

2021 年度 修士論文

深層学習を用いた ILC 崩壊点検出アルゴリズムの開発

九州大学大学院 理学府 物理学専攻
粒子物理学分野 素粒子実験研究室

後藤 輝一

指導教員 川越 清以

2020 年 12 月 16 日



九州大学
KYUSHU UNIVERSITY

概要

素粒子はそれが従う統計によって二種類に分類され、フェルミ統計に従う粒子をフェルミ粒子、ボース統計に従う粒子をボース粒子と呼ぶ。現時点では存在が知られているフェルミ粒子はクォークとレプトンとに分類される。一方、現時点では知られているボース素粒子には、素粒子間の相互作用を伝達するゲージ粒子と、素粒子に質量を与えるヒッグス機構に関連して現れるヒッグス粒子がある。ゲージ粒子のうち、重力を媒介するとされる重力子は未発見である。素粒子の大きさは分かつておらず、大きさが無い（点粒子）とする理論と、非常に小さいがある大きさを持つとする理論がある。標準模型（標準理論）では素粒子には大きさが無い（点粒子）ものとして扱っており、現時点では実験結果と矛盾が生じていない。ただし、点粒子は空間が最小単位の存在しない無限に分割可能な連続体であることを前提としているが、標準模型で扱うスケールより 15 桁以上小さいスケール（プランク長スケール）においては、空間が連続的であるか離散的であるかは判明していない。離散的である場合には点粒子として扱えない。超弦理論においては全ての素粒子は有限の大きさを持つひもの振動状態であるとされる。我々が普段目にする物質は（微小な、あるいは大きさが無い）素粒子からできているにも関わらず、有限の大きさを持っている。それは、複数の素粒子が運動する有限の領域が、ハドロンや原子などの大きさを持つ粒子を構成することによる。素粒子のうちほとんどのものは、自然界に単独で安定的に存在しているわけではないので、宇宙線の観測や加速器による生成反応により発見・研究された。素粒子の様々な性質を実験で調べ、それを理論的に体系化していくこと、及び理論的に予言される素粒子を実験で探索していくことが、素粒子物理学の研究目的である。

目次

第 1 章	序論	10
1.1	標準模型	10
1.2	ILC 計画	11
1.3	ILC の物理	13
1.4	ILC の検出器 -International Large Detector (ILD)-	14
1.5	ILC のソフトウェアと事象再構成	14
1.5.1	ソフトウェア	16
1.5.2	飛跡の再構成	16
1.5.3	ジェットの再構成	17
1.6	本研究の目的	18
1.7	本論文の流れ	19
第 2 章	深層学習	20
2.1	機械学習と深層学習	20
2.2	パーセプトロン	22
2.2.1	単純パーセプトロン	22
2.2.2	多層パーセプトロン	23
2.3	ニューラルネットワーク	24
2.3.1	ニューラルネットワークの構造	24
2.3.2	ニューラルネットワークの学習	27
2.3.3	ディープニューラルネットワーク	30
2.4	リカレントニューラルネットワーク	30
2.4.1	リカレントニューラルネットワークの構造と学習	30
2.4.2	リカレントニューラルネットワークの問題点	33
2.4.3	長・短期記憶 (Long Short-Term Memory, LSTM)	33
2.5	Attention	38
2.5.1	エンコーダー・デコーダーモデル	38
2.5.2	Attention	38

2.6	ハイパーパラメータ	41
第3章	崩壊点検出の為のネットワーク	43
3.1	データ	43
3.1.1	データ全体の性質	43
3.1.2	飛跡の情報と前処理	47
3.2	深層学習を用いた崩壊点検出の実現	50
3.3	飛跡対についてのネットワーク	51
3.3.1	ネットワークの構造	52
3.3.2	ネットワークの学習と戦略	52
3.3.3	ネットワークの評価	55
3.4	任意の数の飛跡についてのネットワーク	62
3.4.1	ネットワークの構造	62
3.4.2	ネットワークの学習と戦略	67
3.4.3	ネットワークの評価	68
第4章	深層学習を用いた崩壊点検出	75
4.1	崩壊点検出アルゴリズム	75
4.2	崩壊点検出の最適化と評価	75
4.3	崩壊点検出の性能	75
第5章	現行の手法との比較	76
5.1	崩壊点検出単体での比較	76
5.2	フレーバータギングを含めた比較	76
第6章	まとめと今後の展望	77
付録A	ソースコード	79
A.1	飛跡対についてのネットワーク	79
A.2	任意の数の飛跡についてのネットワーク	79
A.3	崩壊点の再構成	79
参考文献		80

図目次

1.1	標準模型の素粒子	11
1.2	ILC の全体像	12
1.3	ILC 計画の今後	12
1.4	重心系エネルギーとヒッグス事象生成断面積の関係	13
1.5	International Large Detector (ILD)	15
1.6	Primary Vertex と Secondary Vertex の図示	17
1.7	深層学習によるジェットの再構成	19
2.1	機械学習の中の深層学習の位置付け	21
2.2	単純パーセプトロン	22
2.3	ヘヴィサイドの階段関数	23
2.4	多層パーセプトロン	23
2.5	活性化関数	25
2.6	ニューラルネットワーク	25
2.7	リカレントニューラルネットワーク	31
2.8	リカレントニューラルネットワークの重みの明示的な表現	32
2.9	リカレントニューラルネットワークの出力方法	32
2.10	LSTM の流れ	34
2.11	単体の LSTM	34
2.12	LSTM の各ゲートについての図解	36
2.13	Stacked LSTM	37
2.14	双方向 LSTM	37
2.15	LSTM によるエンコーダー・デコーダーモデル	39
2.16	Attention と LSTM によるエンコーダー・デコーダーモデル	39
2.17	Additive Attention と Dot-Product Attention	41
3.1	終状態 $b\bar{b}$ での崩壊点の例	46
3.2	事象に含まれる飛跡の本数と崩壊点の個数	46
3.3	トラック・パラメータの定義 [1]	47

3.4	変数の分布の例	48
3.5	LCFIPlus によって予想される崩壊点の位置の分布	49
3.6	LCFIPlus によって予想される崩壊点の位置と χ^2 値の相関	49
3.7	終状態 $b\bar{b}$ での崩壊点	52
3.8	飛跡対についてのネットワークの概略図	53
3.9	各終状態での分類クラスのデータ数の比	53
3.10	訓練データでの分類クラスのデータ数の比	54
3.11	評価のための飛跡対についてのネットワーク	57
3.12	ネットワークのスコアとクラス分類の効率の関係	58
3.13	各モデルの ROC 曲線	59
3.14	各モデルの混合行列と各モデルの相対値	61
3.15	t-SNE による次元削減の比較	62
3.16	リカレントニューラルネットワークを用いた崩壊点生成	63
3.17	崩壊点生成のためのリカレントニューラルネットワーク構造	63
3.18	系列 1 ステップについての独自リカレントニューラルネットワーク構造	64
3.19	独自リカレントニューラルネットワーク構造の解釈	65
3.20	Attention を組み込んだエンコーダー・デコーダーモデルへの拡張	66
3.21	独自リカレントニューラルネットワークの Attention への拡張	67
3.22	飛跡順のシャッフル	69
3.23	標準的な LSTM と独自のネットワークの比較	71
3.24	各データ属性の効率とスコアの関係	72
3.25	各データ属性の ROC 曲線	72
3.26	Attention Weight	74

表目次

1.1	ILD サブディテクターの詳細なパラメータ (バレル)	15
1.2	ILD サブディテクターの詳細なパラメータ (エンドキャップ)	15
3.1	MC シミュレーションデータの性質	44
3.2	データサンプルの事象数と用途	45
3.3	ソフトウェア・ハードウェアの環境	51
3.4	飛跡対についてのネットワークにおける訓練可能なパラメータ	56
3.5	評価のための飛跡対についてのモデル	56
3.6	任意の数の飛跡についてのネットワークの入力変数の大きさ	68
3.7	任意の数の飛跡についてのネットワークにおける訓練可能なパラメータ	70

第 1 章

序論

本章では、まず 1.1 節で素粒子の振る舞いを記述する理論である、標準模型 (Standard Model, SM) について紹介する。

次に 1.2 節にて、この標準模型や標準模型を超える物理 (Physics beyond the Standard Model, BSM) を探索するための国際線形加速器 (International Linear Collider, ILC) 計画についての説明を行う。そのような ILC が目指す物理について 1.3 節で述べる。また、ILC で使用される予定の検出器やソフトウェアに関する説明を 1.4 節と 1.5 節で行う。

更に本研究の目的について 1.6 節で、本論文の流れについて 1.7 節でそれぞれ述べ本論文の序論とする。

本章の作成にあたり、参考文献 [2, 3] を使用した。

1.1 標準模型

宇宙の誕生や、生物の発生と同様に、物質の起源は人類の根元的な問いの一つである。そのような物質の素となる粒子のことを素粒子といい、その素粒子の振る舞いを記述する理論を標準模型という。標準模型によると、素粒子はスピン半整数のフェルミ粒子とスピン整数のボース粒子に分類される。

フェルミ粒子は全てスピン $1/2$ の粒子で構成され、陽子や中性子などのハドロンを構成するクォークと電子やニュートリノなどのレプトンに分けられる。更に、クォークは電荷が $+2/3$ のアップクォーク系列と $-1/3$ のダウントクォーク系列に、レプトンは電荷が -1 の荷電レプトンと中性電荷の中性レプトン (ニュートリノ) に分けられる。また、それぞれ世代と呼ばれるものを構成し現在は 3 つの世代が確認されている。クォークの場合はアップクォーク u 、チャームクォーク c 、トップクォーク t 、ダウントクォーク d 、ストレンジクォーク s 、ボトムクォーク b が存在する。レプトンの場合は荷電レプトンとして、電子 e^- 、ミュー粒子 μ^- 、タウ粒子 τ^- 、中性レプトンとして、電子ニュートリノ ν_e 、ミューニュートリノ ν_μ 、タウニュートリノ ν_τ が存在している。これらの系列や世代間の違いをフレーバーと呼んでいる。

ボース粒子は基本的な四つの相互作用である、強い相互作用、弱い相互作用、電磁相互作用、

重力相互作用の内、重力相互作用を除いた三つの相互作用をそれぞれ媒介するスピン 1 のゲージ粒子と、対称性を破り素粒子に質量を与えるスピン 0、中性電荷のヒッグス粒子 H で構成される。電磁相互作用を媒介する粒子として、中性電荷の光子 γ 、強い相互作用を媒介する粒子として、中性電荷のグルーオン g、弱い相互作用を媒介する粒子として、電荷 ± 1 の W ボソン W^\pm 、中性電荷の Z ボソン Z が存在する。

更に素粒子には、質量やスピンが等しく電荷の正負のみが反転した反粒子が存在する。これら反粒子と通常の粒子が衝突すると対消滅が起こり、その質量が全てエネルギーへと変換される。一方、ある粒子の質量の二倍以上のエネルギーを生じさせた場合は対生成が起こり、その粒子と反粒子の対が生成されることがある。

以上の素粒子を標準模型の素粒子といい、一般に図 1.1 のように纏められている。

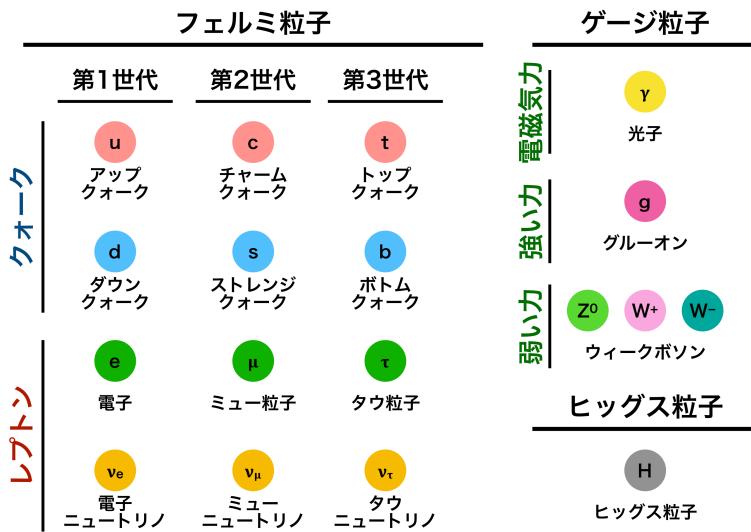


図 1.1: 標準模型の素粒子

標準模型は様々な実験で非常によく確かめられているが、ダークマターをはじめとする幾つかの物理現象を説明できていない。これらの標準模型で説明できない物理現象を BSM といい、現在は様々な実験施設で BSM の探索が行われている。次節の ILC 計画はそのような試みの一つである。

1.2 ILC 計画

ILC 計画とは、日本の東北にある北上山地に図 1.2 のような全長 20.5 km の国際線形加速器 (ILC) を建設する計画である。ILC 計画は国際共同研究であり、2013 年に出版された The Technical Design Report (TDR[4]) には 2400 人の研究者、48ヶ国、392 の研究機関と大学のグループが著名している。また、技術開発はリニアコライダーコラボレーション (The Linear Collider Collaboration, LCC) によって推進され、LCC の活動は国際将来加速器委員

会 (The International Committee for Future Accelerator, ICFA) の下、リニアコライダー国際推進委員会 (Linear Collider Board, LCB) によって監督されている。2021年現在、ILC 計画は準備段階へ向けて計画が進められており、ILC 準備研究所 (ILC Pre-Lab) の為の準備として ICFA は ILC の国際推進チーム (International Development Team, IDT) の設立を承認した。今後は LCC や LCB に代わり、この ILC 国際推進チームが ILC 計画の推進を行なっていく予定となっている。ILC 計画の今後の大まかな流れを図 1.3 に示す。

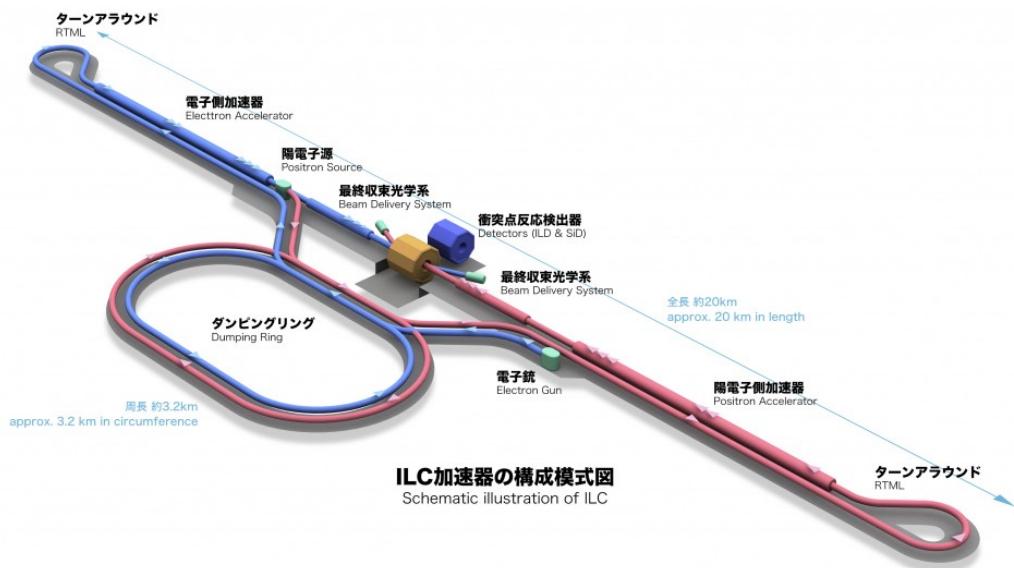


図 1.2: ILC の全体像 [5]

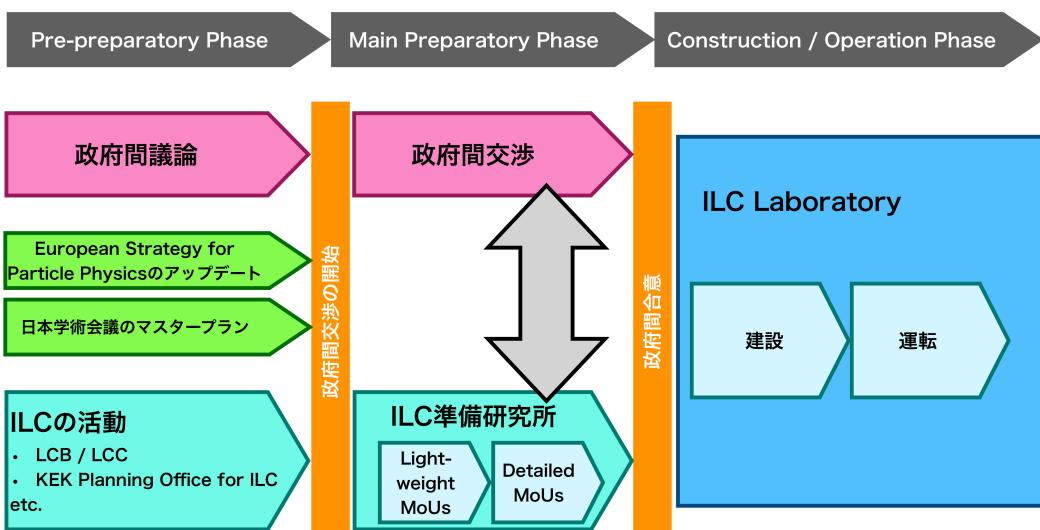


図 1.3: ILC 計画の今後 [6]

1.3 ILC の物理

ヒッグス粒子が 2012 年に欧州原子核研究機構 (CERN) の大型ハドロン衝突型加速器 (Large Hadron Collider, LHC) で発見されて以降、ヒッグス粒子の性質について、より詳細な調査が行われている。ヒッグス粒子は標準模型の中で電弱相互作用の対称性を破り、素粒子に質量を与える役割を担っており、更に質量に結合するという特徴を持っている。このような性質からヒッグス粒子の振る舞いは標準模型によって詳細に決定される。したがって BSM によって標準模型の素粒子の振る舞いに変化が生じた場合、ヒッグス粒子はその影響を受けると予想されている。特にヒッグス粒子の結合定数は、そのような BSM の仮定によって異なった変化をする事が示唆されている。

ILC は、このヒッグス粒子の性質を詳細に調べる為のヒッグスマクトリーとしての役割を期待されている。LHC が陽子-陽子を衝突させる加速器であるのに対し、ILC は電子-陽電子を衝突させる加速器である。これは粒子と反粒子の関係となっており、ILC は LHC と比較して目的とする事象に対しエネルギーをより効率的に使うことができる。更に電子-陽電子衝突の場合は相互作用の関係から、背景事象が少ないという特徴を持っている。

ILC は $e^+e^- \rightarrow Zh$ 事象の反応断面積が最大となる重心系エネルギー $\sqrt{s} = 250$ GeV での運転開始 (ILC250) を予定している (図 1.4)。また、ILC には様々な物理目標を達成する為に多数のアップグレードオプションが存在し、重心系エネルギーについてはメインリニアックを延長することで 1 TeV までの拡張が可能である。

Figure 2.7
Production cross
section for the
 $e^+e^- \rightarrow Zh$ process
as a function of the
center of mass energy
for $m_h = 125$ GeV,
plotted together with
those for the WW and
 ZZ fusion processes:
 $e^+e^- \rightarrow \nu\bar{\nu}H$ and
 $e^+e^- \rightarrow e^+e^-H$.

