現すること手法。ほぼ精度を落とさずモデル圧縮が可能である。高速な計算処理を可能にするための手法の一つである。
「学習時に、ある割合でいくつかのノードを無効にする手法のことである。」
この説明はドロップアウトについてのものであるので誤り。
「寄与の小さいパラメータを0にすることで、精度を保ったままパラメータを減らすことが可能である。」
この説明はプルーニングについてのものであるので誤り。
「既に学習済みのモデルの入出力を用いてより軽量なモデルを学習する手法である。」
この説明は蒸留についてのものであるので誤り。

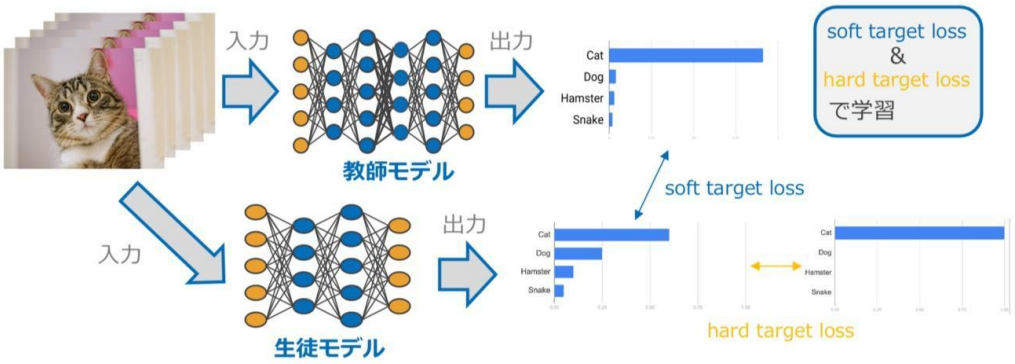
モデル圧縮

【問題】
以下はあるモデル圧縮の手法についての説明である。この手法の名称として正しいものを選べ。
「まず、教師モデルとして学習済みの大きいモデルやアンサンブルしたモデルなどを用意する。そのモデルにおける入出力を用いて生徒モデルと呼ばれる小さいモデルの学習を行うことで、ある程度の精度を持つ計算量の少ないモデルを構成する手法である。Noisy　Studentなどの近年の強力な手法にも用いられている。」

【選択肢】
(a)マルチタスク学習
(b)モデル並列
(c)蒸留
(d)ファインチューニング

【解答】
(c)蒸留

【解説】
蒸留は大きなモデルの入出力を用いて小さいモデルの学習を行うことで、高精度な軽量モデルを作成しようとする手法。通常の損失関数に加え、生徒モデルの出力分布が教師モデルの出力分布に近くなるようにする損失も考える。前者はhard target loss、後者はsoft target lossと呼ばれる。



選択肢の各手法について簡単に説明する。
マルチタスク学習：単一のモデルで複数のタスクを同時に学習することで、メインのタスクの汎化性能を向上させる手法。
モデル並列：モデルをいくつかに分割して、それぞれを複数の計算資源で処理することで学習を高速化させる手法。
ファインチューニング：あるタスクで既に学習済みのモデルのパラメータを初期値として、別のタスクでモデル全体の再学習を行う手法。

