１．　section1_強化学習

DQN は、深層学習を用いた強化学習モデルであり、深層強化学習の発展に大きく貢献したモデルの１つである。

DQN を用いて Atari2600 の中の７つのゲームを学習させたところ、ゲームごとのチューニングを行わずに、人間の専門家に匹敵する性能を達成したと述べられている。

DQN では、畳み込みニューラルネットワークを用いて行動価値関数を推定する。

（１）DQN での学習を安定化させるため工夫

①体験再生（experience replay）

タイムステップ t における状態を st、タイムステップ t でエージェントがとる行動を at、タイムステップ t での報酬を rt とし、各タイムステップ t におけるエージェントの経験 et=(st,at,rt,st+1)をデータ集合 D={e1…1n}に蓄積する。このデータ集合 D から取り出されたものを再生記憶（replay memory）という。学習時には、蓄積されたサンプルの中から、経験をランダムに取出し、損失の計算に用いる。

②目標 Q ネットワークの固定

③報酬のクリッピング

報酬の値を{-1,0,1}の３値に制限すること。これにより「報酬のスケールが与えられたタスクによって大きく異なる」という問題が解消されゲームごとに学習率を調整する必要がなくなった。

（２）体験学習には、通常のオンライン Q 学習に比べていくつかの利点がある。

①パラメータの更新時に、同じ経験を何回も使えるため、計算量の大きなエピソードの振興の回数を抑制できること。

②更新の分散を軽減できること。強い相関性を持つ入力系列に対して学習を行うと、直近の入力に引きずられてパラメータが修正されるため、過去の入力に対する推定が悪化し収束性が悪くなるが、体験再生では経験をランダムに取り出すため、系列方向の相間を断ち切ることができる。

③過去の様々な状態で行動分布が平均化されるため、直前に取得したデータが次の行動の決定に及ぼす影響が軽減できる。これにより、パラメータの振動や発散を避けることができる。

（３）「価値関数が小さく更新されただけても、選ばれる行動が大きく変わってしまう」という問題に対して DQN では、目標 Q ネットワークを固定するという工夫を行った。目標値の算出に用いる Q ネットワークのパラメータを固定し、一定周期でこれを更新することで学習を安定させる。

損失関数

$L(\theta)=Es,a,r,s'\sim D[(r+\gamma \max Q(s',a';\theta^-)-Q(s,a;\theta))^{**}2]$

２． section2_AlphaGo

囲碁のようなゲームに対して、盤面ｓに対して最適な状態価値関数ｖ*(ｓ)を評価するのは一般的に難しい。ｂをゲームの幅（状態毎の合法手の数)、ｄをゲームの深さ（ゲームの長さ）として、状態価値を探査木で計算するには、ｂｄ乗のオーダーの計算が必要だが、囲碁の場合ｂ～２５０、ｄ～１５０となるため、現実的な時間で計算できないからである。そのため深さと幅を減らすことが必要になる。

深さを減らすためには、盤面ｓの状態価値の評価を上手く近似しなければならない。幅を減らすためには、全行動からのサンプリングの仕方を工夫する必要がある。

AlphaGo では、畳み込みニューラルネットワークでモデル化した状態価値関数および方策関数を用いることで、探索の際の蓋さと幅を減らしている。

AlphaGo の学習過程は３つのステージで構成される。

（１）教師あり学習によって囲碁の熟練者の手をニューラルネットワークに学習させる。

このニューラルネットワークをSL方策ネットワークと呼ぶ。SLはSupervised Learning(教師あり学習)。

このネットワークの入力は盤面であり、出力はそれぞれの合法な手を選ぶべき確率ｐσ（ａ｜ｓ）とする。

ここでａは手の種類、σはネットワークのパラメータである。学習時には、教師データを｛ｓｋ、ａｋ｝k=1~m（ｍはミニバッチのサンプル数）で与え、ネットワークのパラメータの更新は

$$\Delta\sigma = \alpha / m \Sigma (\delta \log p_\sigma (ak|sk) / \delta \sigma)$$