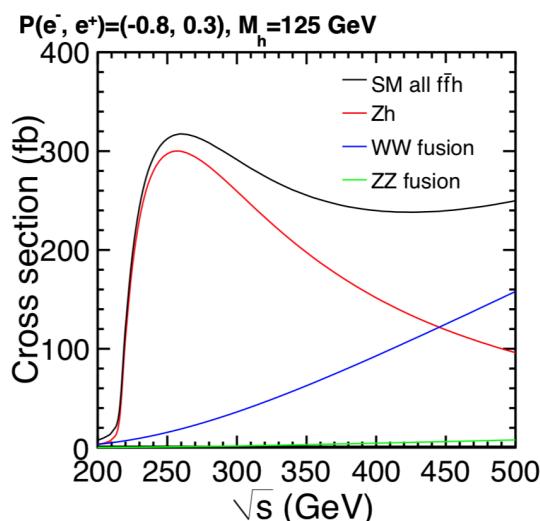


図 1.4: 重心系エネルギーとヒッグス事象生成断面積の関係 [7]。ヒッグス粒子の質量を 125 GeV とした時の $e^+e^- \rightarrow Zh$ 事象の生成断面積を赤線でプロットしている。また、WW fusion $e^+e^- \rightarrow \nu\bar{\nu}h$ 事象、ZZ fusion $e^+e^- \rightarrow e^+e^-h$ 事象をそれぞれ青線、緑線で表現している。

$e^+e^- \rightarrow Zh$ 事象はヒッグス粒子の反跳粒子である Z 粒子の識別をすることによって、ヒッグス粒子の崩壊モードに寄らず事象を選別できるという点で非常に重要である。また、背景事

象である $e^+e^- \rightarrow Z\gamma$ や $e^+e^- \rightarrow ZZ$ に関してもよく理解されており、電弱相互作用の計算によって 0.1 % 程度に抑えることができる [2]。この反跳粒子を使用した解析では $e^+e^- \rightarrow Zh$ 事象の全断面積を得ることができ、絶対正規化されたヒッグス粒子の結合定数やヒッグス粒子のエキゾチック崩壊についての測定が可能となる。

$e^+e^- \rightarrow Zh$ 事象での終状態の Z 粒子は、レプトン対 ($Z \rightarrow l^+l^-$) またはクォーク対 ($Z \rightarrow q\bar{q}$) に崩壊する。レプトン対にはおよそ 30% 程度の割合で崩壊し、クォーク対には残りの 70% 程度の割合で崩壊する為、統計量を大きくするという点でクォーク対をより精度よく識別することは非常に重要である。これらのクォーク対はエネルギー効率のために、それぞれ真空中でクォークの粒子反粒子対を生成・結合しハドロンとなる。この過程で生成されたクォークも同様にハドロンを形成するため、初めのクォーク対のそれぞれの進行方向には多数のハドロン粒子が生成されることとなる。これをジェットといい、Z 粒子を始めとする様々な粒子のクォークへの崩壊は、このジェットを用いて識別される。このジェットの再構成については、1.5.3 項にて説明する。

1.4 ILC の検出器 -International Large Detector (ILD)-

ILC では Silicon Detector (SiD) と International Large Detector (ILD) 二つの検出器コンセプトが検討されている。本研究は ILD のデータを用いている為、本節では ILD について簡単な解説を行う。ただし、本研究の基本的な発想はそのような検出器の違いに寄らず使用することができる。

ILD はヒッグス粒子や電弱相互作用の物理からの要求値を満たすように設計され、また後述する Particle Flow (1.5.2 項) によって最適化されている。また ILD は様々なサブディテクターによって構成され、ビームの衝突点 (図??の右下) を包む様に内側から順に、Vertex Detector (VTX)、Silicon Internal Tracker (SIT)、Time Projection Chamber (TPC)、Electromagnetic Calorimeter (ECAL)、Hadron Calorimeter (HCAL)、Iron Yoke (Muon) が並んでいる。また、HCAL と Iron Yoke の間には Solenoid Coil があり、3.5T の磁場をかけている。ILD では、VTX や SIT、TPC を用いて荷電粒子の飛跡を測定し、ECAL によって電子や光子、HCAL によってハドロン粒子のエネルギーを測定する。

衝突点の前方方向には、Forward Tracking Detector (FTD)、Luminosity Calorimeter (LumiCAL)、LHCAL、Beam Calorimeter (BeamCAL) が並んでいる。それぞれのサブディテクターの詳細については表 1.1 と表 1.2 にまとめれる。

1.5 ILC のソフトウェアと事象再構成

ここでは ILC で使用されるソフトウェアと事象再構成について述べる。事象再構成とは、加速器実験によって得られるデータから飛跡やジェットなどの物理情報を再構成するアルゴリズムである。そのような再構成は電子-陽電子の衝突毎に行われ、この衝突一回分の事を事象

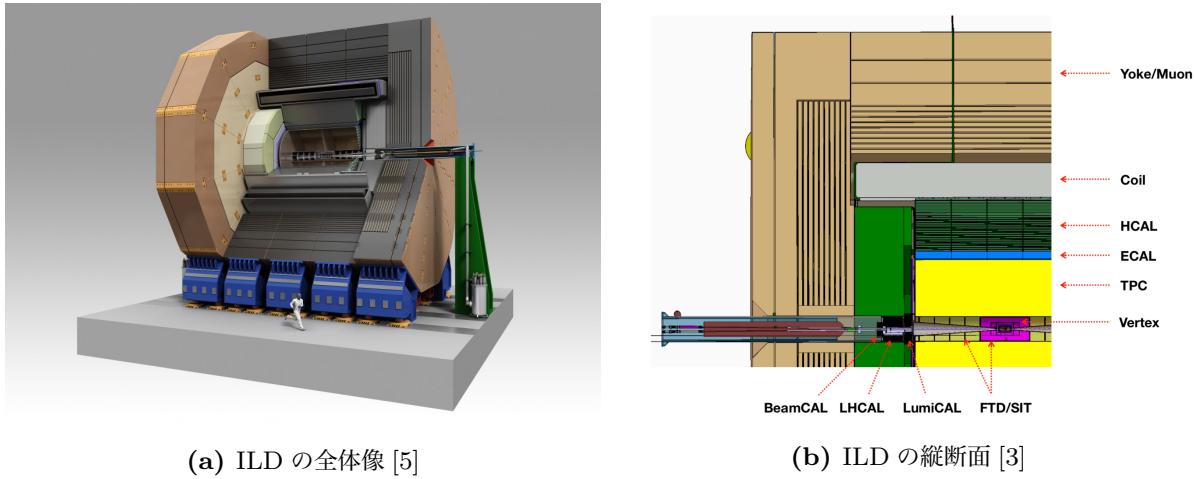


図 1.5: International Large Detector (ILD)

	r_{in} [mm]	r_{out} [mm]	z_{max} [mm]	要素技術
VTX	16	60	125	シリコンピクセルセンサー
SIT	153	303	644	シリコンピクセルセンサー
TPC	329	1770	2350	マイクロパターンガス検出器
SET	1773	1776	2300	シリコンストリップセンサー
ECAL	1805	2028	2350	吸収層：タングステン センサー：シリコン/シンチレーター
HCAL	2058	3345	2350	吸収層：スチール センサー：シンチレーター/RPC ガス
Coil	3425	4175	3872	
Muon	4450	7755	4047	センサー：シンチレーター

表 1.1: ILD サブディテクターの詳細なパラメータ (バレル) [3]

	z_{min} [mm]	z_{max} [mm]	r_{in} [mm]	r_{out} [mm]	要素技術
FTD	220	371		153	シリコンピクセルセンサー
	645	2212		300	シリコンストリップセンサー
ECAL	2411	2635	250	2096	吸収層：タングステン センサー：シリコン/シンチレーター
HCAL	2650	3937	350	3226	吸収層：スチール センサー：シンチレーター/RPC ガス
Muon	4072	6712	350	7716	センサー：シンチレーター
BeamCAL	3115	3315	18	140	吸収層：タングステン GaAs 読み出し
LumiCAL	2412	2541	84	194	吸収層：タングステン センサー：シリコン
LHCAL	2680	3160	130	315	吸収層：タングステン

表 1.2: ILD サブディテクターの詳細なパラメータ (エンドキャップ) [3]

という。

ILC のソフトウェアは iLCSoft[8] と呼ばれるソフトウェアエコシステムにまとめられている。

ILC における事象再構成は、トラッキングや Particle Flow といった 1.5.2. 粒子の再構成と、更にそれらによって再構成された粒子を使った 1.5.3. ジェットの再構成に区分できる。ILC ではジェットの再構成は崩壊点検出、ジェットクラスタリング、フレーバータギングという行程で行われる。これらジェットの再構成は iLCSoft 内の LCFIPlus[9] が使用されている。

1.5.1 ソフトウェア

ILC は実際の実験データを取得できないため、本研究で使用するデータは全てシミュレーションデータである。シミュレーションデータは標準模型と BSM を用いて、モンテカルロ (Monte Carlo, MC) 法によって生成されており、シミュレーションにおける検出器の情報は DD4hep[10, 11] という検出器記述の為のツールキットによって提供されている。それらのシミュレーションデータは LCIO と呼ばれる階層型の Event Data Model (EDM) によって管理されている。LCIO では、MC 情報から事象の生データ、デジタル化、解析や後述する再構成に至るまでが紐づけられており、階層的に取り扱うことができる。

ILC のソフトウェアは二つの検出器コンセプト (ILD, SiD) で共通しており、前述したように現在は iLCSoft というソフトウェアエコシステムによって統括されている。提供されている API の言語は C++・java・Fortran である。

それらソフトウェアモジュールは Marlin[12] という C++ アプリケーションフレームワークによって運用されており、プロセッサーと呼ばれるモジュールを作成・組み込むことにより、様々な再構成・解析アルゴリズムを簡単に別のモジュールへ置き換えることができる。また、データの入出力は LCIO、ROOT フォーマットによって行われる。

1.5.2 飛跡の再構成

飛跡の再構成は、シミュレーションされた検出器のデータから、粒子 (飛跡) を再構成するトラッキングと、そのようにして得られた個々のトラッキング検出器 (VTX, SIT, TPC など) の飛跡や粒子についての情報を繋ぎ合わせ、より高精度の粒子情報を提供する Particle Flow という手順によって行われる。

トラッキングでは Kalman-Filter が使用され、まず荷電粒子の軌跡をパターン認識を用いて再構成し、次にそれらの軌跡について運動学的な物理量をフィッティングによって抽出している。ILD において異なるサブディテクターのトラッキングは異なるアルゴリズムが使用されている。

トラッキングによって運動学的な物理量を得られるのは荷電粒子のみである。中性粒子は VXD や TPC に飛跡を残さない為、他のサブディテクターを用いた再構成が必要である。このように粒子の性質によって、再構成を行うべき最適なサブディテクターは異なる。ILC に

おける粒子種毎の最適な測定手法や再構成手法の選択は、Particle Flow によって提供されている。Particle Flow では荷電粒子をトラッキング検出器によって測定し、光子や中性ハドロンの再構成をそれぞれ ECAL や HCAL で行う。これらの再構成について、iLCSoft では PandoraPFA と呼ばれるアルゴリズムが使用されている。このアルゴリズムでは、まずカロリメータのヒットをクラスター化し、それらクラスターとトラッキング情報を関連づけ粒子識別を行っている。

以上が飛跡の再構成である。飛跡の再構成では、検出器で得られた情報から粒子を再構成するまでを行なっている。実際には、注目すべき物理事象は前述したジェットのような特徴的なシグネチャを残す為、次項のジェットの再構成による更なる事象再構成が必要である。

1.5.3 ジェットの再構成

事象中に生じたクォークは 1.3 節で述べたようにジェットを形成する。ジェットには多数の粒子（飛跡）が含まれ、それら多数の粒子の親となる粒子が崩壊した点を崩壊点（Vertex）という。特に、ハドロンのような準安定な親粒子の崩壊点の事を Secondary Vertex といい、電子-陽電子の衝突点を Primary Vertex という（図 1.6）。ジェットの再構成では、まずこの崩壊点を崩壊点検出（Vertex Finder）を用いて探索し、そのようにして得られた崩壊点を用いてジェット中の粒子を分離するジェットクラスタリング（Jet Clustering）が行われ、そのジェット中の飛跡から崩壊した準安定な親粒子のクォーク・フレーバーを識別するフレーバータギング（Flavor Tagging）が行われる。

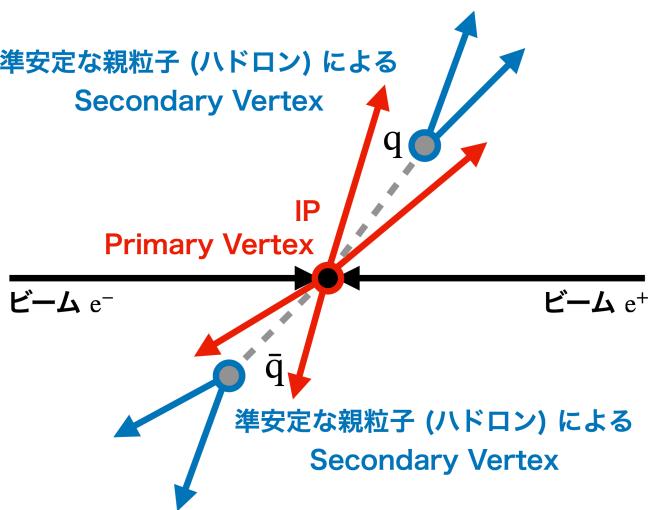


図 1.6: Primary Vertex と Secondary Vertex の図示。左右から電子・陽電子ビームが入射され、図中央で衝突したと仮定している。灰色の破線は準安定なハドロンを表現しており、赤線の飛跡と共に IP (Primary Vertex) から生じている。ハドロンは更に図右上と左下で崩壊し、Secondary Vertex を残す。青線はこの Secondary Vertex 由来の飛跡である。

崩壊点検出では、二本以上の飛跡について交点を求めるフィッティングを用いている。この時、フィッティングの健全性は χ^2 値によって把握している。

まず Primary Vertex の再構成を Tear-Down 法によって行う。具体的には予想されるビームスポットと事象中の全飛跡についてフィッティングを行い、 χ^2 値が一定以下になるまで χ^2 値への寄与が大きい飛跡を一本ずつ取り除くことによって、残った飛跡を Primary Vertex 由来であると判定している。次に Secondary Vertex の再構成を Build-Up 法によって行う。ここでは、Primary Vertex に含まれていない飛跡について、二本の飛跡(飛跡対)の全ての組み合わせを作りフィッティングを行う。そのようにして得られた χ^2 値と運動量の方向、不变質量などをカットベースに判定し、Secondary Vertex についての飛跡対を選別する。更に、この飛跡対について飛跡を加えていくことで Secondary Vertex を再構成している。

崩壊点検出で得られた崩壊点を用いて、ジェットクラスタリングが行われる。ジェットクラスタリングでは Durham アルゴリズム [13] を使用し、事象中の中性粒子を含めたより多くの情報を用いてジェット中の粒子をクラスター化している。

また、そのようにクラスター化されたジェット中の粒子についてフレーバータギングが行われる。フレーバータギングでは Boosted Decision Trees (BDTs) を用いて親粒子のフレーバーを識別している。

以上がジェットの再構成である。様々な物理解析において、ジェットの個数やそのフレーバーの識別は信号事象と背景事象の区別や物理解析などに使用されている。したがって、ジェットの再構成の性能向上はあらゆる物理解析の性能向上と直結していると言える。

1.6 本研究の目的

本研究の目的は、深層学習を使用して 1.5.3 項で紹介した崩壊点検出を開発・改善することである。ILC では現在 LCFIPlus 内の崩壊点検出が使用されているが、Primary Vertex や Secondary Vertex の選別に人が定めた閾値が多く含まれており、カットベースな評価が行われている。このような人が定めた閾値は最適ではなく、何らかの情報を欠損してしまっている可能性がある。本研究では深層学習を用いたパターン認識の技術から、新しい崩壊点検出アルゴリズムを提案しより柔軟な識別を行うことを目標とする。

また、この研究は深層学習を用いて事象再構成を改善していくプロジェクトの一つである。最終的には概念図 1.7 に示すように、全ての事象再構成アルゴリズムを深層学習や機械学習に置き換えることを目指している。これまで ILC の事象再構成ではあまり深層学習は使われておらず、特に崩壊点検出に関しては本研究が殆ど初めての試みである。

したがって、本研究のソフトウェア開発としての目的は、LCFIPlus への導入や次世代の LCFIPlus への深層学習実装における起点となる事である。ここでは、深層学習の実装・構築から iLCSoft への導入を行い、ILC 研究における深層学習導入の先駆けとなることを目指す。

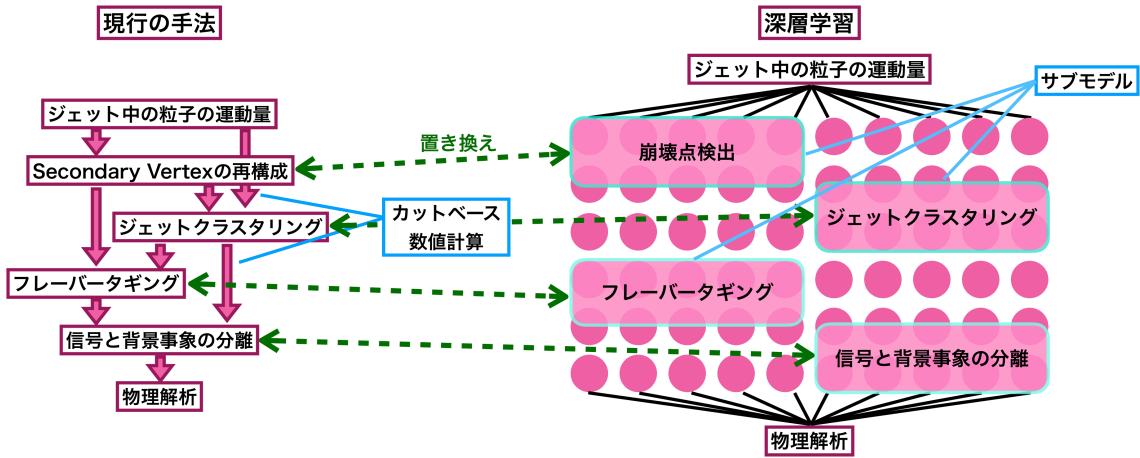


図 1.7: 深層学習によるジェットの再構成。図左は現行の手法、図右は目標としている深層学習を用いた再構成手法の概念図である。現行の手法では多くの場合、数値計算やそこから得られる変数に閾値を課してカットベースな評価が行われている。現在は深層学習への置き換えの第一段階として、それぞれの役割に特化したサブモデルと入れ替えを行い性能の改善を図っている。

1.7 本論文の流れ

本章と 2 章は本論文の導入である。

2 章では本論文の核となる技術である深層学習について解説を行う。ここでは、本研究を理解する為に必要な技術領域や背景理論について簡潔な導入を行い、3 章以降では、この 2 章の内容を前提とした議論を行う。ただし深層学習に関しての種々のテクニックについては経験則によるものが多いため、本研究で使用したものについては 2 章では説明せず、3 章で述べる。また、具体的な実装に関しても同様に 2 章では記載せず、付録 A にまとめる事とする。

3 章と 4 章、5 章は本論文の本題である。

3 章では本研究で使用するデータと作成した深層学習のネットワークについて、その構造の解説や評価を行う。また、深層学習を使用した崩壊点検出についての発想や計算環境についても 3 章で述べる。

4 章では 3 章で作成したネットワークを用いた崩壊点検出について、アルゴリズムや各種最適化、簡単な評価を行う。

5 章ではその様にして得られた崩壊点検出と、LCFIPlus で使用されている現行の崩壊点検出との比較について述べる。

6 章は本論文の結論である。ここでは、本研究をまとめると共に今後の展望について述べる。

第 2 章

深層学習

本章では、深層学習 (Deep Learning, DL) について述べる。

まず深層学習の導入として、2.1 節で機械学習 (Machine Learning, ML) と深層学習の概要について簡単に紹介する。

次に 2.2 節では、深層学習を理解する上で前提となるパーセプトロン (Perceptron) というネットワークを解説する。パーセプトロンは、後述するニューラルネットワークの先駆けとなる技術である。

2.3 節では、深層学習の基礎技術であるニューラルネットワーク (Neural Network, NN) を導入する。主に、2.3.1 項でニューラルネットワークを構築するために必要な計算手順について、2.3.2 項でニューラルネットワークの学習に関して重要な技術要素についてそれぞれ説明する。

深層学習は、2.3 節までの基盤的な技術を使用するだけでも様々な問題に対応できるが、扱うデータや問題の性質によって、更に応用的な使い方が求められる。本研究ではそのような応用技術を幾つか用いてネットワークの作成を行なっている。その為、2.4 節や 2.5 節ではそのような深層学習の応用技術の解説を行う。

2.4 節では、系列データを取り扱うためのリカレントニューラルネットワーク (Recurrent Neural Network, RNN) を導入する。2.5 節では、近年注目されている注意機構 (Attention) と呼ばれる技術について説明する。

最後に 2.6 節にて、ニューラルネットワークが持つハイパラメータについてまとめる。

本章の作成にあたり、参考文献 [14, 15, 16] を使用した。

2.1 機械学習と深層学習

深層学習とは、機械学習の技術の一つである。本節ではまず、この機械学習について簡単に説明し、その後、機械学習における深層学習の位置付けを述べる。

機械学習とは、データに現れるパターンや統計情報を計算機 (学習器) に「学習」させることによって、逐一プログラミングをすることなく未知の問題に対応させる為の技術である。こ

これは人間の持つ知性を機械に実現する、人工知能 (Artificial Intelligence, AI) に関する研究の一分野であると言える。このような研究は、1956 年のダートマス会議 [17] から始まり、現在は第三期の AI ブームと言われている。機械学習は、機械 (計算機) が独自に未知の問題を解く為の技術や手法の総称であるが、問題に対するアプローチの仕方によって、教師あり学習、教師なし学習、強化学習などに分類することができる (図 2.1)。

- 教師あり学習

教師あり学習とは、訓練データ (Training data) と呼ばれる正解がラベル付けされたデータを用い、学習器の出力を正解に近付けるように学習器を更新していく手法である。主にクラス分類を行う分類問題や、連続値を予測する回帰問題などの問題を解くことができる。具体的な例としてサポートベクターマシン (Support Vector Machine, SVM[18, 19]) や決定木などが用いられる。

- 教師なし学習

教師なし学習とは、訓練データを用いず、データの持つ数学モデルや構造を抽出する技術である。主にクラスタリングや次元圧縮などに使用され、代表的な手法は、k 平均法や主成分分析などである。

- 強化学習

強化学習とは、環境とのやり取りから報酬を受け取り、エージェントを構築していく手法である。学習は報酬を最大化するように進み、教師あり学習の一分野のようにみなす事も出来るが、強化学習は一連の行動に対しての報酬を考慮する点で異なる。強化学習は様々な分野で使用されているが、主に長期的な戦略が必要となるゲームなどの領域で用いられている。

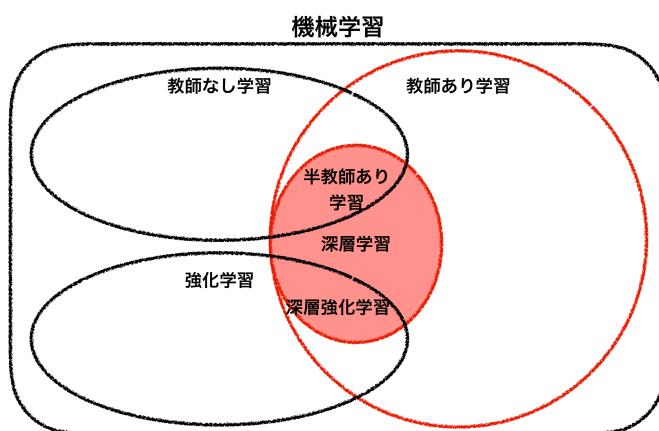


図 2.1: 機械学習の中の深層学習の位置付け。深層学習は基本的に教師あり学習であるが、半教師あり学習分類やクラスタリングなど応用的な研究が行われている。

深層学習は、このような機械学習の中で、基本的には、回帰問題や分類問題などを解く教師

あり学習に分類される。しかし近年では、半教師あり学習やディープクラスタリング、深層強化学習といった様々な技術的応用が提案されている。次節以降では、この深層学習の基盤技術について紹介する。