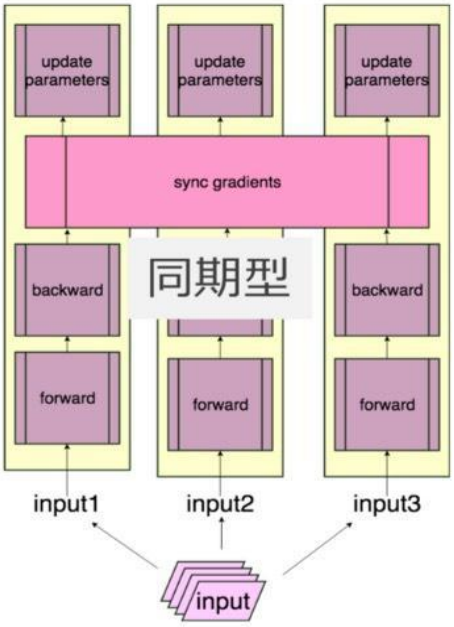
データ並列

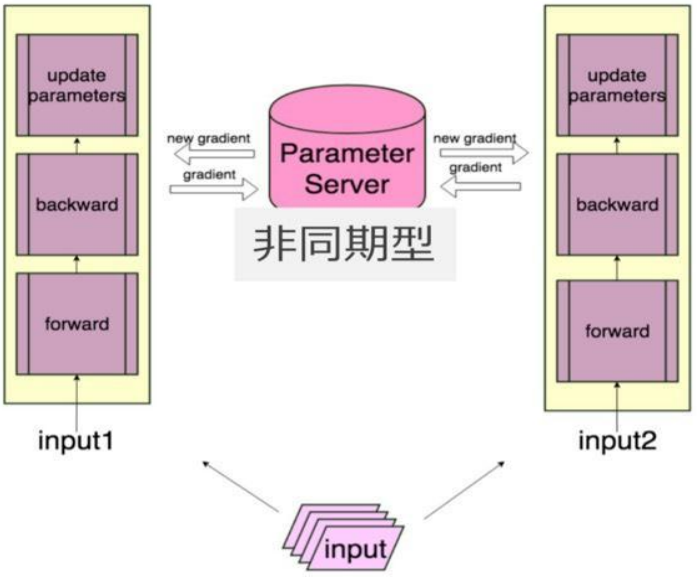
【問題】
学習を高速化させる分散処理におけるデータ並列についての説明として正しいものを選べ。

【選択肢】
(a)学習の際は、一つのモデルをいくつかの小さい部分モデルに分けて、それぞれの分割されたモデルを用いて同一のデータの処理を行う。
(b)非同期型においては、各レプリカが求めた勾配をパラメーターサーバに送信し、その勾配を用いてパラメータサーバが平均勾配を更新する。
(c)非同期型のデータ並列は、同期型のデータ並列よりもスループットは低いが、精度の面で優れている。
(d)同期型のデータ並列では、一つ以上の計算資源の勾配計算が完了したらその時点でモデルを更新する。

【解答】
(b)非同期型においては、各レプリカが求めた勾配をパラメーターサーバに送信し、その勾配を用いてパラメータサーバが平均勾配を更新する。

【解説】
データ並列は別々の計算資源上に異なるデータを割り当てて、その数分だけをモデルのレプリカを用意して同時に学習を行うことで学習を高速化させる手法である。
データ並列には同期型と非同期型の二種類が存在する。





同期型は全計算資源の勾配計算が完了してからモデルを更新するが、非同期型は一つ以上の勾配計算が完了したらその時点でモデルを更新する。そのため、非同期型ではパラメータサーバを用いてそれぞれの勾配の情報を保存しておく。
「学習の際は、一つのモデルをいくつかの小さい部分モデルに分けて、それぞれの分割されたモデルを用いて同一のデータの処理を行う。」

この説明はモデル並列の説明であるため、誤り。
「非同期型のデータ並列は、同期型のデータ並列よりもスループットは低いが、精度の面で優れている。」
非同期型のデータ並列は一つ以上の勾配計算が完了次第モデルの更新が行われるが、同期型では全ての勾配計算が終わるまでモデルの更新を待たなければならないので、非同期型の方がスループットは優れている。しかし、精度の面では同期型の方が優れている。よってこの選択肢の説明は誤りである。
「同期型のデータ並列では、一つ以上の計算資源の勾配計算が完了したらその時点でモデルを更新する。」
これは非同期型のデータ並列についての説明であるため、誤り。

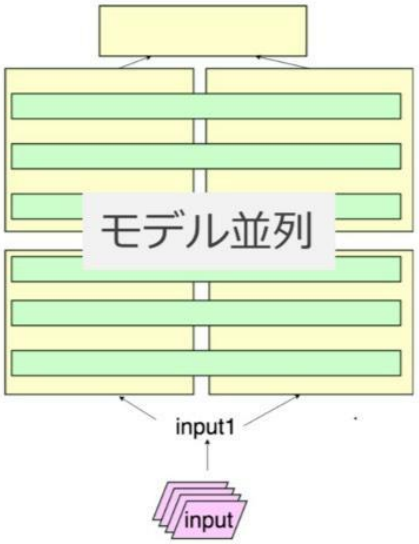
モデル並列

【問題】
学習を高速化させる分散処理におけるモデル並列についての説明として誤っているものを選べ。

【選択肢】
(a)モデルの分割の仕方は層毎やチャンネル毎など様々な方法が存在し、最適な分割法はモデルに依存する。
(b)データ並列と比較すると、順伝播と逆伝播の両方で通信が必要なこともあり、通信コストが高い。
(c)データ並列と同様にモデルのレプリカを複数用意する必要がある。
(d)モデルのサイズが大きすぎて、一つのGPUのメモリでは処理し切れない様なケースに有用な手法である。