をσに加えることによって行う。α：学習率、Tiはｉ番目の対戦の最終タイムステップである。

－＞対数尤度を最大化するため、勾配上昇法となる。

（２）（１）で最適化されたパラメータを初期値にして、新たなニューラルネットワークを学習する。

このネットワークをRL方策ネットワークと呼ぶ。RLはReinforcement Learning（強化学習)。

RLネットワークの構造は、SL方策ネットワークと同一であり、パラメータをρ、ネットワークの出力は方策

$p_\rho(a|s)$とする。パラメータρの初期値は、SL方策ネットワークのパラメータσとする。RL方策ネットワークの学習では、現在の方策とこれまでの方策からランダムに選択された方策同士を戦わせる。ランダムに選択する理由は、現在の方策に対して過剰適合しないようにするためである。パラメータの更新は

$$\Delta\rho = \alpha / n \Sigma\Sigma (\delta \log p_\rho (a|s) / \delta \rho)(z - v(s))$$

をρに加えることによって行う。この式はREINFORCEアルゴリズムに基づいて導出されている。

ｉ番目の対戦のタイムステップｔにおける報酬ｚはｚ＝ｒ（ｓTi）で与えられる。勝ちの場合１、負けの場合－1である。

つまりｚは、そのゲームに勝つ場合は常に１が与えられ、そのゲームに負ける場合は常に－１が与えられることになる。

ここでｖ（ｓ）は、勾配の分散を低減させるために導入されており、ベースラインと呼ばれる。

（３）盤面ｓに関する価値関数ｖｐ（ｓ）を評価する。

価値関数ｖｐ（ｓ）は、盤面ｓから始めて、両方のプレイヤーが方策ｐを用いて指し手を選んだ場合のゲームの結果

$V_p(s) = E[zt|st=s, at....aT\sim P]$である。

最適方策はわからないため、RL 方策ネットワーク p ρに関する価値関数 v pρを評価する。また、価値関数をパラメータθの価値ネットワーク v θ(s)でモデル化する。つまり、最適な価値関数 v*(s)を v θ(s)~v Pρ(s)~v*(s)となるよう学習を行う。価値ネットワークの構造は、方策ネットワークの構造と類似のものであるが、出力は v*(s)の値である。価値ネットワークのパラメータは、盤面と目標出力のペア(s、z)へ回帰させることで更新する。更新は

$$\Delta\theta = \alpha / m \Sigma (zk - v\theta(sk)) \cdot \delta v\theta(sk) / \delta\theta$$

をθに加えることによって行う。

目標出力 z を得るために行う対戦は“t＝1 .....U-1 の時間ステップは SL 方策ネットワーク pσから、t＝U の時間ステップはあらゆる手からランダムに(手が合法手になるまでサンプリング)指し手を選び、t＝U+1...T の時間ステップは RL 方策ネットワーク pρから指し手を選ぶ”という処理を行う

AlphaGo で用いられているモンテカルロ木探索（Monte Carlo Tree Search,MCTS）のアルゴリズムである APV-MCTS の概略

・選択（selection）
・展開（expension）
・評価（evaluation）
・記録（backup）

３． section3_軽量化_高速化技術

（１）蒸留（distillation）

蒸留とは大きなモデルが獲得した知識を小さなモデルに転移する方法でり、一般に軽量化の手法に位置づけられる。

分類問題の学習の目的は通常、正解クラスに対する確率の対数の平均を最大にすることである。ニューラルネットワークなどで構成されたモデルは、正解クラスに対する確率だけでなく、不正解クラスに対する確率も出力する。

蒸留では、この不正解クラスの確率を利用する。

（２）軽量化（量子化）（quantization）

①パラメータと活性化の両方を２値化すると、メモリ消費量、速度ともにメリットがある。

②３値化は２値化と比べると情報量が多くなり、精度が上がる。

③パラメータ、活性化、勾配は量子化の対象になる。

④活性化を二値化することは、活性化にステップ関数に似た形状の関数（ｘ＜０でｆ（ｘ）＝－１、ｘ≧０でｆ（ｘ）＝１となる関数）を適用することと同じであり、ほとんどの勾配が０となる問題が生じる。これを回避するため、逆伝播時にストレート・スルー・エスティメータ（Straight-Through Estimator STE）を適用すると、勾配０がなくなり学習が進む。STE は逆伝播時だけ、対象となる関数を恒等写像関数として扱うという方法である。