2.2 パーセプトロン

パーセプトロンは深層学習の基礎となる技術であり、1958年、Rosenblattによって提案された[20]。ここでは、このパーセプトロンについて解説することで、次節のニューラルネットワークへの導入とする。

2.2.1 単純パーセプトロン

パーセプトロンとは、情報を伝達するネットワークである。ここでこのネットワークとは、ある情報を受け取り、それを後方へ伝達するような構造のことを言うものとする。まず、最も簡単なパーセプトロンとして、図2.2のような構造を考える。図2.2は、二つの入力 x_1, x_2 を受け取り、一つの出力 y を行なっているネットワークである。このような入力や出力の数や入力や出力そのものの事をノードやニューロンという。また、図2.2のように、ただ入力と出力のみを持っているパーセプトロンを特に単純パーセプトロン (Simple Perceptron) という。

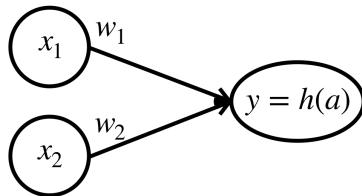


図2.2: 単純パーセプトロン。入力を x_1, x_2 、出力を y と置いた。また、各ノードの繋がりを矢印で表現し、同時にそれぞれの線は学習可能な重みを表している。

単純パーセプトロンの情報伝達は、簡単な計算で定義される。出力 y は、 x_1, x_2 とそれぞれの重み w_1, w_2 を用いて、

$$\begin{aligned} y &= h(a) \\ a &= w_1x_1 + w_2x_2 \end{aligned} \tag{2.1}$$

と計算される。ここで、出力 y は関数 h によって変換されている。このような関数を活性化関数 (Activation function) という。特に単純パーセプトロンでは、活性化関数 h としてヘヴィサイドの階段関数 (図2.3) を用いる。

その閾値を θ とすると、出力は更に、

$$y = h(a) = \begin{cases} 0 & (a = w_1x_1 + w_2x_2 \leq \theta) \\ 1 & (a = w_1x_1 + w_2x_2 > \theta) \end{cases} \tag{2.2}$$

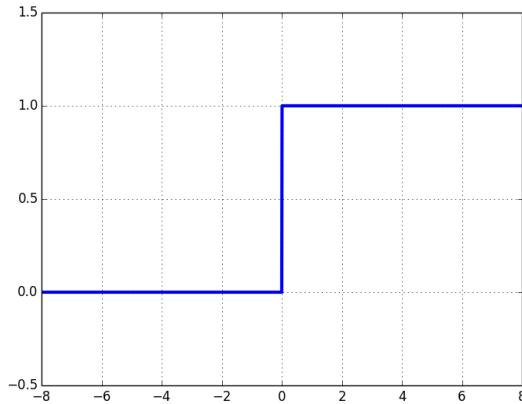


図 2.3: ヘヴィサイドの階段関数

と書ける。この時、単純パーセプトロンはある一定値 θ までは”0”、それ以上であれば”1”を返す二値信号のネットワークであると考えることが出来る。パーセプトロンやニューラルネットワークにおいて、学習可能なパラメータは重み w_1, w_2 であり、これら重みを更新していく操作を学習（トレーニング、Training）という。

2.2.2 多層パーセプトロン

単純パーセプトロンは線形な問題を解く事しか出来なかつたが、層を重ねることで非線形に対応できるという点で非常に高い発展性を持っていた [21]。そのように単純パーセプトロンを重ねたネットワークの事を多層パーセプトロン (Multi Layer Perceptron, MLP) という。多層パーセプトロンは図 2.4 のように表現出来る。

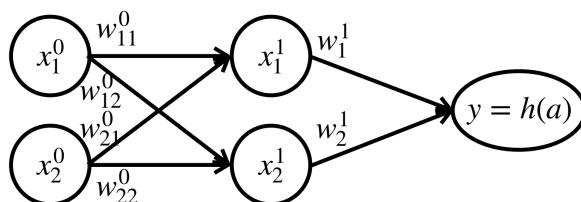


図 2.4: 多層パーセプトロン。図 2.2 と同様にノードを繋ぐそれぞれの線は学習可能な重みを表現している。前方のノードは後方のノードの全てと接続されている点にも注意が必要である。

多層パーセプトロンは単純パーセプトロンとは異なり、入力、出力以外に、中間層（隠れ層）を持っている。ここで、中間層を x_1^1, x_2^1 と置くと、単純パーセプトロンと同様に入力 x_1^0, x_2^0 とそれぞれの重み $w_{11}^0, w_{12}^0, w_{21}^0, w_{22}^0$ を用いて

$$\begin{aligned} x_1^1 &= w_{11}^0 x_1^0 + w_{21}^0 x_2^0 \\ x_2^1 &= w_{12}^0 x_1^0 + w_{22}^0 x_2^0 \end{aligned} \tag{2.3}$$

と計算でき、また出力 y についても、 x_1^1, x_2^1 とそれぞれの重み w_1^1, w_2^1 を用いて、

$$\begin{aligned} y &= h(a) \\ a &= w_1^1 x_1^1 + w_2^1 x_2^1 \end{aligned} \tag{2.4}$$

となる。

以後、特に断らない場合はあるノード x 、重み w について、層の深さ、前後のノードを次のように表現する。

$$\begin{aligned} x_{(\text{ノード番号})}^{(\text{層の深さ})} \\ w_{(\text{前のノード番号}) (\text{後ろのノード番号})}^{(\text{層の深さ})} \end{aligned} \tag{2.5}$$

多層パーセプトロンは、学習の手法や層を重ねるに連れて重みが更新出来なくなる勾配消失問題など様々な課題を抱えていた。次節ではこれらの問題を後述する誤差逆伝播法(Backpropagation) や活性化関数によって解決したニューラルネットワークについて解説する。

2.3 ニューラルネットワーク

本節ではニューラルネットワークについて解説を行う。ただし、ここでは計算手法や言葉の定義についてのみ述べる。実装方法については現在様々なフレームワーク [22, 23, 24, 25] があり、それぞれで実装の仕方が異なっている。本研究では tensorflow-keras を用いた。^{*1}

ニューラルネットワークに関する技術は「ニューラルネットワークの構造」についてと「ニューラルネットワークの学習」についてに大きく分けられると考えている。前者は主に入力から出力までのネットワークの構築を、後者は構築されたネットワークの重み更新についての技術である。

2.3.1 ニューラルネットワークの構造

ニューラルネットワークは様々な技術によって支えられているが、その基本構造は前節の多層パーセプトロンと全く同じである。ニューラルネットワークと多層パーセプトロンとの大きな構造の違いは活性化関数である。ニューラルネットワークでは様々な活性化関数が提案されており^{*2}、これが勾配消失問題を解消する鍵となっている。活性化関数は重み更新のために微分可能な関数である必要があるが、どのような関数を選ぶかはユーザーに委ねられている。勾配消失や重み更新についての詳しい解説は 2.3.2 節で行う。以下に活性化関数の例を示す(図 2.5)。

^{*1} 具体的なコードに関しては付録 A を参照。

^{*2} 多層パーセプトロンはニューラルネットワークの内、活性化関数に階段関数を使った特別なネットワークであると再定義できる。

- 階段関数

$$h(a) = \begin{cases} 0 & (a \leq \theta) \\ 1 & (a > \theta) \end{cases} \quad (2.6)$$

- シグモイド関数

$$h(a) = \frac{1}{1 + \exp(-a)} \quad (2.7)$$

- \tanh 関数

$$h(a) = \tanh(a) \quad (2.8)$$

- ReLU (Rectified Linear Unit, ランプ) 関数 [26]

$$h(a) = \begin{cases} 0 & (a \leq \theta) \\ a & (a > \theta) \end{cases} \quad (2.9)$$

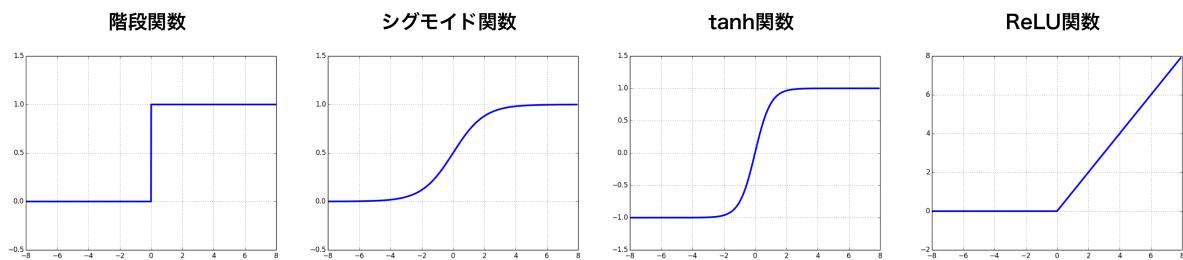


図 2.5: 活性化関数

一般に、ニューラルネットワークは多層パーセプトロンと同様の構造であるので、図 2.6 のように表現出来る。

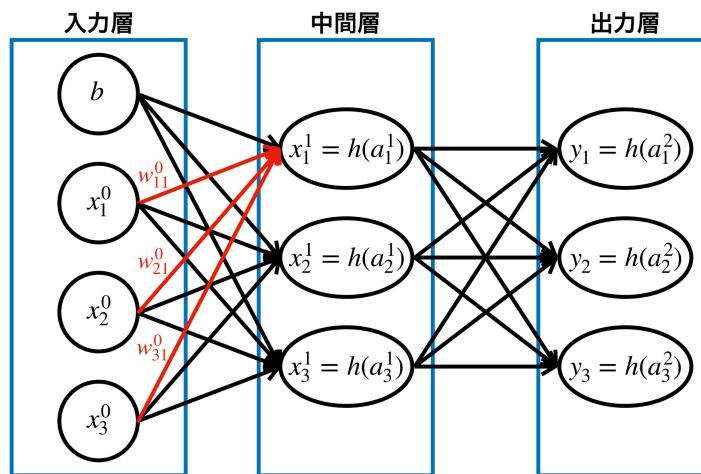


図 2.6: ニューラルネットワーク。基本構造は多層パーセプトロンと同様であるとわかる。赤線は中間層 x_1^1 へ入力される重みである。

ここで、中間層 x_1^1 は入力 x_1^0, x_2^0, x_3^0 とそれぞれの重み $w_{11}^0, w_{21}^0, w_{31}^0$ を用いて、

$$\begin{aligned} x_1^1 &= h(a_1^1) \\ a_1^1 &= w_{11}^0 x_1^0 + w_{21}^0 x_2^0 + w_{31}^0 x_3^0 + b_1^0 \end{aligned} \quad (2.10)$$

と計算できる。また、バイアスとして b を導入している。これはパーセプトロンの閾値 θ に対応している。中間層 x_2^1, x_3^1 についても同様に、

$$\begin{aligned} x_2^1 &= h(a_2^1) \\ a_2^1 &= w_{12}^0 x_1^0 + w_{22}^0 x_2^0 + w_{32}^0 x_3^0 + b_2^0 \\ x_3^1 &= h(a_3^1) \\ a_3^1 &= w_{13}^0 x_1^0 + w_{23}^0 x_2^0 + w_{33}^0 x_3^0 + b_3^0 \end{aligned} \quad (2.11)$$

と書ける。

また、これら x_1^1, x_2^1, x_3^1 の計算は行列とベクトルを用いて、より簡潔に表現できる。

$$\begin{aligned} \mathbf{x}^1 &= \begin{pmatrix} x_1^1 \\ x_2^1 \\ x_3^1 \end{pmatrix} = h(\mathbf{a}^1) \\ \mathbf{a}^1 &= \begin{pmatrix} a_1^1 \\ a_2^1 \\ a_3^1 \end{pmatrix} = W^0 \mathbf{x}^0 + \mathbf{b}^0 = \begin{pmatrix} w_{11}^0 & w_{21}^0 & w_{31}^0 \\ w_{12}^0 & w_{22}^0 & w_{32}^0 \\ w_{13}^0 & w_{23}^0 & w_{33}^0 \end{pmatrix} \begin{pmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{pmatrix} + \begin{pmatrix} b_1^0 \\ b_2^0 \\ b_3^0 \end{pmatrix} \end{aligned} \quad (2.12)$$

行列 W^0 とベクトル \mathbf{b}^0 は学習可能な重みであり、得られた新たな状態ベクトル \mathbf{a}^1 は微分可能な任意の活性化関数 h によって、中間層 \mathbf{x}^1 へと変換される。以下、これを繰り返すことによって、ネットワークは構築されている。

出力層における活性化関数は一般に回帰問題では恒等関数を、分類問題ではソフトマックス関数と呼ばれる関数を使用する。回帰問題において、最終的な出力は数値（連続値）であるため、恒等関数によって変換を行わず、そのまま出力することが一般的である。

$$y_k = h(a_k^2) = a_k^2 \quad (2.13)$$

一方で、分類問題では、最終的な出力は分類されたクラスとなるため、以下のようない softmax 関数を使用する。

$$y_k = h(a_k^2) = \frac{\exp(a_k^2)}{\sum_{i=1}^N \exp(a_i^2)} \quad (2.14)$$

この softmax 関数は、分母が総和、分子がその一要素の形をしており、 y_k を k について足し合わせると 1 になることがわかる。このことから、出力 y_k は k 番目のクラスについての確率として解釈でき、分類問題において、どのクラスにどの程度該当するかを表現することに相当している。

前述したように、これは最も基本的なニューラルネットワークであり、このような全結合（Fully connected, Dense）な層を重ねたネットワークを順伝播型（フィードフォワード）ニューラルネットワーク（Feedforward Neural Network）という。

2.3.2 ニューラルネットワークの学習

教師あり学習であるニューラルネットワークにおける学習は、損失関数 (コスト関数, Loss function) を最小化するように、重みを更新していく事で行われる。損失関数とは、訓練データの正解ラベルとネットワークの出力がどの程度離れているかを計算するための関数である。この損失関数は取り組む問題や訓練データの性質によって適切に選択する必要がある。ここでは、よく使用される損失関数として以下の二つを挙げる。

- 交差エントロピー誤差

$$L = - \sum_k^N t_k \log (y_k) \quad (2.15)$$

- 平均二乗誤差

$$L = \frac{1}{N} \sum_k^N (t_k - y_k)^2 \quad (2.16)$$

t_k, y_k はそれぞれ k 番目の正解ラベルとクラスの出力 (確率や値) を示している。分類問題については交差エントロピー誤差が、回帰問題については平均二乗誤差が主に使用される。

分類問題において、正解ラベル t は、あるクラスに関して 0 か 1 かのベクトル (one-hot vector) で表現されることが一般的である。例えば、赤、青、緑について分類を行う場合 (3 クラス分類という)、赤を $(1, 0, 0)$ 、青を $(0, 1, 0)$ 、緑を $(0, 0, 1)$ と定義する。また、ネットワークの出力 y はどのクラスに属するかの確率となっている。例えば、赤、青、緑がそれぞれ 80 %、10 %、10 % の場合は出力 y は $(0.8, 0.1, 0.1)$ と書ける。したがって、正解ラベルを赤とすると損失関数 L は

$$\begin{aligned} L &= - \sum_k^3 t_k \log (y_k) \\ &= -t_1 \log y_1 - t_2 \log y_2 - t_3 \log y_3 \\ &= -1 \cdot \log 0.8 \\ &= 0.22314... \end{aligned} \quad (2.17)$$

と計算される。

回帰問題において、出力 y 、正解ラベル t は共に連続値であるため、平均二乗誤差のような二つの差を用いる損失関数が一般的である。

前述したように、ネットワークの学習はこの損失関数を最小化するように進む。今、損失関数は変数 y_k の関数で表現出来ており、このような関数の最小値を求めるためには、単に変数 y_k を用いて偏微分を行い勾配を求めれば良い。計算機において、このような勾配を求め、徐々に関数を最小化していく手法を勾配降下法 (Gradient Descent Method) という。勾配降下法において、次のステップの変数 y'_k は次のように計算される。

$$y'_k = y_k - \eta \frac{\partial L}{\partial y_k} \quad (2.18)$$

ここで、ステップ幅を決定する定数 η をニューラルネットワークにおいて学習率 (learning rate) という。学習率は 0.001 などの定数を問題やネットワークによって適切に選ぶ必要がある。このようなネットワークについて更新されない初期設定のパラメータをハイパーパラメータという。ハイパーパラメータについては後の 2.6 節で述べる。

ニューラルネットワークにおいて、各重みを更新するための勾配は連鎖律 (chain rule) を用いて計算される。これは更新する重みと最小化される損失関数の間に output 層と活性化関数が存在しているためである。^{*3} 具体的には、ある重み行列 W に対して、勾配降下法、連鎖律を考慮すると、次のステップの重み行列 W' は

$$\begin{aligned} W' &= W - \eta \frac{\partial L}{\partial W} \\ &= \begin{pmatrix} w_{11} & w_{21} & w_{31} \\ w_{12} & w_{22} & w_{32} \\ w_{13} & w_{23} & w_{33} \end{pmatrix} - \eta \begin{pmatrix} \frac{\partial L}{\partial w_{11}} & \frac{\partial L}{\partial w_{21}} & \frac{\partial L}{\partial w_{31}} \\ \frac{\partial L}{\partial w_{12}} & \frac{\partial L}{\partial w_{22}} & \frac{\partial L}{\partial w_{32}} \\ \frac{\partial L}{\partial w_{13}} & \frac{\partial L}{\partial w_{23}} & \frac{\partial L}{\partial w_{33}} \end{pmatrix} \\ &= W - \eta \frac{\partial L}{\partial y} \frac{\partial y}{\partial a} \frac{\partial a}{\partial W} \end{aligned} \quad (2.19)$$

と計算される。

このような最適化問題に関して、いくつかのアルゴリズムが提案されている。現在は確率的勾配降下法 (Stochastic Gradient Descent, SGD[27]) やそれを基礎とした RMSProp[28]、Adam[29] などの手法がよく使用されている。

学習方法における、ニューラルネットワークと多層パーセプトロンの大きな違いは、重みの更新を出力層から逆伝播させる誤差逆伝播法 (Backpropagation[30]) という手法の有無である。再度、図 2.6 を考える。全ての出力、中間層を行列計算を用いて記述すると、

$$\begin{aligned} \mathbf{a}^1 &= \begin{pmatrix} a_1^1 \\ a_2^1 \\ a_3^1 \end{pmatrix} = W^0 \mathbf{x}^0 + \mathbf{b}^0 = \begin{pmatrix} w_{11}^0 & w_{21}^0 & w_{31}^0 \\ w_{12}^0 & w_{22}^0 & w_{32}^0 \\ w_{13}^0 & w_{23}^0 & w_{33}^0 \end{pmatrix} \begin{pmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{pmatrix} + \begin{pmatrix} b_1^0 \\ b_2^0 \\ b_3^0 \end{pmatrix} \\ \mathbf{x}^1 &= h(\mathbf{a}^1) \\ \mathbf{a}^2 &= \begin{pmatrix} a_1^2 \\ a_2^2 \\ a_3^2 \end{pmatrix} = W^1 \mathbf{x}^1 = \begin{pmatrix} w_{11}^1 & w_{21}^1 & w_{31}^1 \\ w_{12}^1 & w_{22}^1 & w_{32}^1 \\ w_{13}^1 & w_{23}^1 & w_{33}^1 \end{pmatrix} \begin{pmatrix} x_1^1 \\ x_2^1 \\ x_3^1 \end{pmatrix} \\ y_k &= \sigma(a_k^2) = \frac{\exp(a_k^2)}{\sum_{i=1}^n \exp(a_i^2)} \end{aligned} \quad (2.20)$$

と書ける。ここで、出力部分のソフトマックス関数を σ と書いた。

ある重み w_{11}^1 について考える。損失関数 L の重み w_{11}^1 による偏微分は、連鎖律を考慮して、

$$\begin{aligned} y_1 &= \sigma(a_1^2) \\ a_1^2 &= w_{11}^1 x_1^1 + w_{21}^1 x_2^1 + w_{31}^1 x_3^1 = \sum_{i=1}^N w_{i1}^1 x_i^1 \end{aligned} \quad (2.21)$$

^{*3} 損失関数はクラスの出力の関数であり、クラスの出力は活性化関数によって計算され、活性化関数は出力層の関数である。

より、

$$\begin{aligned}\frac{\partial L}{\partial w_{11}^1} &= \frac{\partial L}{\partial y_1} \frac{\partial y_1}{\partial a_1^2} \frac{\partial a_1^2}{\partial w_{11}^1} \\ &= \frac{\partial L}{\partial y_1} \frac{\partial \sigma(a_1^2)}{\partial a_1^2} x_{11}^1\end{aligned}\quad (2.22)$$

と計算できる。

ここで、勾配の計算に活性化関数の偏微分が常に積の形で含まれていることがわかる。この活性化関数の偏微分が 0 になり、そこから抜け出せなくなると、その勾配は常に 0 になり消失してしまう、これが勾配消失である。勾配消失に陥った場合は、重みが適切に更新されず、学習が不十分になってしまう。このような問題は活性化関数を変更することによって改善され、現在は ReLU 関数がよく用いられている。

また、更に浅い層の重み w_{11}^0 について考えると、

$$\begin{aligned}x_1^2 &= h(a_1^1) \\ a_1^1 &= w_{11}^0 x_1^0 + w_{21}^0 x_2^0 + w_{31}^0 x_3^0 = \sum_{i=1}^N w_{i1}^1 x_i^0\end{aligned}\quad (2.23)$$

より、

$$\begin{aligned}\frac{\partial L}{\partial w_{11}^0} &= \sum_k^N \frac{\partial L}{\partial y_k} \frac{\partial y_k}{\partial a_k^2} \frac{\partial a_k^2}{\partial x_1^2} \frac{\partial x_1^2}{\partial a_1^1} \frac{\partial a_1^1}{\partial w_{11}^0} \\ &= \sum_k^N \frac{\partial L}{\partial y_k} \frac{\partial y_k}{\partial a_k^2} \frac{\partial a_k^2}{\partial x_1^2} \frac{\partial h(a_1^1)}{\partial a_1^1} x_{11}^0 \\ &= \sum_k^N \left(\frac{\partial L}{\partial y_k} \frac{\partial y_k}{\partial a_k^2} w_{1k}^1 \right) \frac{\partial h(a_1^1)}{\partial a_1^1} x_{11}^0\end{aligned}\quad (2.24)$$