【解答】
(c)データ並列と同様にモデルのレプリカを複数用意する必要がある。

【解説】
データ並列ではモデルのレプリカを複数用意してそれぞれで処理を行うが、モデル並列は一つのモデルを何かしらの方法で分割して複数のプロセスに配置して処理を行う手法であるので誤り。
モデル並列は、一つのモデルの何かしらの方法で分割し、それぞれの部分を別々の計算資源に割り当てて学習を行うことで学習を高速化させる手法。必要に応じて計算資源間で通信を行って処理を継続する。データの行き来が頻繁に行われてしまうことから、通信コストが増加してしまうことが問題とされている。



他の選択肢については正しい記述である。

【参考】
https://logmi.jp/tech/articles/285424

GPU

【問題】
GPUについての特徴として正しいものを選べ。

【選択肢】
(a)計算処理速度向上のために量子化を利用している。
(b)単精度浮動小数点演算と倍精度浮動小数点演算の両方が可能であるが、計算速度も考慮して倍精度を用いることが基本である。
(c)画像処理ではなく汎用演算を目的としてGPUの演算機能を用いる技術のことをGPGPUと呼ぶ。
(d)並列計算よりも直列計算の方が得意であるという特徴がある。

【解答】
(c)画像処理ではなく汎用演算を目的としてGPUの演算機能を用いる技術のことをGPGPUと呼ぶ。

【解説】
GPUは並列計算が高速であることが特徴であり、この特徴がディープラーニングにおける行列計算の高速化に非常に有用であり、現在は無くてはならない計算資源となっている。
元々画像処理を目的とした計算機であったGPUを、ディープラーニングのような画像や映像とは直接関係のない計算用途を目的として用いることをGPGPUと呼ぶ。
以下、各選択肢が誤っている理由について説明する。
「計算処理速度向上のために量子化を利用している。」
GPUでは量子化は用いられていないため、誤り。量子化とは浮動小数点で表されるパラメータを低いbitで近似することで精度を落とさずに計算速度を向上させる手法である。TPUにおいて計算高速化のために量子化が用いられている。
「単精度浮動小数点演算と倍精度浮動小数点演算の両方が可能であるが、計算速度を考慮して倍精度を用いることが基本である。」
前半の記述は正しいが、基本的には計算速度を考慮して単精度を用いるため、誤り。
「並列計算よりも直列計算の方が得意であるという特徴がある。」
GPUが得意であるのは並列計算であるため、誤り。

GPUとCPU

【問題】
GPUとCPUの違いに関する説明として誤っているものを選べ。

【選択肢】
(a)CPUのコア数は数個であるのに対して、GPUは数千個のコア数を持つ。
(b)GPUはCPUと異なり、SIMDと呼ばれる単一命令形で複数データの処理に適した設計になっている。
(c)一つあたりのコア性能で見るとCPUよりもGPUの方が高性能であるため、複雑な処理に向いているのはGPUである。
(d)計算の際に利用されるメモリに関して、CPUはメモリ容量が重要視され、GPUでは単位時間あたりのデータ読み出し量が重要視される傾向がある。

【解答】
(c)一つあたりのコア性能で見るとCPUよりもGPUの方が高性能であるため、複雑な処理に向いているのはGPUである。

【解説】
コア数で比較すると、GPUが数千個のコアを持つのに対してCPUは数個しか持たないが、コア性能で見るとCPUの方が高性能であるため複雑な処理に向いているため、誤り。
GPUは単純な処理を並列に、CPUは複雑な処理を直列に処理するのが得意であるというのがそれぞれの特徴である。
GPUはSIMDと呼ばれる単一命令形で複数データの処理に適した設計になっているもの特徴の一つである。一つの命令を複数のプロセッサで実行して並列処理する方式。なお、CPUはMIMDという並列命令形で複数データの処理に適した設計になっている。こちらは一つの命令に対して、一つのプロセッサで処理する方式である。
また、GPUにおいて単位時間あたりのデータ読み出し量は計算速度を握る鍵になっている。なお、画像を用いた大きなモデルの場合はGPUでもメモリ容量が重要となるので注意が必要。