（３）剪定（pruning）

①He et al., 2017 で提案されたチャネル剪定（Channel pruning）ではラッソに基づいて代表的なチャネルを選択し、冗長なチャネルを刈り取った上で、最小二乗法で残ったチャネルを再構成する。

②宝くじ仮説

ニューラルネットワークには「部分ネットワーク構造」と「初期値」の組み合わせに当たりが存在し、それを引き当てると効率的に学習が可能という説。

「剪定によって得られる疎なネットワークは、なぜ剪定を使わずに最初から学習できないのか？」

③剪定は学習と同時又は学習後に行われる。

④剪定とは、精度に大きく影響を与えないノードやエッジなどを刈ることによって、データ量と計三回数を削減する方法。

（４）データ並列（data parallelism）・モデル並列（model parallelism）

分散深層学習のアプローチの手法。

①データ並列

親モデルをコピーしたｎ個のレプリカ（子モデル）を各 GPU ワーカに配置し、異なるバッチを各 GPU ワーカに供給する。

②データ並列

同期型と非同期型に分けられる。

４． section4_応用モデル

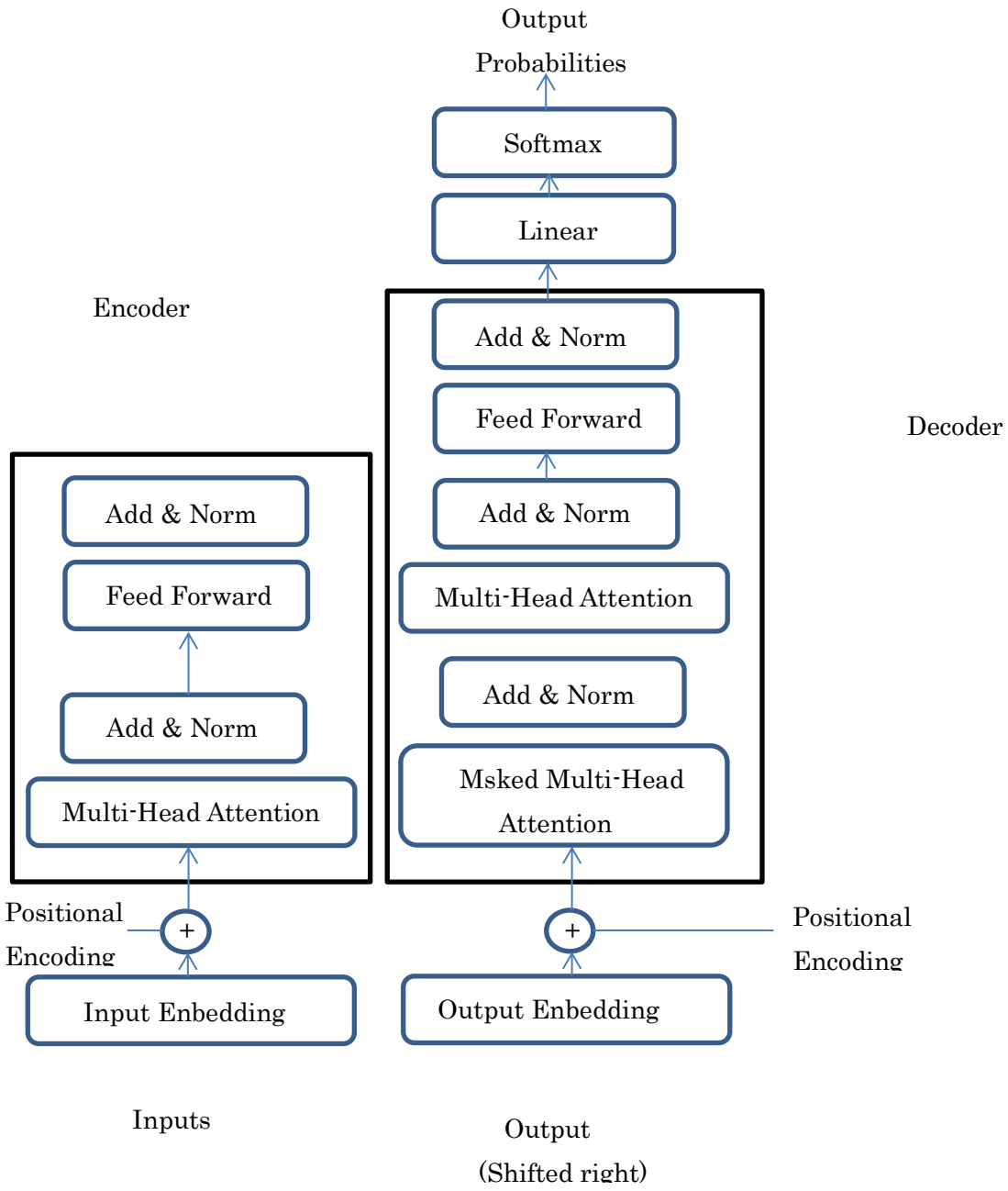
画像分類における高性能な CNN 構造とその特徴

名称	
VGG	Simonyan & Zisserman,2014 は２０１４年に提案されたモデルで、フィルタサイズ３×３の畳み込み層を複数積み重ねることで、少ないパラメータで大きなフィルタを使用した場合と同じ範囲を畳み込むことを提案したモデルである。構造としてはシンプルだが、実務でよく用いられるモデル。
GoogLeNet	Szegedy et.,2014 は２０１４年に提案されたモデルで、フィルタサイズの異なる畳み込み層を並列に配置するインセプション（inception）モジュール、ネットワークの途中で分岐させたサブネットワークで予測を行う補助的分類器（auxiliary classifier）といた工夫が導入されている。 サブネットワークでも予測を行い、そこからも逆伝播することで勾配消失を回避している。
ResNet	He et al.,2015 は２０１５年に提案された恒等写像を介して、特徴マプ同士を足し合わせるショートカット結合が特徴なモデルである。このショートカット結合によって、勾配消失を引き起こすことなく、より深く層を重ねることができるようになった。（１００層を超える）