と計算できる。この計算は更に層を重ねた場合でも同様の手順で行うことができる。

重みの更新は基本的に全ての訓練データを使用するのではなく、訓練データをいくつかの塊に分け、その塊について損失関数を計算することで行われる。このような手法をミニバッチ学習と呼ばれる。(訓練データ全てを用いたものをバッチ学習という。) ミニバッチ学習に使用されるデータの数をミニバッチサイズ(あるいは単にバッチサイズ) という。これも後述するハイパーパラメータの一つである。ミニバッチ学習はバッチ学習と比較して二つの利点が存在する。

一つは膨大なデータを直接処理しなくても良いという点である。一般に深層学習で使用されるデータは非常に膨大であり、GPUなどのメモリに乗らない場合があるが、ミニバッチ学習ではこれを回避することができる。

もう一つは学習が停滞しづらいという点である。訓練データと比較してサイズの小さいミニバッチは上記の勾配が 0 になりづらく、局所的な最小点での学習の停滞を回避することができる。

ただしミニバッチ学習のバッチサイズが小さくなつた場合には、損失関数が平均化されず学習が不安定になる（収束しなくなる）という問題が生じる場合がある。^{*4}

2.3.3 ディープニューラルネットワーク

ディープニューラルネットワーク (Deep Neural Network, DNN) という言葉の定義は非常に曖昧である。^{*5} 2.1節で述べたように、現在は第三期 AI ブームであると言われている。これは上述してきた技術的成熟に加え、計算機の性能が向上したことにより、より層を重ねた（深い）ニューラルネットワークの学習が可能になった結果であると言える。2006 年の Hinton らによる auto-encoder[31] や 2014 年に Ian によって提案された敵対的生成ネットワーク (Generative Adversarial Network, GAN[32]) など、様々な発展的な応用がなされ、現在においても毎年のように新しいネットワークが提案されている。本節で述べたニューラルネットワークはその基盤技術の一部である。次節以降では、系列データを扱うためのニューラルネットワークの応用について紹介する。

2.4 リカレントニューラルネットワーク

前節で紹介したようなフィードフォワードニューラルネットワークは系列データを扱う際、重み行列が固定的な大きさでしか保持出来ないという点と直前の系列に依存した学習が出来ないという点に関して課題を抱えている。これらの課題を解決するために提案されたのが、リカレントニューラルネットワークというネットワーク構造である。本節では、このリカレントニューラルネットワークについて解説を行う。リカレントニューラルネットワークは主に系列データ、特に時系列データを取り扱うためのネットワークである。このような時系列に関するニューラルネットワークは自然言語処理などの分野で発展し、音声認識や機械翻訳といった技術に使用されている。リカレントニューラルネットワークの構造は、フィードフォワードニューラルネットワークと比較すると複雑であるが、要素計算は全結合であり、基本的にはその組み合わせで理解できる。2.4.1 項では、そのようなリカレントニューラルネットワークの構造と学習について述べる。その後、リカレントニューラルネットワークの抱える問題と、ゲート (Gate) と記憶セル (Cell) 呼ばれる技術によってその問題を克服した長・短期記憶 (Long Short-Term Memory, LSTM[33]) というネットワークを 2.4.2 項と 2.4.3 項でそれぞれ紹介する。

2.4.1 リカレントニューラルネットワークの構造と学習

リカレントニューラルネットワークの構造は、これまでのフィードフォワードニューラルネットワークとは大きく異なる。そのネットワーク構造はいくつかの表現方法が存在している

^{*4} バッチサイズが 1 (1 サンプルのみ) のミニバッチ学習をオンライン学習という。

^{*5} 少なくとも私ははつきりとした定義を存じない。

が、本論文では時間について展開した図で書くこととする。図 2.7 の左側が時間について展開していない図、右側が時間について展開した図である。左側では、時間についての構造をループで表現し、任意の時間 t についての入力 x_t と出力 h_t を持つネットワークとして表現している。右側では、時間についての構造を展開し、出力や入力を系列情報とともに表現している。これまでのフィードフォワードネットワークは情報の伝達を左右に描いていたが、このリカレントニューラルネットワークは上下に書き、系列の流れを左右で表現されることが多い。図 2.7 の右側では、出力 h が二つ存在し、上に進むものを出力に、右に進むものが次の系列への入力になっていることがわかる。このように、現在の出力を次の系列の入力として使うことで、直前の系列情報への依存性を導入している。

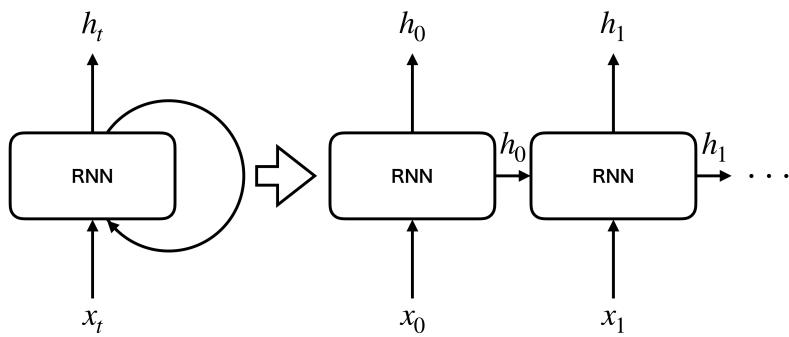


図 2.7: リカレントニューラルネットワーク。図左は時間について展開していない描画方法、図右は時間について展開した描画方法である。下方から入力され、上方に出力している。横方向は系列を表現しており、入力 x_t は全ての系列で同じ形状でなければならない。

これまでと同様に明示的にネットワークの重みを描画すると、図 2.8 のようになる。図 2.7 の RNN に当たる部分が展開され、重みを線で表現した図になっている。図より、一つ前に系列の出力 \mathbf{h}_{t-1} と現在の系列の入力 \mathbf{x}_t を用いて、現在の系列の出力 \mathbf{h}_t が生成されていることがわかる。 \mathbf{h} についての赤い線に関する重み行列を W_h 、 \mathbf{x} についての黒い線に関する重み行列を W_x と置くと、出力 \mathbf{h}_t は次のように計算できる。

$$\begin{aligned} \mathbf{h}_t &= \tanh(W_h \mathbf{h}_{t-1} + W_x \mathbf{x}_t) \\ \mathbf{a}_t &= \begin{pmatrix} w_{h,11} & w_{h,21} & w_{h,31} \\ w_{h,12} & w_{h,22} & w_{h,32} \\ w_{h,13} & w_{h,23} & w_{h,33} \end{pmatrix} \begin{pmatrix} h_{t-1,1} \\ h_{t-1,2} \\ h_{t-1,3} \end{pmatrix} + \begin{pmatrix} w_{x,11} & w_{x,21} & w_{x,31} \\ w_{x,12} & w_{x,22} & w_{x,32} \\ w_{x,13} & w_{x,23} & w_{x,33} \end{pmatrix} \begin{pmatrix} x_{t,1} \\ x_{t,2} \\ x_{t,3} \end{pmatrix} \end{aligned} \quad (2.25)$$

リカレントニューラルネットワークでは活性化関数として \tanh 関数を使用している。前述したように、個々の要素計算はフィードフォワードニューラルネットワークの様に全結合で構成されていることがわかる。ここで、非常に重要な性質として、学習可能な重み行列 W_h, W_x は全ての系列 t について同一のものであり、大きさが不变であることに注意する。

このように再帰的に重み行列を使用することで、行列の大きさが可変でないという性質を回

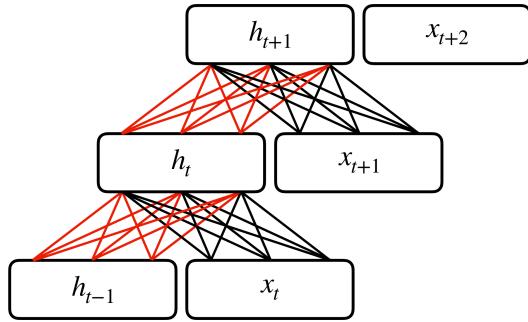


図 2.8: リカレントニューラルネットワークの重みの明示的な表現。赤線で示した重みは隠れ状態 h についての重み W_h 、黒線で示した重みは隠れ層 x についての重み W_x である。図 2.7 で示した”RNN”部はここでは線に相当している。ネットワークは上に積み重なっているように表現しているが、実際にはこれらの重み行列は全ての系列で同一のものである。

避し、系列情報を取り入れることに成功している。また、このような入力の系列長についての柔軟性は、長さが不定であるリアルタイムな時系列データを扱えるという点で重要である。

リカレントニューラルネットワークの出力方法は、問題によっていくつかのパターンが存在する。(図 2.9) 例えば、語句の分類の様な問題の場合は、一つの入力に対して、一つの出力を得る Many to Many という出力の作り方を行う。また、機械翻訳のデコーダーなど、一つの入力に対して、複数の出力を得たい場合は One to Many を用いる。最後に、感情分析の様に複数の入力に対して、一つの出力を得たい場合は Many to One を用いる。

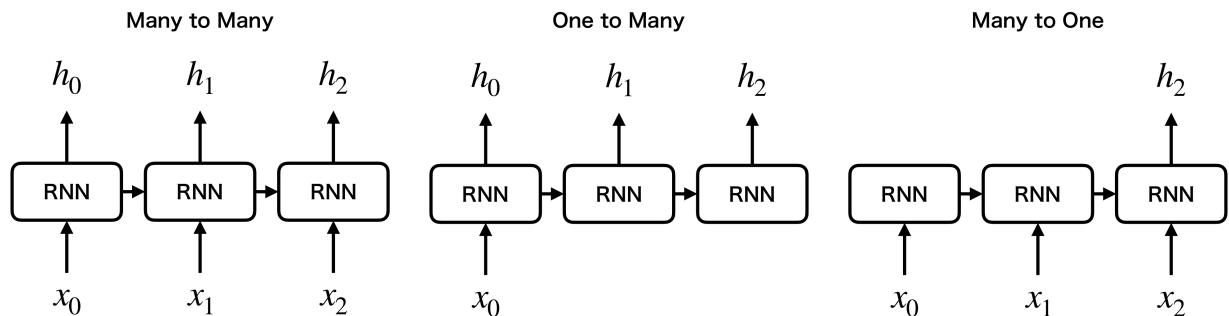


図 2.9: リカレントニューラルネットワークの出力方法

リカレントニューラルネットワークの学習は基本的にフィードフォワードニューラルネットワークと同様であるが、図 2.8 に見る様にネットワークは系列にしたがって深くなっている。この為、重み更新はこの系列を遡って行う必要がある。このような誤差逆伝播法の事を Backpropagation Through Time (BPTT) という。実際には計算リソースの削減のため、Truncated BPTT という、時系列方向に適当な長さで切り取り計算を行う手法が使用される。

2.4.2 リカレントニューラルネットワークの問題点

リカレントニューラルネットワークは時間方向に展開し、それを遡ることによって学習を行っているため、系列の長さに依存して非常に深いネットワークが構築される。したがって、リカレントニューラルネットワークは真に深いネットワークであると言えるが、深い層からの勾配は非常に消失あるいは爆発しやすく、容易に勾配消失・爆発を招いてしまうという問題が生じている。勾配消失はリカレントニューラルネットワークの活性化関数である \tanh 関数に起因している。2.3.2 項で解説したように、誤差逆伝播法は連鎖律によって計算され、その計算には活性化関数の微分が含まれている。ここで \tanh 関数の微分は、

$$\begin{aligned} \frac{\partial y(x)}{\partial x} &= \frac{\partial}{\partial x} \tanh(x) = \frac{1}{\cosh^2(x)} \\ &= 1 - \frac{\cosh^2(x) - 1}{\cosh^2(x)} = 1 - \frac{\sinh^2(x)}{\cosh^2(x)} = 1 - \tanh^2(x) = 1 - y^2 \end{aligned} \quad (2.26)$$

と計算される。

$1 - y^2$ は、 $y = 0$ 以外において常に 1 より小さい値を取ってしまう。リカレントニューラルネットワークでは系列長に応じて、この 1 より小さい値 ($1 - y^2$) が複数回掛けられてしまうため、勾配消失が生じやすくなっている。同時にリカレントニューラルネットワークでは、連鎖律によって系列長に応じて同じ重み行列 W_h を複数回掛けしており、この重み行列の値に応じて、勾配が発散あるいは消失してしまうことが考えられる。

また、リカレントニューラルネットワークは時系列上の複数の入出力から、矛盾した重み更新を受け取ってしまう入力重み衝突、出力重み衝突という問題も抱えている。

更に、リカレントニューラルネットワークはその構造上、長期的な系列情報を保持できないという課題も存在している。

以上のような問題を解決するため、ゲートとセルと呼ばれる技術を導入したものを、ゲート付きリカレントユニット (Gated Recurrent Unit, GRU[34]) という。次項では、このゲートを用いたリカレントニューラルネットワークの一つである LSTM について紹介する。

2.4.3 長・短期記憶 (Long Short-Term Memory, LSTM)

LSTM のネットワーク構造全体を図 2.10 に示す。リカレントニューラルネットワークとの最も大きな違いは、LSTM は隠れ状態 (出力) を二つ h_t, c_t 持っているという点である。 c_t は長期的な記憶セルを示しており、図 2.10 では上部の赤い線で表現されている。一方、 h_t は、リカレントニューラルネットワークと同様に短期的な系列情報の伝達と出力に使用されている。LSTM は三つの入力 h_{t-1}, c_{t-1}, x_t を受け取り、二つの出力 h_t, c_t を提供するネットワークであるとみなすことができる。

また、LSTM は内部に前項で述べた勾配消失や勾配爆発、入力重み衝突、出力重み衝突といった様々な問題を解決するためのゲートと呼ばれる構造を四つ持っている (図 2.10)。

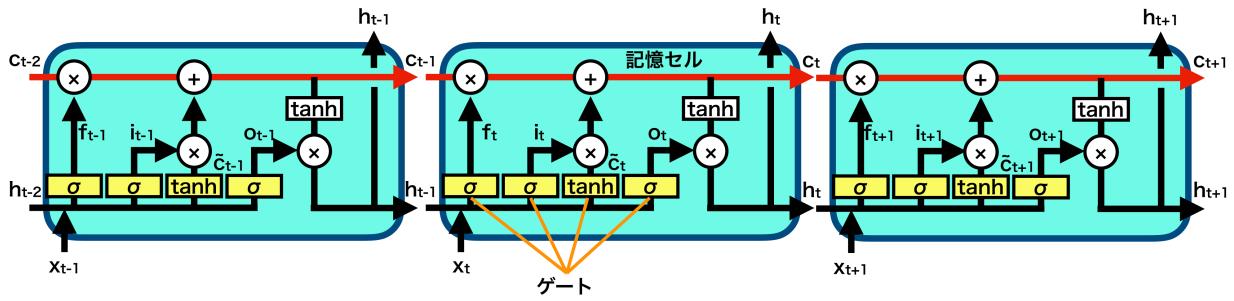


図 2.10: LSTM の流れ。リカレントニューラルネットワークの様に時間について展開した図である。赤線が記憶セル、黄色の各部がそれぞれゲートである。出力と隠れ状態である h, c は直前の系列から受け取り、入力 x は下方から導入されている。

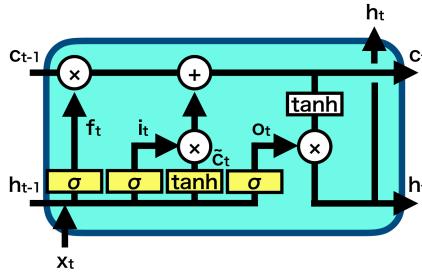


図 2.11: 単体の LSTM。系列 1 ステップ分の LSTM を切り出したものである。リカレントニューラルネットワークと同様に重み行列は各ステップに渡って同一のものが使用されている。

四つのゲートはそれぞれ忘却ゲート、入力ゲート、セルの更新、出力ゲートと呼ばれている。それぞれの役割と演算を以下に示す。

- 忘却ゲート

図 2.12a の赤い線で表現している領域では、入力 h_{t-1}, x_t を用いて、どの程度、直前の長期記憶セル c_{t-1} を忘れるかの度合いである f_t を計算している。これを忘却ゲートという。 f_t は、次のようにかける。

$$f_t = \sigma(W_f x_t + R_f h_{t-1}) \quad (2.27)$$

したがって最終的には、

$$c_{t-1} f_t = c_{t-1} \sigma(W_f x_t + R_f h_{t-1}) \quad (2.28)$$

として、次のゲート以降で長期記憶を更新するための容量を確保していると解釈できる。

- 入力ゲート

入力ゲートは忘却ゲートと全く同じ構造 (図 2.12b) をしており、次のように計算で

きる。

$$\mathbf{i}_t = \sigma(W_i \mathbf{x}_t + R_i \mathbf{h}_{t-1}) \quad (2.29)$$

忘却ゲートではどの程度、長期記憶セルを忘れるかを計算していたように、入力ゲートでの \mathbf{i}_t は、次のセルの更新時に新しい長期記憶セルをどの程度重視するかを表現していると解釈できる。

- セルの更新

図 2.12c では、まず更新された長期記憶セル $\tilde{\mathbf{c}}_t$ を計算している。

$$\tilde{\mathbf{c}}_t = \tanh(W_c \mathbf{x}_t + R_c \mathbf{h}_{t-1}) \quad (2.30)$$

次にこれまでの三つのゲートの結果をまとめることで、新しい隠れ状態である長期記憶セル \mathbf{c}_t を計算できる。

$$\begin{aligned} \mathbf{c}_t &= \mathbf{c}_{t-1} \mathbf{f}_t + \tilde{\mathbf{c}}_t \mathbf{i}_t \\ &= \mathbf{c}_{t-1} \sigma(W_f \mathbf{x}_t + R_f \mathbf{h}_{t-1}) + \tanh(W_c \mathbf{x}_t + R_c \mathbf{h}_{t-1}) \sigma(W_i \mathbf{x}_t + R_i \mathbf{h}_{t-1}) \end{aligned} \quad (2.31)$$

第一項では、直前の長期記憶セル \mathbf{c}_{t-1} をどの程度忘れるかを \mathbf{f}_t によって制御し、第二項では、新しく計算された長期記憶セル $\tilde{\mathbf{c}}_t$ をどの程度重視するかを \mathbf{i}_t によって制御している。

- 出力ゲート

最後に出力ゲートについて述べる。出力ゲートではここまでに計算された \mathbf{c}_t と入力 $\mathbf{x}_t, \mathbf{h}_{t-1}$ を用いて、最終的な出力となる \mathbf{h}_t を計算している。ただし、出力ゲート \mathbf{o}_t 自体は入力ゲートや忘却ゲートと全く同じ形(図 2.12d)をしている。具体的な計算は次のように書ける。

$$\begin{aligned} \mathbf{o}_t &= \sigma(W_o \mathbf{x}_t + R_o \mathbf{h}_{t-1}) \\ \mathbf{h}_t &= \tanh(\mathbf{c}_t) \mathbf{o}_t \\ &= \tanh(\mathbf{c}_t) \sigma(W_o \mathbf{x}_t + R_o \mathbf{h}_{t-1}) \\ &= \tanh(\mathbf{c}_{t-1} \sigma(W_f \mathbf{x}_t + R_f \mathbf{h}_{t-1}) + \tanh(W_c \mathbf{x}_t + R_c \mathbf{h}_{t-1}) \sigma(W_i \mathbf{x}_t + R_i \mathbf{h}_{t-1})) \\ &\quad \sigma(W_o \mathbf{x}_t + R_o \mathbf{h}_{t-1}) \end{aligned} \quad (2.32)$$

以上のように、LSTM の内部構造は煩雑であるが、その構成要素は全てリカレントニューラルネットワークと同様に全結合で計算できる。以下に最終的な出力の計算についてまとめる。

$$\begin{aligned} \mathbf{c}_t &= \mathbf{c}_{t-1} \sigma(W_f \mathbf{x}_t + R_f \mathbf{h}_{t-1}) + \tanh(W_c \mathbf{x}_t + R_c \mathbf{h}_{t-1}) \sigma(W_i \mathbf{x}_t + R_i \mathbf{h}_{t-1}) \\ \mathbf{h}_t &= \tanh(\mathbf{c}_t \sigma(W_f \mathbf{x}_t + R_f \mathbf{h}_{t-1}) + \tanh(W_c \mathbf{x}_t + R_c \mathbf{h}_{t-1}) \sigma(W_i \mathbf{x}_t + R_i \mathbf{h}_{t-1})) \\ &\quad \sigma(W_o \mathbf{x}_t + R_o \mathbf{h}_{t-1}) \end{aligned} \quad (2.33)$$

ここで、隠れ状態(出力) $\mathbf{h}_t, \mathbf{c}_t$ の計算は全て入力 $\mathbf{h}_{t-1}, \mathbf{c}_{t-1}, \mathbf{x}_t$ によって計算できている。また、学習可能な重み行列は $W_f, W_i, W_c, W_o, R_f, R_i, R_c, R_o$ の八つである。一般に、適宜バイアス b が加えられる。

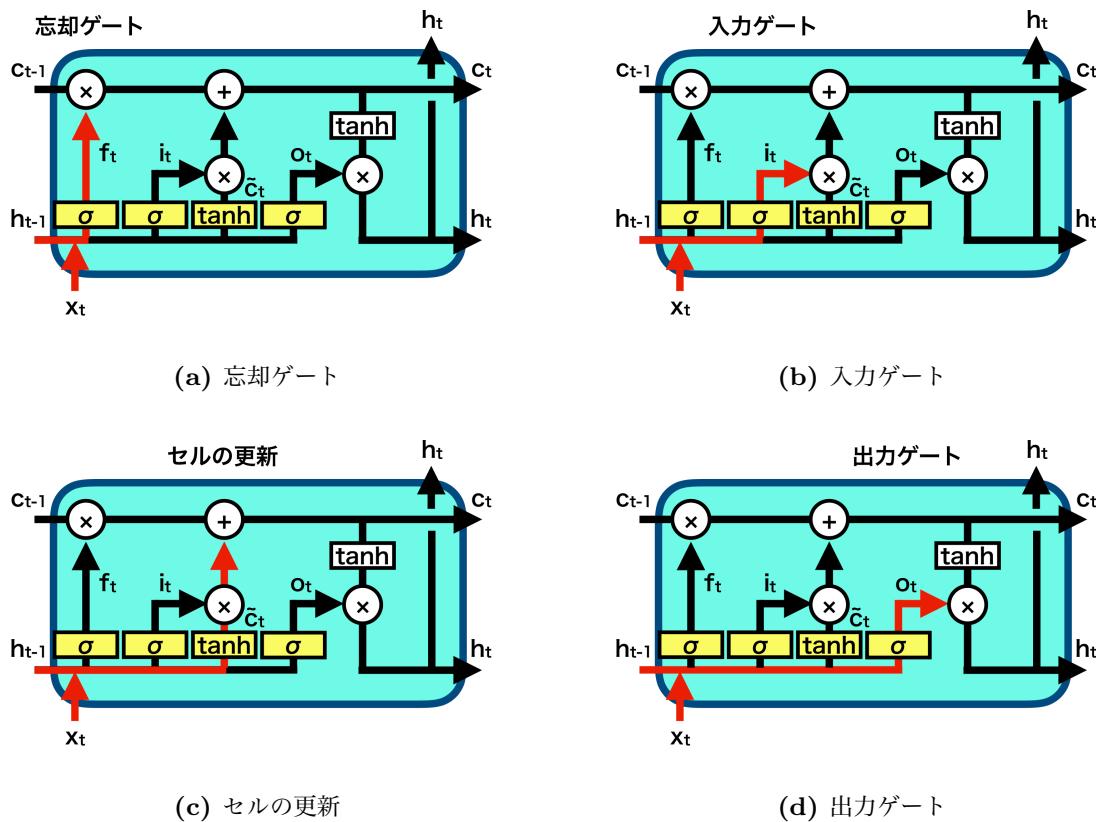


図 2.12: LSTM の各ゲートについての図解。それぞれ赤線部に沿って情報が伝達される。

リカレントニューラルネットワークや LSTM は更に重ねられる (Stacked) という非常に強力な性質を持っている。そのようなネットワークを図 2.13 に示す。図からわかるように、一段目の LSTM の出力が二段目の LSTM の入力になっている。一方で、セルは両者で共有されず、独立した状態を保持している。それら以外の基本的な構造は一段であった時の LSTM と変わっていない。このように、LSTM は系列としての深さだけでなく、フィードフォワードニューラルネットワークと同様に重ねることによる深さの確保が可能である。勿論この重ねる操作は二段以上への拡張が可能であり、その場合は二段目の出力を三段目の入力に使うことによって実現できる。どの程度重ねるかはハイパーパラメータである。

更に、双方向 (Bidirectional) LSTM と呼ばれるネットワークについても解説する。これは、図 2.14 のように、片側を順方向に、もう一方を逆方向に系列を読み込むことにより、自然言語処理などに見られる、将来的な系列情報への依存を導入することができる。ここで、前述の LSTM を重ねる手法と異なり、それぞれの LSTM の入力はそれぞれ独立しており、前段の出力を使用していないことに注意が必要である。また、この双方向 LSTM の構造上の性質として、系列データを全て持つ必要があるという点にも留意しなければならない。したがって、リアルタイムな問題については、未来の情報を得ることができないため、この双方向 LSTM を用いることはできない。また、この双方向 LSTM を重ねることも可能である。

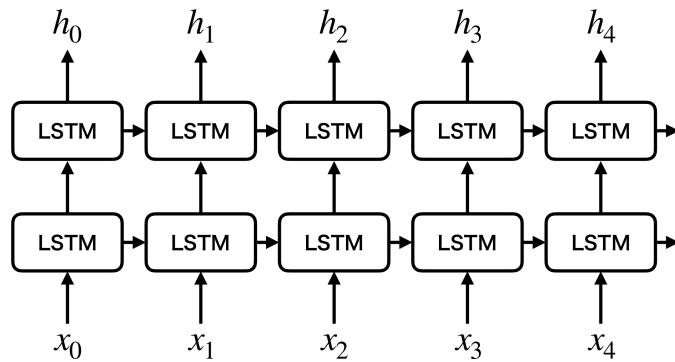


図 2.13: Stacked LSTM。ここでは LSTM の二つの隠れ状態を一つの線で表現している。下方を一段目、上方を二段目と呼ぶと、一段目の入力 x と比べ、一段目の出力・二段目の入力はより抽象度の高い情報となっている。

リカレントニューラルネットワークはこのように次々と拡張され、より複雑で難解な系列情報の処理について、高い性能を発揮している。

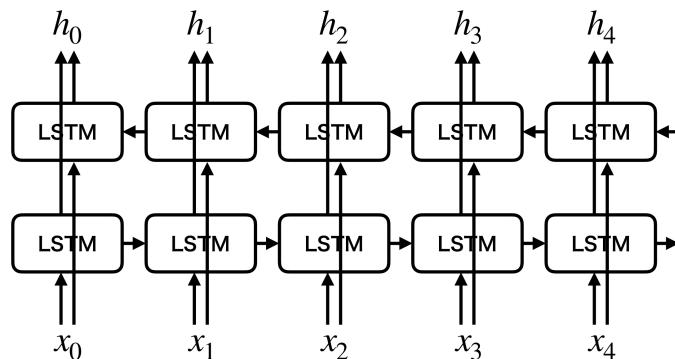


図 2.14: 双方向 LSTM。Stacked LSTM とは異なり、一段目の LSTM の出力は二段目の入力に使用されておらず、計算はそれぞれ並列的に行われる。よって出力は LSTM 二つ分の大きさになっている。

リカレントニューラルネットワークはその構造が再帰的であるという点（並列化が困難である）から、学習が遅く、重いという課題を抱えている。また、後述するエンコーダー・デコーダーモデルにおいては、データの系列の長さに応じた情報を確保できないという欠点が存在している。^{*6}次節では、このような問題を解決するための注意機構（Attention）という技術を解説する。

^{*6} 例えば、機械翻訳を用いて 100 単語分の英文を日本語に翻訳する場合と 10 単語分の英文を翻訳する場合において、100 単語分の英文が持つ情報の方が 10 単語分の英文と比べて多いことは明らかであるが、リカレントニューラルネットワークはこれらの情報の多さの違いに対応できず、常に同じ量の情報から日本語を生成してしまう。

2.5 Attention

Attention[35, 36] はその名の通り、ある系列データのどこに注意するかを計算する機構である。主に機械翻訳や意味理解などのエンコーダー・デコーダーモデルに使用されており、様々な応用が議論されているが、ここでは基本的な Attention の理論と考え方についてのみ解説する。

2.5.1 項では、まず Attention を理解する上で必要不可欠なエンコーダー・デコーダーモデルについて述べ、その中で前節のリカレントニューラルネットワークが抱える問題について紹介する。次に、2.5.2 項で Attention の理論や計算について述べる。

2.5.1 エンコーダー・デコーダーモデル

Attention は主に機械翻訳などのエンコーダー・デコーダーモデルに使用されている。LSTM を用いたエンコーダー・デコーダーモデルの大まかな構成を図 2.15 に示す。LSTM ではエンコーダーとデコーダーを繋ぐ情報はエンコーダーの最後の層の出力を使用することが多い。つまり、エンコーダーは図 2.9 の Many to One、デコーダーは One to Many を使用しており、エンコーダーによって抽出された情報はこの One の部分に集められることになる。この出力は Many to One の内、Many であるデータの系列の長さ（系列長）に依存せず、常に同じ大きさとなる。これは長い系列長であっても、短い系列長であっても同じ量の情報を使用しているということを意味している。したがって、より長い系列長や短い系列長を扱う場合は、そのデータが持つ情報を完全に保持、表現することは難しく、情報の欠損により性能が下がってしまうという問題があった。

この問題を解決するための技術が Attention である。Attention を組み込んだ LSTM のエンコーダー・デコーダーモデルを図 2.16 に示す。ここで、図中の Attention と示している部分は、実際にはエンコーダーの全ての系列の出力を集めた行列である。このようにして、Attention は系列長に依存した情報量を確保できる。次項ではより詳細な Attention の計算について説明する。

2.5.2 Attention

図 2.16 では、エンコーダーとデコーダーの間に Attention を表現しているが、実際には Attention はデコーダーの個々の系列に応じて計算される。これは、デコーダーのある系列 t がエンコーダー全体のどの情報に注目しているかを逐次計算しなくてはならない為である。Attention の実装の仕方は多岐にわたるが、その計算は次のような手順にまとめられる。

1. Query、Key、Value を用意する。
2. Query と Key を用いて、Attention Weight を計算する。

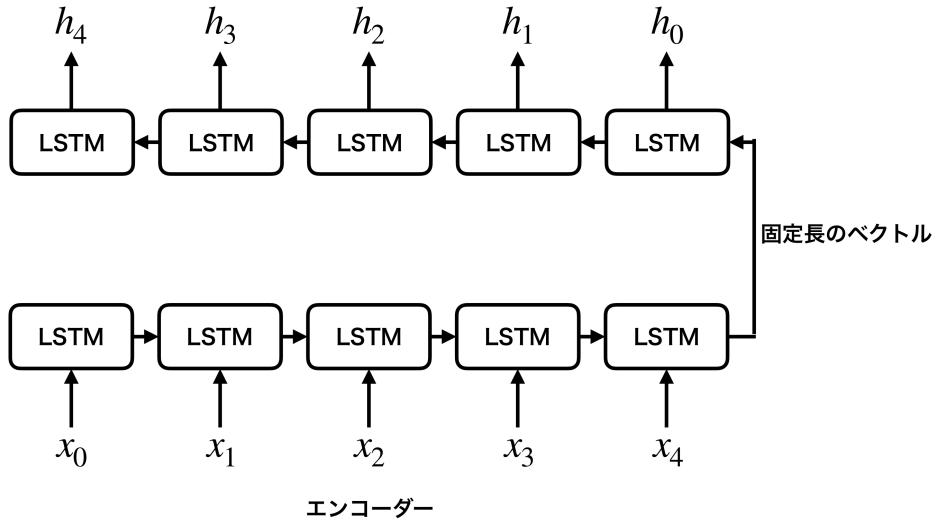


図 2.15: LSTM によるエンコーダー・デコーダーモデル。図下部がエンコーダー部、図上部がデコーダー部である。エンコーダー部の入力 x から抽出された情報は図右の固定長のベクトルへと集約される。デコーダー部ではこの固定長のベクトルを初期状態にして出力 h の生成を行う。

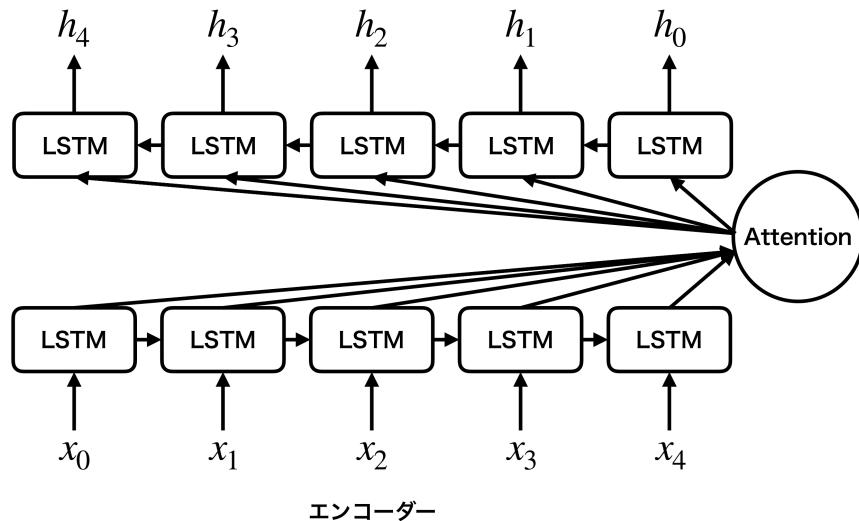


図 2.16: Attention と LSTM によるエンコーダー・デコーダーモデル。図下部がエンコーダー部、図上部がデコーダー部である。エンコーダー部で抽出された情報は図右の Attention に集約される。これは後述する Key や Value に相当する。Attention の大きさはエンコーダー部の入力系列長に依存しており、常に適切な情報量を確保できる。

3. Attention Weight と Value を用いて、コンテキストを作成する。

Query、Key、Valueにおいて、Key と Value は同じものが用いられることが多い。Key と Value は図 2.16 のエンコーダーの全ての系列の出力を集めた行列に相当している。Query とはここでは、デコーダーの個々の系列である $t - 1$ 番目の出力 (隠れ状態) である。したがって、Query と Key を用いて、Attention Weight を計算するという操作は $t - 1$ 番目の隠れ状

態 (Query) とエンコーダーの全ての系列の出力を集めた行列 (Key) を用いて、どこに注意するかを表現した重み (Attention Weight) を計算していることを意味している。次に、この Attention Weight を、もう一度エンコーダーの全ての系列の出力を集めた行列 (Value) に掛けることで、エンコーダーのそれぞれの系列の情報を注意しながら取り出すことができる。この取り出された情報をコンテキストといい、Attention ではこのコンテキストをデコーダー部に使用することで、系列長に依存した情報量を保持している。

Attention Weight の計算方法はいくつかの手法が存在しているが、ここでは最も単純な手法を二つ述べる。

- Additive Attention[35]

Additive Attention の利点は Query と Key がどのような大きさであっても計算できるという点である。N 番目の Query と Key がそれぞれ、大きさ $[F]$ と $[E, M]$ のベクトルと行列であるとすると、その計算は図 2.17 の上段のように表現できる。ここで N 番目の Query は系列長の M 回分反復されており、最終的に大きさ $[F, M]$ の行列となっている。Additive Attention の特徴は Attention の内部に独自の学習可能な重みを保持している点である。N 番目の Query 行列と Key はそれぞれ重み行列と掛けられ、得られたそれぞれの行列要素について和を取っている。その後、更に大きさ $[D]$ の重みベクトルと掛けられ、大きさ $[M]$ のエネルギー (図 2.17 中のベクトル e_N) を得る。このベクトルについて、ソフトマックスと同様の関数で規格化し、Attention Weight を作成している。この計算はフィードフォワードニューラルネットワークそのものであり、Additive Attention はこのように内部にネットワークを持っているため、後述する Dot-Product Attention と比較して計算が遅くなるという欠点を抱えている。

- Dot-Product Attention[36]

Dot-Product Attention は Additive Attention と比較して非常にシンプルな構造をしている。その計算を図 2.17 の下段に示している。図からも分かるように Additive Attention とは異なり、内部に重みを保持しておらず、Query と Key から直接 Attention Weight を計算している。ただし、このように Query と Key を掛けるためには、それぞれの大きさを揃える必要があり、具体的には $E = F$ の時のみこの Dot-Product Attention を使用することができる。

コンテキストをどのように以降の計算へ組み込むのかについての具体的な例は、本研究の?? 項に示している。上記のように計算された Attention Weight を確認することでネットワークの理解をより深めることができる。この Attention Weight はあるデコーダー部の 1 系列がエンコーダー部の全系列に対して何処に注意しているかを表現している事に再度注意しておく。したがって、デコーダー部の全系列に対して Attention Weight を計算した場合、それは行がデコーダー部の系列長、列がエンコーダー部の系列長の行列となる。Attention Weight に関する具体的な図についても本研究の 3.4.3 項で確認する。

Attention は”Attention is all you need[37]”と言われるほど、近年非常に注目されている技

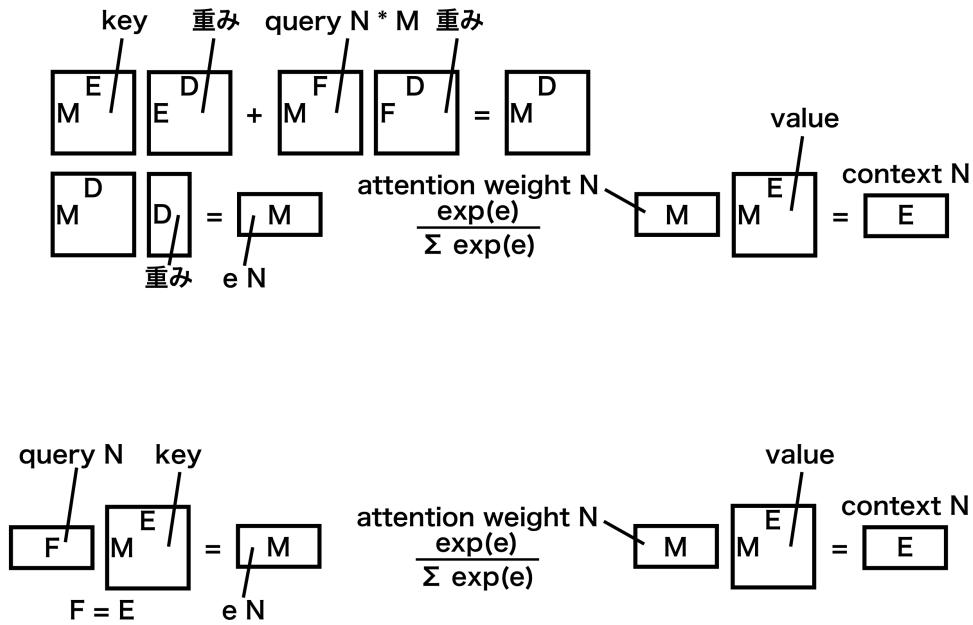


図 2.17: Additive Attention と Dot-Product Attention

術である。ここでは、Attention を LSTM の補助として使用している例を挙げたが、Attention は様々な技術と組み合わせられる。またそれだけでなく、Attention のみを用いて構成された、Transformer と呼ばれるモデルは 2021 年現在において、自然言語処理の標準的なネットワークであると言われるほどの性能と、LSTM では実現できなかった速さを実現している。

2.6 ハイパーパラメータ

ここまで述べたように深層学習は教師あり学習であるため、訓練データを用いて重みの更新を行い、ネットワークの重み（パラメータ）を調整（チューン）していく。しかし、ネットワークはネットワーク自体を構築するため、学習によって更新されないパラメータをいくつも持っている。このようなパラメータをハイパーパラメータという。ハイパーパラメータは学習前に設定しておく必要があり、ネットワークの性能を引き出すためには適切に最適化（ハイパーパラメータ・チューニング）する必要がある。ハイパーパラメータの種類や数はネットワークの構造によって大きく異なるが、一般にチューニングが必要とされるハイパーパラメータについて以下に示す。

- 最適化手法 (Optimizer) : RMSProp や Adam といった重み更新の為の最適化手法
- 学習率 (Learning rate) : 重み更新のステップ幅
- エポック数 (Epochs) : 学習回数・訓練データを一周学習することを 1 エポックという
- バッチサイズ (Batch size) : ミニバッチ学習における訓練データサンプルの大きさ
- ノード数 (Node) : 重み行列の大きさ

- 層の数 (Layer) : フィードフォワードネットワークにおける全結合層の数

これらはハイパーパラメータの一例であり、それについて適切に選ぶ必要がある。ハイパーパラメータの探索手法も幾つか提案されており、ランダムサーチやグリッドサーチ、ベイズ最適化などが用いられている。本研究におけるハイパーパラメータ・チューニングに関しては 3.4.3 項で述べる。

以上が本論文のための深層学習の導入である。この 2 章を前提として、3 章での本研究で使用したネットワークの構造を解説していく。以降の章では、ここで挙げた深層学習の用語を説明せずに使用するが、その場合は適宜、本章を参照していただきたい。また、本章の初めや 2.3 でも述べたように、深層学習の実装に関しては様々なフレームワークがあるため、ここでの記載は省かせていただく、本研究の実装に関しては付録 A や私の GitHub[38] にまとめている。

第3章

崩壊点検出の為のネットワーク

本章では、深層学習を用いた崩壊点検出のために構築したネットワークの詳細について述べる。

まず 3.1 節では、本研究で取り扱うデータの特性について述べる。1.5.1 項で言及したように本研究で使用するデータは全て MC シミュレーションデータである。3.1 節ではそのようなシミュレーションデータについてソフトウェアの性質や物理的な性質について解説する。ただし、ネットワークの学習に使用した訓練データについては個々のネットワークの解説で適宜述べる。

次に、3.2 節では、深層学習を使用して、どのように崩壊点検出を実現するかについて、発想と構築したネットワークの役割について概説する。また、使用したハードウェアやソフトウェア・フレームワークの環境などについてもここで述べる。

3.3 節と 3.4 節では、本研究のために構築した個々のネットワークについて、構造・学習・性能と評価について議論する。

3.1 データ

本研究で使用したシミュレーションデータについて 3.1.1 項では、データ全体についてのソフトウェアや物理的性質に関して述べる。3.1.2 項では、本研究で使用する飛跡についての情報と深層学習に使用するための前処理について紹介する。

3.1.1 データ全体の性質

本研究ではシミュレーションデータとして、イベントジェネレーター WHIZARD[39] を用いて生成された ILD フルディテクターシミュレーションデータを使用した。これらのデータは後述する LCFIPlus での性能評価 [40] で使用されたデータと同一のものである。MC シミュレーションデータの性質を表 3.1 に示す。

ここで、終状態の b はボトム・フレーバー、 c はチャーム・フレーバー、 q はアップ、ダウン、ストレンジ・フレーバーのクォークをそれぞれ表している。これらのクォーク対は自然界で直

イベントジェネレーター	WHIZARD
検出器	ILD フルディテクターシミュレーション
重心系エネルギー	Z 粒子の質量 (91.2 GeV)
終状態	$e^+ e^- \rightarrow b\bar{b}, c\bar{c}, q\bar{q}$
Beamstrahlung/ISR	なし
ビーム偏極	なし