DenseNet	Huang et al.,2017 はショートカットを持っている点では ResNet と同じだが、ResNet のようにショートカットされた特徴マップを足し合わせるのではなく、特徴マップをチャネル方向に結合する。さらにそのショートカット結合は、あるブロック内ですでに作られた特徴マップ全てを対象にする。

５．　section5_Transformer

２０１７年、RNN や CNN を使わずに系列データを扱う Transformer というモデルが提案された。
Transformer は Seq2Seq と同じようにエンコーダとデコーダからなる。エンコードとデコードはそれぞれ N 回繰り返される。

●Transformer のネットワーク構成



６． section6_物体検知_セグメンテーション

セマンティックセグメンテーションとは、ピクセルごとにクラス分類問題を解くことで、１枚の画像中に含まれる複数の物体に対し、それぞれが存在する領域に対応するクラスを出力するタスクである。セマンティックセグメンテーションを CNN を用いて実現する場合、ピクセルごとに分類問題を解くため、いくつかのピクセルが孤立して異なるクラスに割り当てられてしまう問題がある。この場合、条件付き確率場（Conditional Random Field,CRF）による後処理を施すことで精度向上が期待できる。

セマンティックセグメンテーションにおいて、モデルの性能は、予測された領域とラベルとして与えられた領域の重複度によって評価されることが多い。条幅度を測る指標の一つとして、IoU が与えられるが、IoU は２つの領域の面積の差が大きいほど値が小さく評価されてしまう欠点がある。そこで IoU の分母を２つの領域の平均値に置き換えた Dice 係数もよく用いられる。

Dice 係数は、IoU の欠点を緩和し共通部分の面積より重要視した指標である。

ラベルとして与えられた領域を Strue、予測された領域を Spred としたとき、Dice 係数は

$$Dice(Strue,Spred)=(2|Strue \cap Spred|)/(|Strue|+|Spred|)$$

で定義される。ここで|S|は、領域 S の面積を求める演算、Strue∩Spred は Strue と Spred の共通領域を指す。

一般に IoU と Dice 係数の関係は IoU≦Dice 係数となる。