表 3.1: MC シミュレーションデータの性質

ちにハドロンを形成し、特にボトム、チャーム・フレーバーの場合は再構成可能な Secondary Vertex を残すジェットとなる。

データについて、終状態 $b\bar{b}$ のデータは 15 個のサンプルに、終状態 $c\bar{c}$ のデータは 13 個のサンプルに分け使用した。終状態 $b\bar{b}$ のデータは 12 個のサンプルに分離したが、深層学習ネットワークの学習には使用していない。深層学習を含めた教師あり学習では、健全性のため学習に使用する訓練データと学習時の性能観測に使用する検証データ、最終的な評価に使用するテストデータは分けなければならない。したがって個々のサンプル毎に用途を明らかにし、訓練データの作成とネットワークの評価に使用するデータの詳細について、サンプルに含まれる事象数や飛跡数、用途を表 3.2 にまとめる。

後述する??章の現行の手法である LCFIPlus との比較において LCFIPlus のフレーバータギング (BDTs) の学習が必要である。その為、BDTs の訓練データ作成に $c\bar{c} - 01, 02, 09, 10, 11, b\bar{b} - 01, 02, 03, 10, 11, 12$ を用いた。また、 $c\bar{c} - 01, b\bar{b} - 01$ は本節のデータ特性の調査に使用し、 $c\bar{c} - 02, b\bar{b} - 02, 03$ はネットワークの動作確認にも使用した。ネットワークの訓練データの作成は $c\bar{c} - 03, 04, 05, b\bar{b} - 04, 05, 06$ を用いて行なった。また、訓練データの正解ラベルは MC 情報や LCFIPlus のフィッターからの出力情報を使用した。

崩壊点検出を行うに当たって、終状態による崩壊点の性質の違いに注意しなければならない。終状態が $c\bar{c}$ の場合はチャーム・フレーバーのハドロンによる Secondary Vertex のみが生じる一方で、終状態が $b\bar{b}$ の場合は $b \rightarrow c$ という崩壊過程を辿り、ボトム・フレーバーのハドロンによる Secondary Vertex と更にそこから派生したチャーム・フレーバーのハドロンによる Tertiary Vertex が生じる。それぞれのハドロン粒子の典型的な飛程は寿命 τ と光速 c を用いて、ボトム・フレーバーの場合は $c\tau = 400 - 500 \mu m$ 、チャーム・フレーバーの場合は $c\tau = 20 - 300 \mu m$ となる (図 3.1)。

また、これら以外の崩壊点としてタウ粒子の崩壊や V^0 粒子の崩壊 ($K_S^0 \rightarrow \pi^+\pi^-$, $\Lambda^0 \rightarrow p\pi^-$)、光子変換 ($\gamma_{\text{conv}}X \rightarrow e^+e^-X$) によるものを考えることができる。これらの崩壊点は Secondary Vertex や Tertiary Vertex と比較して、衝突点から遠い位置で生じる。そのような崩壊点やその崩壊点由来の飛跡を以後 Others と呼ぶ。図 3.2 は一つの事象に含まれる飛跡の本数と崩壊点の個数である。これらの粒子識別は MC 情報を用いた。ここでは飛跡の本数

データ名	事象数	飛跡数	用途
c̄c - 01	69581	1344465	データの調査/フレーバータギングの訓練データの作成
c̄c - 02	42204	814074	動作テスト/フレーバータギングの訓練データの作成
c̄c - 03	38662	748027	飛跡対についてのネットワークの訓練データの作成
c̄c - 04	38712	747625	飛跡対についてのネットワークの訓練データの作成
c̄c - 05	38655	748089	任意の数についてのネットワークの訓練データの作成
c̄c - 06	38645	747548	ネットワーク/崩壊点検出の評価
c̄c - 07	38643	747312	ネットワーク/崩壊点検出の評価
c̄c - 08	38715	748801	ネットワーク/崩壊点検出の評価
c̄c - 09	38705	747725	フレーバータギングの訓練データの作成
c̄c - 10	38721	748025	フレーバータギングの訓練データの作成
c̄c - 11	38587	747819	フレーバータギングの訓練データの作成
c̄c - 12	38723	748904	LCFIPlusとの比較
c̄c - 13	35848	693780	LCFIPlusとの比較
<hr/>			
b̄b - 01	62795	1326168	データの調査/フレーバータギングの訓練データの作成
b̄b - 02	42950	909082	動作テスト/フレーバータギングの訓練データの作成
b̄b - 03	34985	738105	動作テスト/フレーバータギングの訓練データの作成
b̄b - 04	34952	739130	飛跡対についてのネットワークの訓練データの作成
b̄b - 05	35047	741568	飛跡対についてのネットワークの訓練データの作成
b̄b - 06	35008	740662	任意の数についてのネットワークの訓練データの作成
b̄b - 07	34000	718057	ネットワーク/崩壊点検出の評価
b̄b - 08	33978	717972	ネットワーク/崩壊点検出の評価
b̄b - 09	35008	740268	ネットワーク/崩壊点検出の評価
b̄b - 10	34954	739320	フレーバータギングの訓練データの作成
b̄b - 11	35012	740797	フレーバータギングの訓練データの作成
b̄b - 12	34972	739953	フレーバータギングの訓練データの作成
b̄b - 13	34986	739402	LCFIPlusとの比較
b̄b - 14	34910	740933	LCFIPlusとの比較
b̄b - 15	10243	216499	LCFIPlusとの比較

表 3.2: データサンプルの事象数と用途

は親粒子のフレーバーのみ考慮し、同フレーバーの親粒子の違いは区別していない。したがって Secondary Vertex については、一つの崩壊点ではなく b と \bar{b} などの複数の崩壊点の飛跡が含まれている。親粒子の違いを区別しない場合の Secondary Vertex の飛跡の数は 5 本程度である。本研究で使用したデータでは、典型的な一事象に含まれる同フレーバーの崩壊点の数は 2 つである。したがって個々の Secondary Vertex に含まれる飛跡の数は 2 – 3 本程度となる。崩壊点検出では Primary Vertex と個々の Secondary Vertex を見分ける必要がある。

崩壊点の個数では親粒子を全て区別している。典型的な崩壊点の個数は終状態が $c\bar{c}$ の場合は Primary Vertex、チャーム・フレーバーの Secondary Vertex が 2 つ、Others の 3 – 5 個である。終状態が $b\bar{b}$ の場合は Primary Vertex、ボトム・フレーバーの Secondary Vertex が 2 つ、チャーム・フレーバーの Tertiary Vertex が 2 つ、Others の 5 – 7 個である。

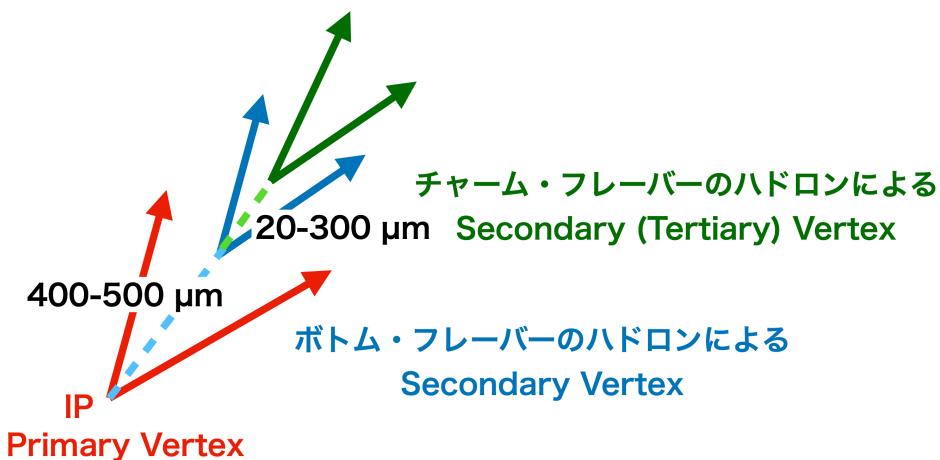


図 3.1: 終状態 $b\bar{b}$ での崩壊点の例。赤線: Primary Vertex 由来の飛跡。青線: Secondary Vertex 由来の飛跡。緑線: Tertiary Vertex 由来の飛跡。

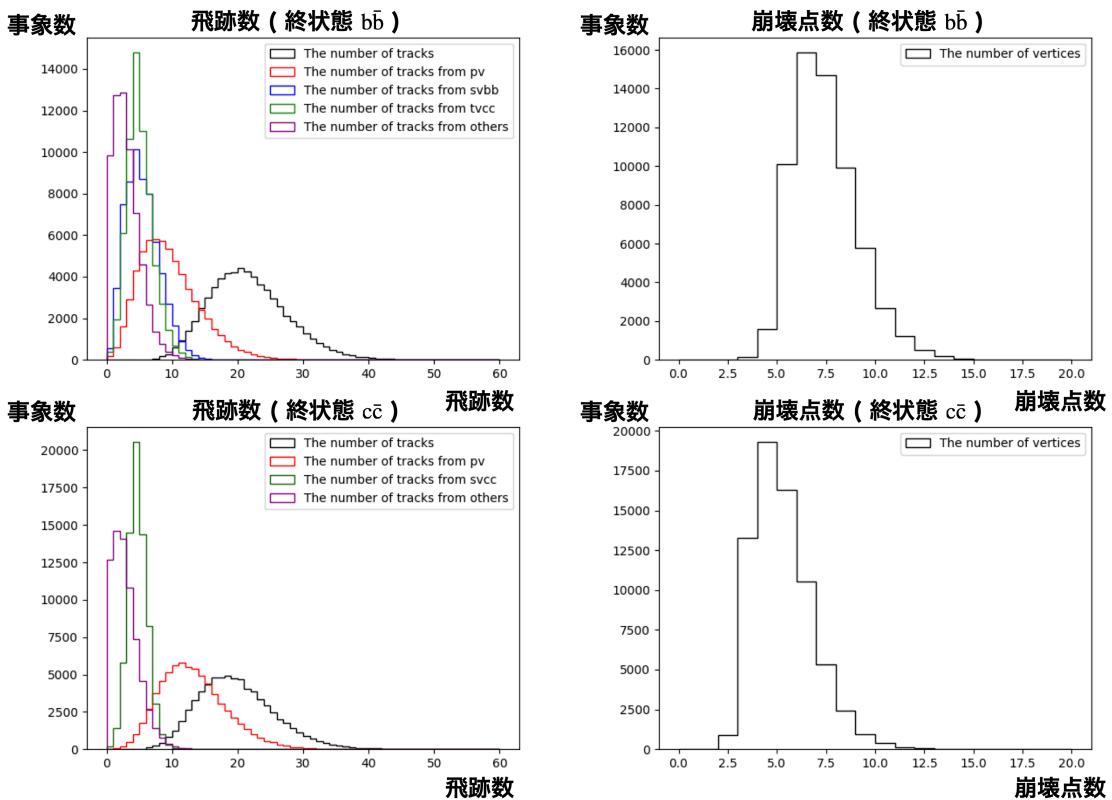


図 3.2: 事象に含まれる飛跡の本数と崩壊点の個数。図左は飛跡の本数、図右は崩壊点の個数である。飛跡の本数において、黒線は事象に含まれる全飛跡数、赤線は Primary Vertex 由来の飛跡数、青線はボトム・フレーバーの Secondary Vertex 由来の飛跡数、緑線はチャーム・フレーバーの Secondary Vertex 由来の飛跡数、紫線は Others の飛跡数を表している。

3.1.2 飛跡の情報と前処理

飛跡一本分の情報として、図 3.3 のような位置や運動量の情報を含んだトラック・パラメータ 5 個 ($d_0, z_0, \phi, \Omega, \tan \lambda$)[1] とその共分散行列 15 値、電荷、エネルギーの計 22 値の変数を使用した。また、加速器の座標系としてビーム衝突点を原点とし、Z 軸をビーム方向にとった座標系を使用する。

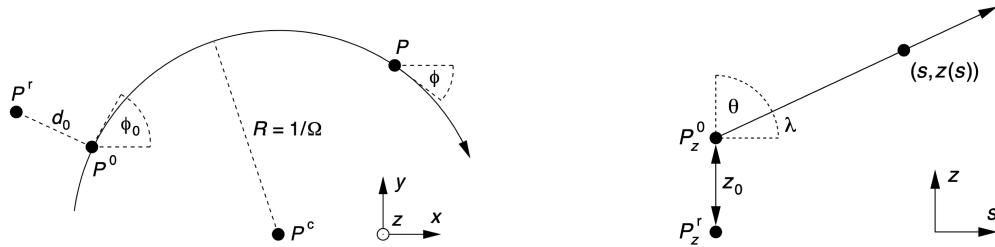


Figure 1: The projection of a helix segment in the xy plane is a part of an arc with centre \mathbf{P}^c and radius R . The direction of the particle is shown with the arrow at the arc. All track parameters are given relative to the reference point \mathbf{P}^r .

Figure 2: The projection of a helix in the sz plane is a straight line (see Eq. 10). The variable s at a point \mathbf{P} is the arc length in the xy plane from \mathbf{P}^0 to \mathbf{P} . This also implies that $s = 0$, if $z = z_0$.

図 3.3: トラック・パラメータの定義 [1]

深層学習の入力に使用する変数は、一般に $[-1, 1]$ の範囲に整形した方が良いと言われている為、変数をそれぞれ以下のような \tanh 関数や線形関数などを用いて整形した。

- トラック・パラメータ

$$d_0 = \tanh(d_0), z_0 = \tanh(z_0), \phi = \phi/\pi, \Omega = \tanh(200 \Omega), \tan \lambda = \tanh(0.3 \tan \lambda)$$

- トラック・パラメータの共分散行列

$$\tanh(8000(x - 0.0005))$$

- エネルギー

$$\tanh(0.5(x - 5.0))$$

トラック・パラメータとエネルギーの整形前と整形後の分布をそれぞれ図 3.4 に示す。整形後の変数の分布が $[-1, 1]$ の範囲になっていることがわかる。

また、LCFIPlus のフィッティングで得られる χ^2 や崩壊点の位置についてもデータを用意した。ここでは、事象中の二本の飛跡（飛跡対）の全ての組み合わせについて計算を行った。ただし、このような高次の情報を含んだ変数は基本的にはネットワークの学習に使用せず、正解ラベルの一つとして使用した。LCFIPlus によって予想される崩壊点の位置について、ビーム衝突点からの距離を図 3.5 に示す。図 3.5 のグラフは両対数グラフとして表現している為、衝突点からの距離 10^{-2} mm から 10^3 mm のプロットとなっており、衝突点は含んでいない。Primary Vertex と Secondary Vertex の衝突点からの距離は大きく異なっているが、各フレー

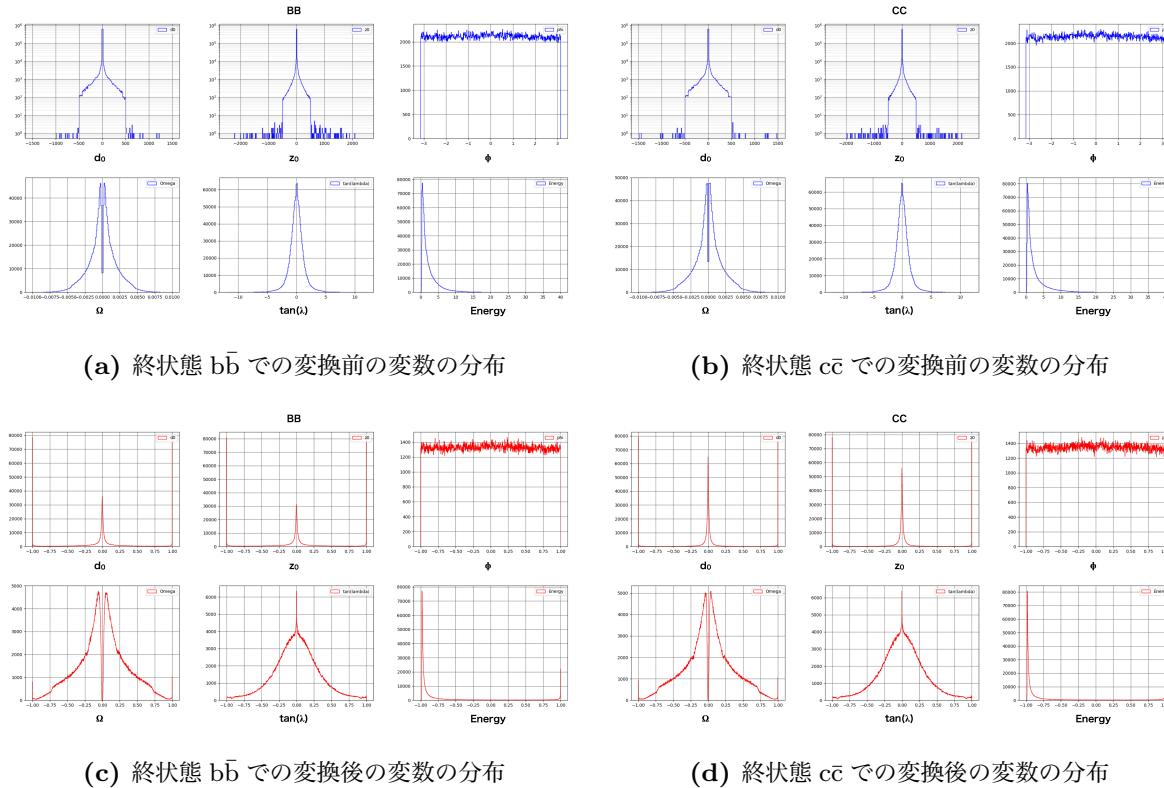


図 3.4: 変数の分布の例

バーの Secondary Vertex の衝突点からの距離は殆ど離れていないことが分かる。また、終状態 $c\bar{c}$ と終状態 $b\bar{b}$ では $b \rightarrow c$ の崩壊過程を辿るか否かの違いがある為、チャーム・フレーバーの崩壊点の位置が少し異なっている。Others は衝突点付近では殆ど起こらず、他の崩壊点と比較して遠い位置で崩壊している。このように崩壊点検出において位置の再構成は非常に重要な情報を含んでおり、逆説的に位置の再構成が出来なければ崩壊点に分離は困難であると言える。

1 mm 付近のピークは LCFIPlus のフィッティングが失敗している飛跡対である。フィッティング健全性を表す χ^2 との相関を見ると、図 3.6 のように、1 mm 付近のデータは大きな χ^2 値を持っていることが分かる。

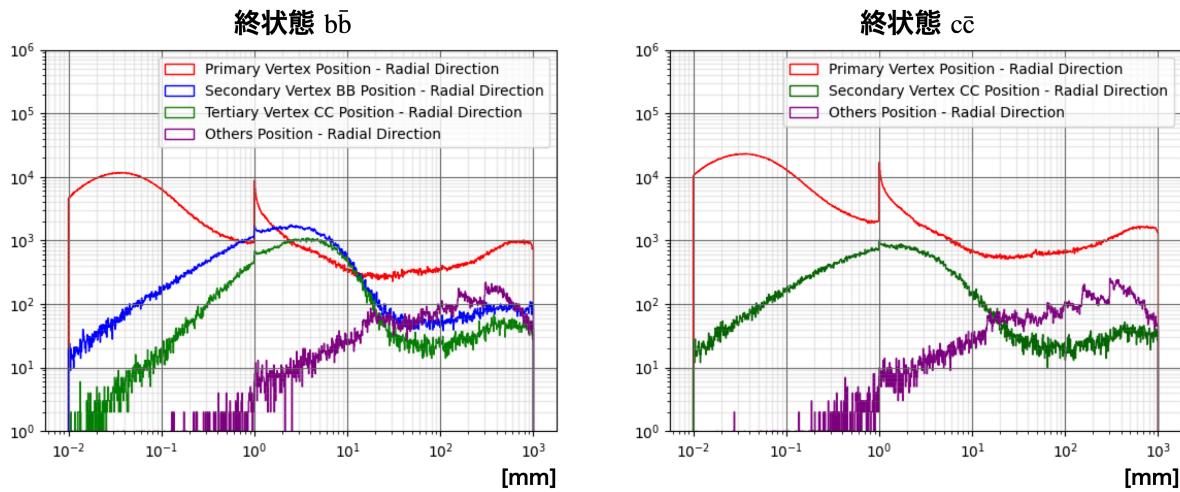


図 3.5: LCFIPlus によって予想される崩壊点の位置の分布。横軸はログスケールの衝突点からの距離である。赤線は Primary Vertex の位置、青線はボトム・フレーバーの Secondary Vertex の位置、緑線はチャーム・フレーバーの Secondary Vertex の位置、紫線は Others の位置を表している。これらは飛跡対についての LCFIPlus の計算値である。

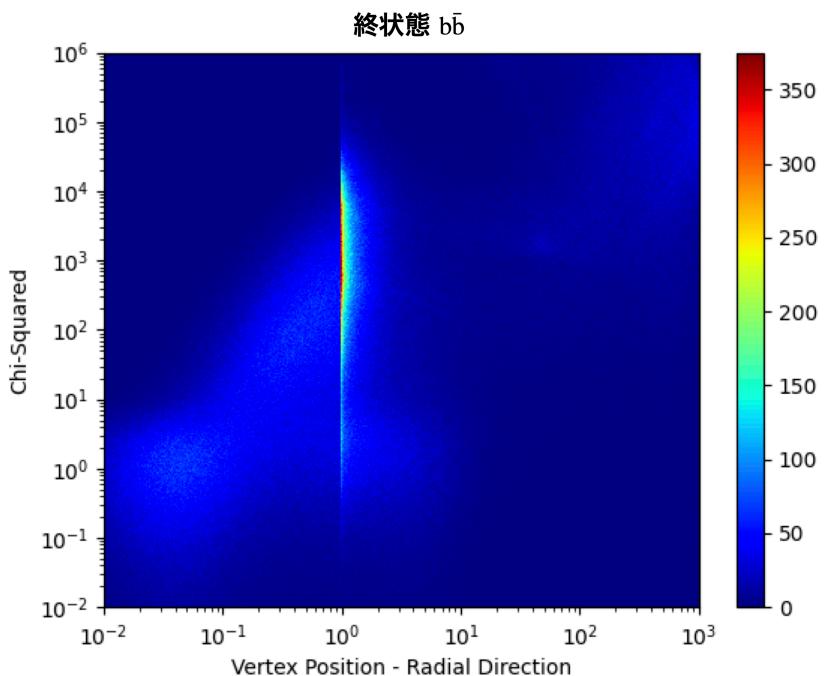


図 3.6: LCFIPlus によって予想される崩壊点の位置と χ^2 値の相関。縦軸、横軸はそれぞれログスケールの χ^2 値、ログスケールの衝突点からの距離である。カラースケールはカウントを示しており、1 mm 付近で非常に大きな値になっていることがわかる。

3.2 深層学習を用いた崩壊点検出の実現

深層学習は分類問題や回帰問題を解けるが、一方で基本的にはクラスタリングなど教師なし学習に対しては不向きである。分類問題や回帰問題では訓練データから「パターン」を学び、データの持つ特徴量の空間内である種の境界を引く必要がある。この「パターン」はあらゆるデータ内で分類可能な決まった性質を持っていなければならない。例えば 3.1.1 項では、同一フレーバーの Secondary Vertex が主に二つあることを解説した。この二つの Secondary Vertex は事象内では位置の違いによって区別できるが、あらゆる事象間で不变的に一番や二番といったラベル付けできる性質を持っていない。^{*1}

崩壊点検出アルゴリズムの目的は同一フレーバー内の複数の Secondary Vertex を含む崩壊点を探索する事である。このような問題は一般にクラスタリングを用いて解くことが多いが、本研究で使用するデータは図 3.2 で示したように事象内に含まれる飛跡の本数や崩壊点の個数も異なっているという性質を持っている。これはクラスター数やクラスターに含まれる要素数が常に変わってしまうということを意味しており、崩壊点検出をクラスタリングで行うことは不適であると判断した。

以上を踏まえた上で私は次の二つのネットワークを用いた崩壊点検出を提案する。

1. 飛跡対についてのネットワーク

- 用途：崩壊点のタネの探索
- 入力：事象内のあらゆる飛跡対（二本の飛跡の全ての組み合わせ）
- 出力：飛跡対の属する崩壊点の種類・崩壊点の位置

2. 任意の数の飛跡についてのネットワーク

- 用途：崩壊点の生成
- 入力：崩壊点のタネ・事象内の全ての飛跡
- 出力：事象内のそれぞれの飛跡が崩壊点のタネに結合しているか否か

飛跡対についてのネットワークは崩壊点のタネとなる飛跡対を探索するネットワークである。したがって、このネットワーク単体では崩壊点を形成することはできない。そこで、私は崩壊点の生成を行う、もう一つのネットワークを構築した。崩壊点の生成を行うネットワークは二本以上の飛跡を取り扱う必要があるが、三本、四本、五本の飛跡を取り扱えるネットワークをそれぞれ構築することはネットワークや組み合わせの数を考える上で適切ではない。よって不定の数の飛跡を再帰的に処理するネットワーク構造として、リカレントニューラルネットワークを使用した。任意の数の飛跡についてのネットワークは崩壊点のタネをリカレントニューラルネットワークの初期状態として、そこに飛跡を一本ずつ加え崩壊点の生成を行うネットワークである。本研究は以上の二つのネットワークを用いることで崩壊点検出を実現し

^{*1} 実際には損失関数を最小にするような順序を与えることで分離することは可能であるが本研究においては様々な Secondary Vertex が存在する為不適である。

た。構造や学習についての、より詳細な個々のネットワークの解説は、後の 3.3 節や 3.4 節で述べる。

本研究のネットワークは Tensorflow/Keras フレームワークを用いて構築・学習を行なった。また、学習に際しては計算機として、弊研究室サーバーの”NVIDIA TITAN RTX”や九州大学情報基盤研究開発センター研究用計算機システムの一般利用を使用した。詳細なソフトウェア・ハードウェアの環境を表 3.3 に示す。

ソフトウェア	
Python	3.6.8
Tensorflow	2.1.0
Keras	2.3.1
ハードウェア	
CPU	AMD EPYC 7402P 24-Core Processor 48 個
メモリ	263694036 kB
GPU	NVIDIA Corporation TU102 [TITAN RTX] 2 個

表 3.3: ソフトウェア・ハードウェアの環境

3.3 飛跡対についてのネットワーク

ここでは 3.2 節で紹介した二つのネットワークの内、飛跡対についてのネットワークに関して述べる。主にネットワークの構造に関しては 3.3.1 項で、学習に関しては 3.3.2 項で解説する。また、そのようにして構築、訓練されたネットワーク単体についての性能と評価に関しては、3.3.3 項で述べることとする。

飛跡対についてのネットワークは、崩壊点のタネを探索するためのネットワークであり、入力は二本の飛跡についての情報、出力は飛跡対についての崩壊点の種類や位置である。この崩壊点の種類を考える上で 3.1.1 項で述べた、終状態によって生じる崩壊点の種類の違いを考慮しなければならない。例えば、終状態 $b\bar{b}$ の場合は $b \rightarrow c$ という崩壊過程を辿り、ボトム・フレーバーの Secondary Vertex とチャーム・フレーバーの Tertiary Vertex が生じ、終状態 $c\bar{c}$ の場合はチャーム・フレーバーの Secondary Vertex のみが生じる。また両方の終状態について、これら以外の崩壊点である Others^{*2}を考える必要がある。更に終状態 $b\bar{b}$ の場合について、ボトム・フレーバーの Secondary Vertex からの飛跡とそこから生じたチャーム・フレーバーの Tertiary Vertex からの飛跡を一本ずつ含んだ飛跡対を準崩壊点として定義する。(図 3.7)

以上より、飛跡対についての崩壊点の種類は”非結合な飛跡対 (Not Connected, NC)”、”Primary Vertex (PV)”、”チャーム・フレーバーの Secondary Vertex (SVCC)”、”ボトム・フレーバーの Secondary Vertex (SVBB)”、”チャーム・フレーバーの Tertiary Vertex (TVCC)”、”

*2 タウ粒子の崩壊やストレンジ (s) → フレーバーのハドロンの崩壊、光子変換

終状態 $b\bar{b}$ での準崩壊点 (SVBC)"、"これら以外の崩壊点 (Others)" の計 7 つとなる。

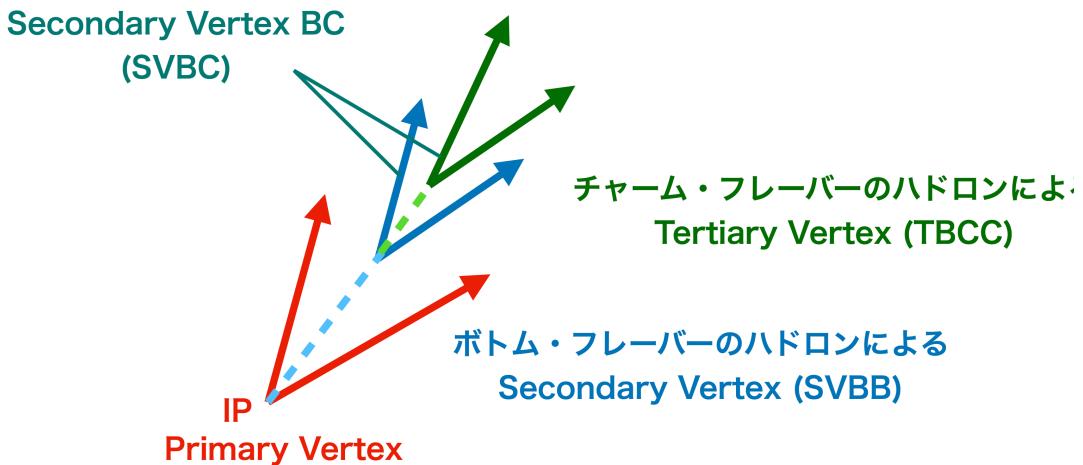


図 3.7: 終状態 $b\bar{b}$ での崩壊点

崩壊点の位置についての訓練データを作成するに当たって、正解ラベルとして図 3.5 の LCFIPlus のフィッティングで得られる計算値を用いた。こちらは回帰によって値を再現する。

3.3.1 ネットワークの構造

飛跡対についてのネットワークとして非常にシンプルなフィードフォーワード構造を使用した。ネットワークの概略図を図 3.8 に示す。

前述したように出力は、7 クラス分類と回帰 1 つである。出力の直前の全結合層で分類問題と回帰問題に分離させた。また、過学習 (Over fitting) を避ける為、Batch Normalization[41] を全結合層の後に配置した。過学習とは、ネットワークが過度に訓練データに適合してしまい、検証データやテストデータへの汎化性能が悪化してしまう、教師あり学習の問題の一つである。また勾配消失への対策として、活性化関数は全て ReLU 関数を使用した。

3.3.2 ネットワークの学習と戦略

訓練データは事象中の全ての飛跡対の組み合わせを考える。よって入力変数は飛跡二本分であるので合計 44 個である。ここで次の二つの事柄に注意しなくてはならない。

- 分類クラスは二つの終状態 $b\bar{b}$ と $c\bar{c}$ を足し合わせたものである
- 分類クラスのデータ数の比が NC や PV が支配的な不均衡データ (Imbalanced Data) となる

各終状態での分類クラスのデータ数の比を図 3.9 に示す。

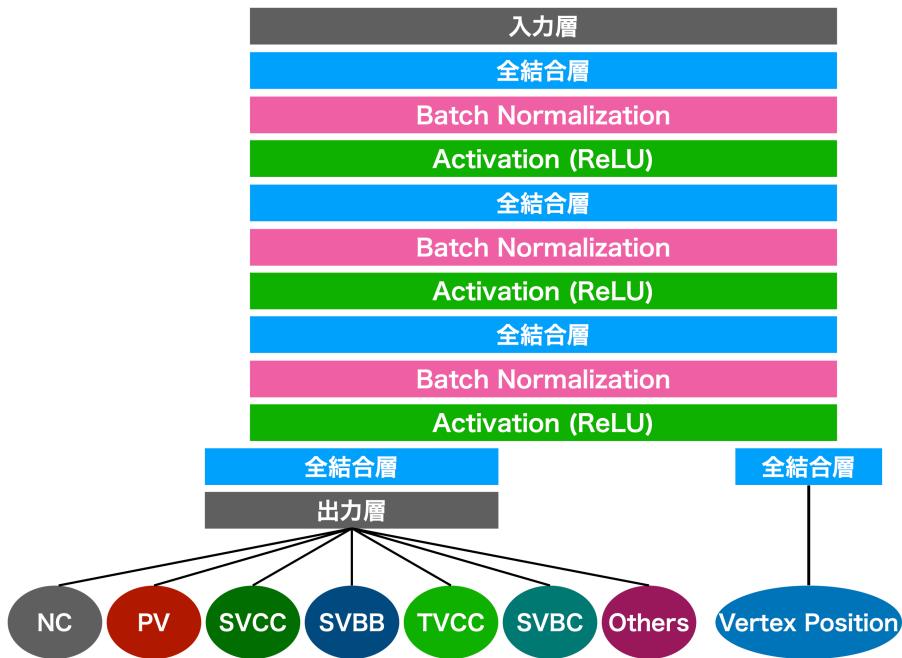


図 3.8: 飛跡対についてのネットワークの概略図。全結合層・Batch Normalization・活性化関数 (ReLU) を三回重ねている。その後、分類問題の為の全結合層と回帰問題の為の全結合層の二つに分離させ、それぞれで出力をを行う。

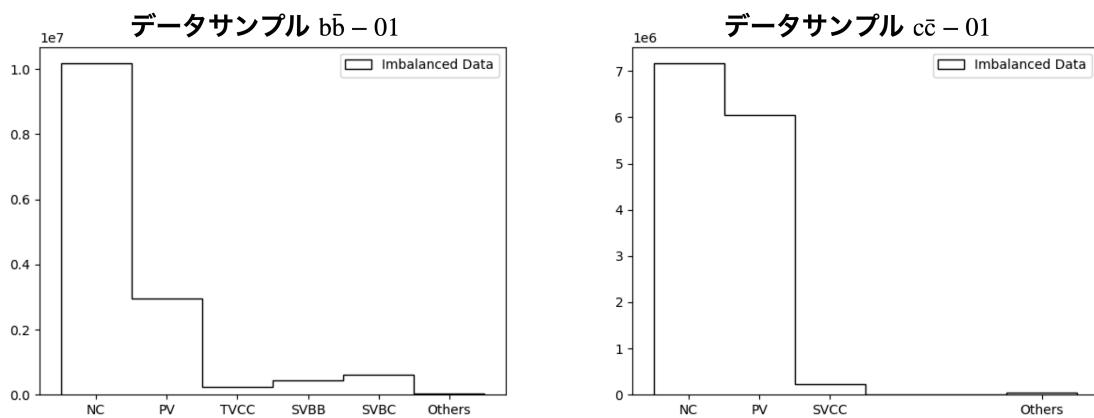


図 3.9: 各終状態での分類クラスのデータ数の比。図左が終状態 $b\bar{b}$ 、図右が終状態 $c\bar{c}$ である。終状態 $c\bar{c}$ では SVCC 以外の SV が基本的に存在しない為、終状態 $b\bar{b}$ と比較して PV の比率が多くなっている。

このような不均衡データについては、少数クラスのデータをかさ増しするオーバーサンプリング、多数クラスのデータを間引くアンダーサンプリング、損失関数のコストに重みをつけるコスト考慮型学習の主に三つの対応策が存在する。オーバーサンプリングやアンダーサンプリングは過学習や情報の欠損などの問題を抱えているため、本研究では基本的にコスト考慮型学習を用いる。ただし、二つの終状態のデータを単純に足し合わせた場合、共通する分類クラスである NC や PV がより顕著になり、他クラスの学習が不十分になると考えられる。このため NC や PV に関しては各終状態毎に半分にランダムサンプリングした後、終状態 $b\bar{b}$ と $c\bar{c}$ を足し合わせた。最終的な訓練データでの分類クラスのデータ数の比を図 3.10 に示す。

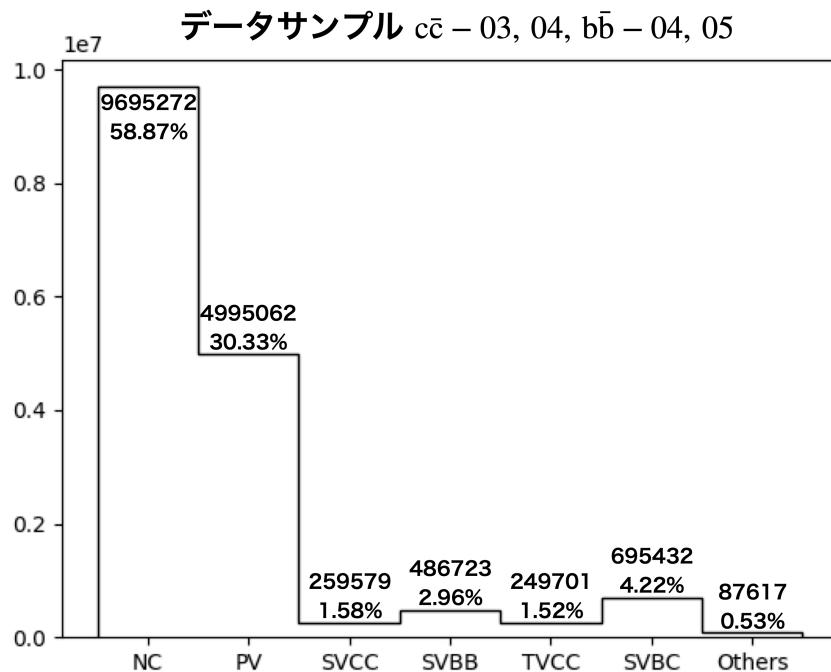


図 3.10: 訓練データでの分類クラスのデータ数の比。実際に使用する訓練データでは NC と PV のカテゴリで全体の 9 割程度となっており、非常に不均衡なデータとなっている。

損失関数 L は分類問題である崩壊点の種類に関しての損失関数 L_{CE} と回帰問題である崩壊点の位置に関しての損失関数 L_{LMSE} の二つの合計となる。

本研究では、コスト考慮型学習として損失関数 L_{CE} の各クラスへの重みは、分類クラスのデータ数の比の逆数を使用し不均衡データへ対策を行なった。ここでは最も数の少ない Others の重みを 1 としている。

$$\begin{aligned}
 L_{CE} = & -0.0090 t_{NC} \log(y_{NC}) - 0.0175 t_{PV} \log(y_{PV}) \\
 & - 0.3375 t_{SVCC} \log(y_{SVCC}) - 0.1800 t_{SVBB} \log(y_{SVBB}) \\
 & - 0.3509 t_{TVCC} \log(y_{TVCC}) - 0.1260 t_{SVBC} \log(y_{SVBC}) \\
 & - 1.0 t_{Others} \log(y_{Others})
 \end{aligned} \tag{3.1}$$

ここで、 t_{NC} , t_{PV} , t_{SVCC} , t_{SVBB} , t_{TVCC} , t_{SVBC} , t_{Others} はそれぞれの分類クラスについての正解ラベルである。また、 y_{NC} , y_{PV} , y_{SVCC} , y_{SVBB} , y_{TVCC} , y_{SVBC} , y_{Others} はそれぞれ

の分類クラスについてネットワークで予想された確率である。

崩壊点の位置についての損失関数 L_{LMSE} として平均二乗誤差を使用した。ただし、図 3.5 からも分かるように、崩壊点の位置は非常に広い範囲に分布しているため、出力や正解ラベルの自然対数を使用し

$$L_{\text{LMSE}} = (\ln(t_{\text{Position}} + 1) - \ln(y_{\text{Position}} + 1))^2 \quad (3.2)$$

とした。ここで、 t_{Position} , y_{Position} はそれぞれ LCFIPlus のフィッターで予想された崩壊点の位置、ネットワークで予想された崩壊点の位置である。

以上より、飛跡対についてのネットワークの損失関数 L は

$$L = w_{\text{vertex}} L_{\text{CE}} + w_{\text{position}} L_{\text{LMSE}} \quad (3.3)$$

となる。ここで、 w_{vertex} , w_{position} は各損失関数 L_{CE} , L_{LMSE} への重みである。

本ネットワークでは崩壊点の位置を学習した後に崩壊点の種類の分類を行なった。これは崩壊点の位置、あるいはそのような情報を抽出できるネットワークのパラメータを利用して、崩壊点の種類の分類を行なって欲しいという狙いからである。これは転移学習 (Transfer Learning, TL) やファインチューニングという手法^{*3}に似た発想であるが、今回はこれら二つの手法とは異なり、ネットワークの重みは全て再学習に使用している。学習は以下の手順を行う。

1. $(w_{\text{vertex}}, w_{\text{position}}) = (0.1, 0.9)$ 、1000 epoch、学習率 LR = 0.001
2. $(w_{\text{vertex}}, w_{\text{position}}) = (0.9, 0.1)$ 、1500 epoch、学習率 LR = 0.001
3. $(w_{\text{vertex}}, w_{\text{position}}) = (0.95, 0.05)$ 、500 epoch、学習率 LR = 0.0001

また、重み更新の最適化手法として SGD を用いた。これは、Adam などでは収束が早すぎ、過学習になる恐れがあったためである。学習には 13175508 サンプル、検証には 3293878 サンプルのデータを使用した。ネットワークの学習可能な重みのパラメータ数を表 3.4 に示す。

3.3.3 ネットワークの評価

ネットワークの評価は 1. ネットワーク間の比較、2. ネットワーク内の理解の二つの観点によって行う。本ネットワークでは、1. ネットワーク間の比較については入力変数やネットワーク構造の異なる幾つかのネットワークのモデルについて比較を行う。2. ネットワーク内の理解については t 分布型確率的近傍埋め込み法 (t-Distributed Stochastic Neighbor Embedding, t-SNE[42]) を用いた次元削減によって、ネットワークによる情報の抽出や各クラスの分離について議論する。

^{*3} これらの手法ではあらかじめ学習済みのネットワークを別の問題解決に活用するテクニックである

層の名称	出力の形状	パラメータ数	接続先
Pair Input	(None, 44)	0	
Dense 1	(None, 256)	11520	Pair Input[0][0]
Batch Normalization 1	(None, 256)	1024	Dense 1[0][0]
Activation ReLU 1	(None, 256)	0	Batch Normalization 1[0][0]
Dense 2	(None, 256)	65792	Activation ReLU 1[0][0]
Batch Normalization 2	(None, 256)	1024	Dense 2[0][0]
Activation ReLU 2	(None, 256)	0	Batch Normalization 2[0][0]
Dense 3	(None, 256)	65792	Activation ReLU 2[0][0]
Batch Normalization 3	(None, 256)	1024	Dense 3[0][0]
Activation ReLU 3	(None, 256)	0	Batch Normalization 3[0][0]
Vertex Dense	(None, 7)	1799	Activation ReLU 3[0][0]
Vertex Output	(None, 7)	0	Vertex Dense[0][0]
Position Output	(None, 1)	257	Activation ReLU 3[0][0]

表 3.4: 飛跡対についてのネットワークにおける訓練可能なパラメータ

1. ネットワーク間の比較

深層学習を取り扱う上で気を付けなければならない問題として、ネットワークの怠け (Lazy) が存在する。ネットワークの怠けとは、ある入力変数について、特定の要素にのみ注目する、あるいは特定の要素を無視してしまうという問題である。飛跡対についてのネットワークでは、深層学習が取り扱いづらい共分散を入力している。ここではネットワークが共分散をどの程度取り扱っているかを評価するため、共分散を含んだ入力変数を用いたモデル A と含んでいないモデル B を構築し比較を行なった。ここでネットワークの構造として、図 3.8 を使用し、そのようなネットワークの構造をネットワーク 1 と呼ぶことにする。

また、3.1.2 節でも述べたように本研究における崩壊点の種類の分離において、崩壊点の位置の再構成は必須である。したがって、そのような崩壊点の位置についての情報を再構成し、適切に処理出来ているかを把握するため、本来回帰問題の正解ラベルとして用いている LCFIPlus によって予想された崩壊点の位置についての情報を出力層の直前で入力するネットワーク 2 を構築した。更に、ネットワーク自体が予想した崩壊点の位置を明示的に入力に使用したネットワーク 3 の構築を行った。これらネットワーク 2、ネットワーク 3 を使用したモデルをそれぞれモデル C、モデル D とし前述のモデル A、モデル B との比較を行なった。それぞれのモデルの詳細を表 3.5 にまとめた。

モデル	入力変数	ネットワーク
モデル A	トラック・パラメータ、共分散行列、電荷、エネルギー	ネットワーク 1
モデル B	トラック・パラメータ、電荷、エネルギー	ネットワーク 1
モデル C	トラック・パラメータ、共分散行列、電荷、エネルギー、崩壊点の位置	ネットワーク 2
モデル D	トラック・パラメータ、共分散行列、電荷、エネルギー	ネットワーク 3

表 3.5: 評価のための飛跡対についてのモデル

比較は各クラス分類についての「ネットワークのスコアとその効率の変化」と「混合行列」

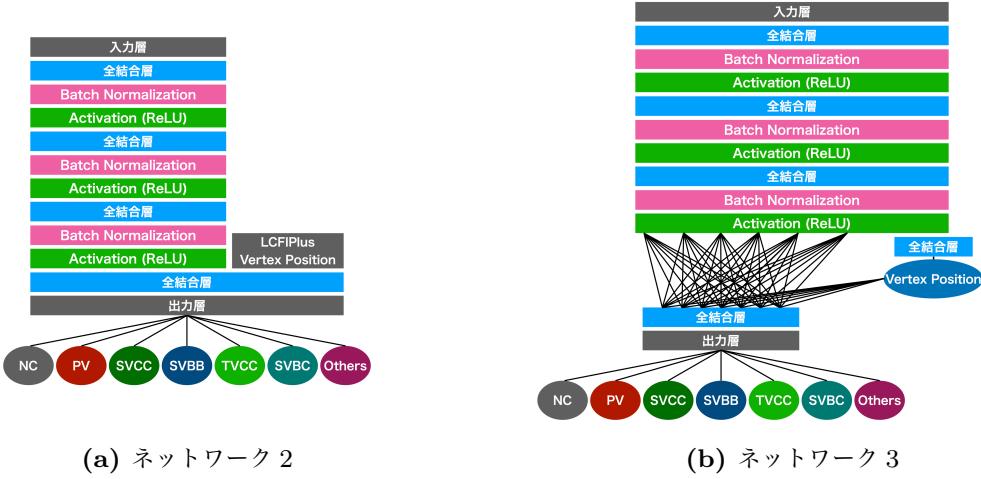


図 3.11: 評価のための飛跡対についてのネットワーク

の二つの指標を用いて行う。また、本ネットワークは7クラス分類を行っているため、ある特定のクラスについてのスコアに閾値を設け、評価することは数学的に厳密に正しくない。しかし、ここでは直感的な理解を優先し評価手法の一つとして採用している。

まず、ネットワークのスコアとクラス分類の効率の関係を図 3.12 に示す。図の横軸はネットワークによって得られる各クラスのスコアに設けた閾値である。また、縦軸は効率を示しており、実線で信号効率を、破線でバックグラウンド効率を表している。ここでは、信号はそれぞれの特定のクラスとし、バックグラウンドはその特定のクラス以外の全てのクラスと定義している。スコアに対して高い閾値を設けると信号・バックグラウンド共に効率は悪化し、逆に低い閾値であれば、全ての要素を信号であると判断してしまうため、両者の効率は最大となっている。したがって、高い信号効率と低いバックグラウンド効率を実現できていれば、良い分類器であると判断できる。図 3.12 では NC・PV・Others については良く分類できているが、各 SV については殆ど見分けられないことが分かる。

これら信号効率を縦軸にバックグラウンド効率を横軸に描画したものが図 3.13 である。このような図を受信者操作特性 (Receiver Operating Characteristic, ROC) 曲線といい、ROC 曲線はその曲線で囲んだ面積が大きければ性能が良いと判断できる。したがって、より左側に張り出している分類器が最も良い性能となる。ここでは、横軸を対数で表現しており、よりバックグラウンド効率の低い領域において、それぞれのモデルの比較を行っている。図 3.13 ではどのクラスについてのモデル間の大きさ差異は見られない。ただし、NC についてのみ Model B の性能が悪いという結果を得られた。

ネットワークがどの程度の効率や純度でクラスを分類出来ているかを更に把握するため混合行列を用いる (図 3.14)。混合行列は横軸をネットワークによって予想されたクラス、縦軸を正解ラベルでのクラスとしてデータを行列化したものである。したがって、対角成分が正答、それ以外は誤答である。ここでは、効率について規格化したものと純度について規格化したものの二つで評価を行う。また、比較のためモデル B,C,D の結果についてはモデル A との相対

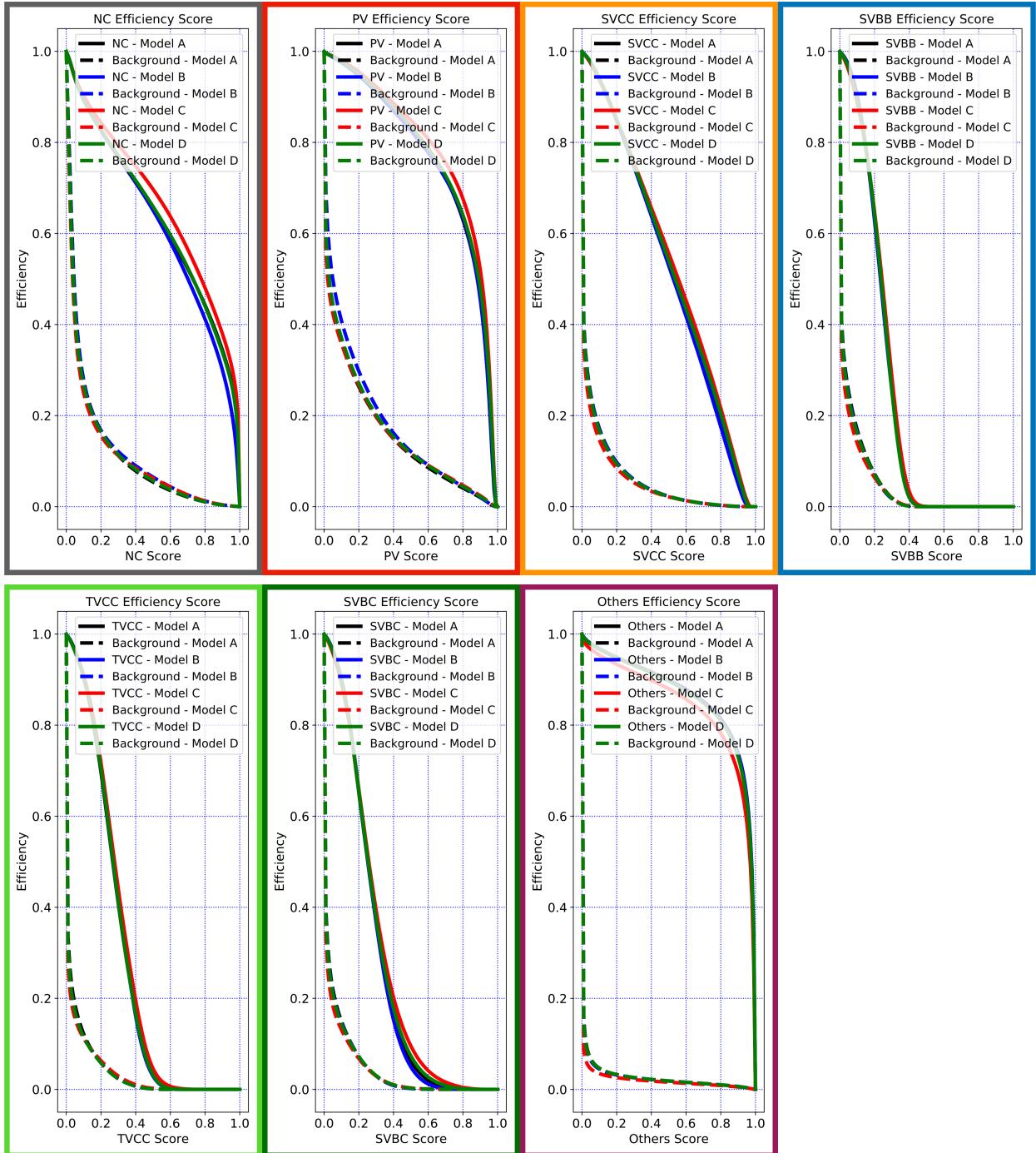


図 3.12: ネットワークのスコアとクラス分類の効率の関係。縦軸は効率、横軸はネットワークのスコアについての閾値である。実線は信号効率、破線はバックグラウンド効率である。黒線はモデル A、青線はモデル B、赤線はモデル C、緑線はモデル D をそれぞれ表している。

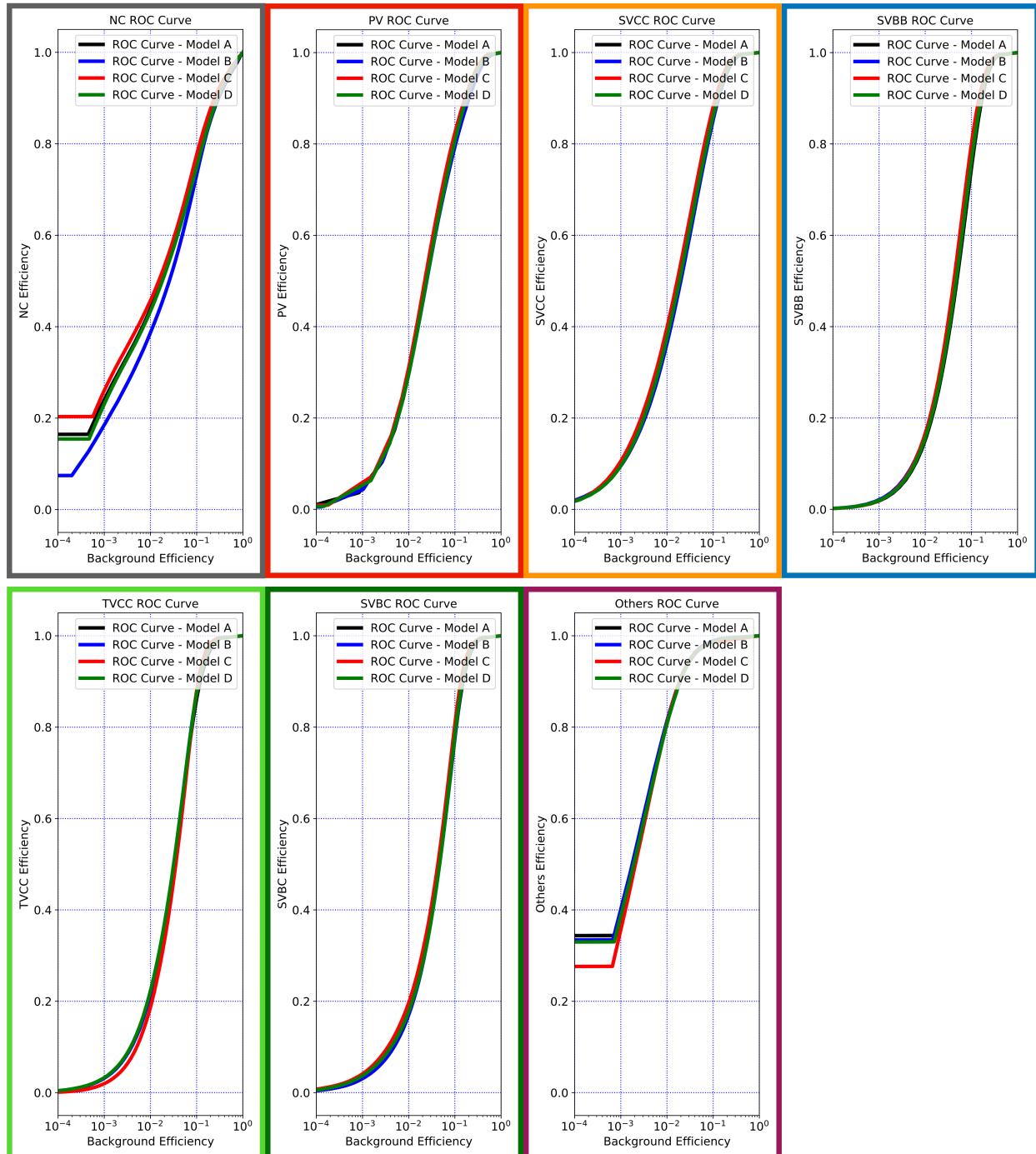


図 3.13: 各モデルの ROC 曲線。縦軸は信号効率、横軸はバックグラウンド効率である。黒線はモデル A、青線はモデル B、赤線はモデル C、緑線はモデル D をそれぞれ表している。

値で表している。

図上部は効率について規格化したものである。そのため正解ラベル方向(行方向)の和が1になっている。混合行列では、NP・PV・SVCC・Othersが比較的高い効率で分離できていることが分かる。一方、各SVについては殆ど分離できておらず、個々のフレーバーや準崩壊点などは区別できていない。ただし、SVとそれ以外との分離は実現できており、SVであることは識別できているが、それがSVのどの崩壊点の種類であるかという点については識別が不十分となっている。

図下部は純度について規格化したものである。そのため予想されたクラス方向(列方向)の和が1になっている。純度については不均衡データとしての特性が強調され、図3.10のようなNCやPVが支配的なクラス配分となっていることから、各SVに対してのNCからの汚染が顕著である。ただし、これは後述する4.2節での崩壊点のタネの選別によってある程度の軽減が可能である。

相対値では対角成分が正であれば性能が高い、それ以外の成分が正であれば性能が低いと判断できる。ModelA・B・Dについてはあまり変化はなく、殆どの差異は0.2%程度に収まっている。ModelCについても同様に大きな違いは確認できないが、効率について規格化した混合行列内のSVBB・TVCC・SVBCの分類性能が向上しており、かつ純度について規格化した混合行列内のNCからのSVCC・SVBB・SVBCへの流入がある程度減少している。ここで、TVCCへの流入については大幅に増大しているが、これはTVCCへの他のSVからの流入が少なくなったことによって相対的に上昇している為であると考えられる。

以上のことから、ネットワークはNC・PV・Othersや単にSVとしては大まかにクラス分類ができると分かる。また、位置に関する程度の予想ができると考えられる(図??)。一方、SV内の各崩壊点の種類への分離は困難でありModelCで多少の改善が見られることから、詳細な位置の再構成には至っておらず、SV内の分離に関しては更なる情報の入力、もしくは抜本的なネットワーク構造の改良が必要であると考えられる。

2. t-SNEによるネットワークの理解

深層学習のネットワーク内部を理解することは非常に難しい課題の一つである。ネットワークが何故、どのようにクラス分類を決定したかを把握する事は容易ではないが、どの程度各クラスを分離して判断できているかは次元削減によってある程度理解が可能である。ここでは、t-SNEという手法を用いて次元削減を行った。

t-SNEとは、高次元の情報についての距離関係を維持しつつ、低次元へマッピングするアルゴリズムである。データの局所的な構造や非線形の次元削減が可能であり、ここではネットワークがどの程度各クラスを分離できているかを二次元の画像として把握することができる。

そのような結果を図3.15に示す。入力変数では殆ど分離できていなかった各クラスが、出力層の直前ではある程度分離できている事が確認できる。また、各SVは非常に近い位置にマップされており、分離が困難であると分かる。更に、分布は線対称な形をしており、これ

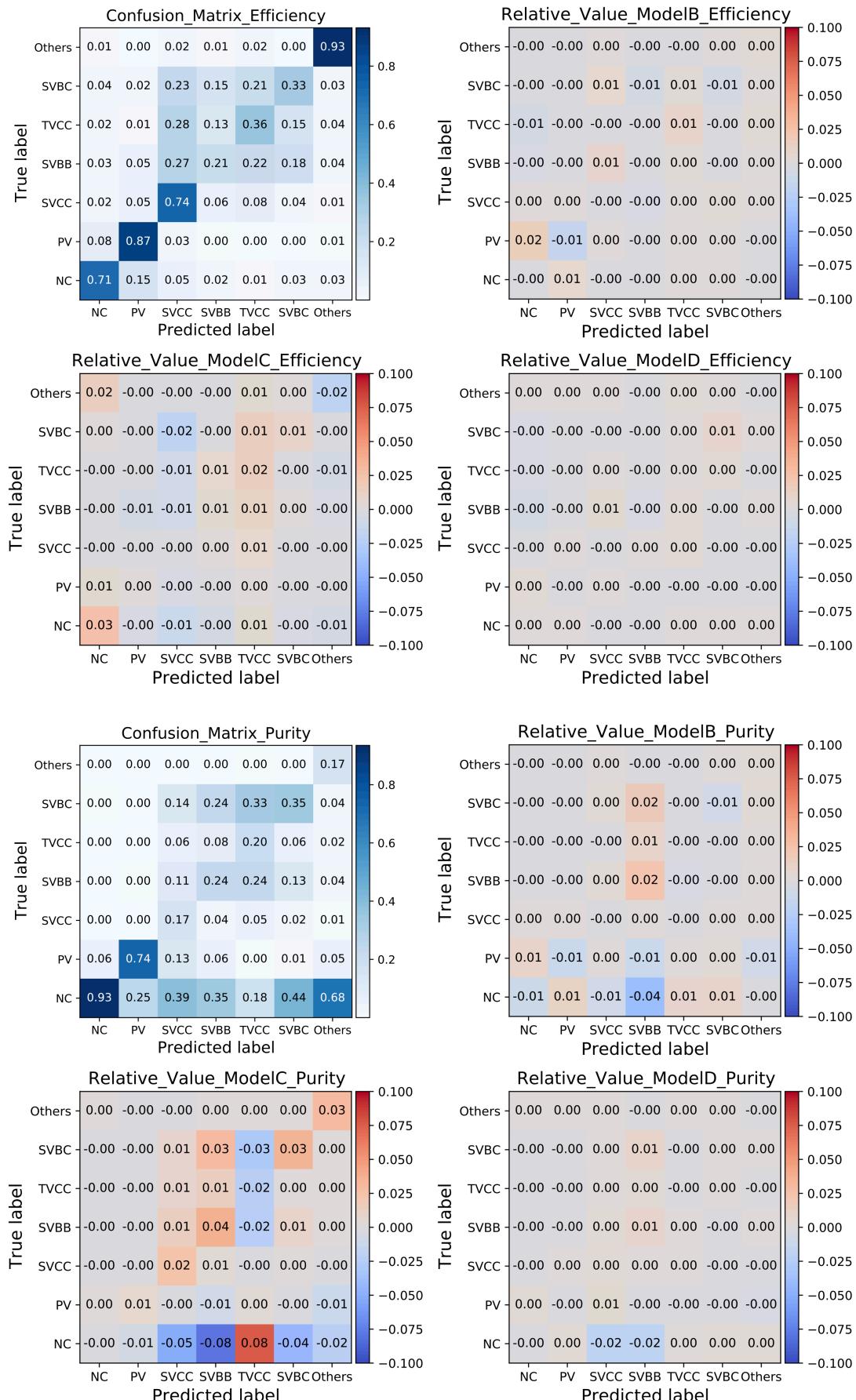


図 3.14: 各モデルの混合行列と各モデルの相対値。図上部は効率について規格化、図下部は純度について規格化した行列である。それぞれ左上がモデル A、右上がモデル B、左下がモデル C、右下がモデル D の結果である。モデル B・C・D に関してはモデル A との相対値で表現している。

は二本の飛跡についての情報を十分に混合できていないということを意味していると考えられる。

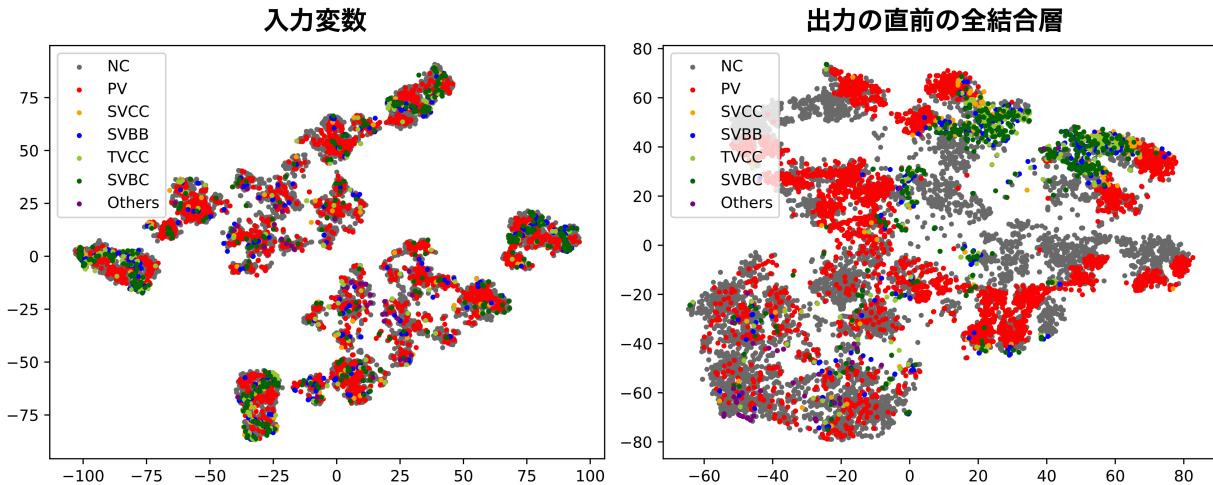


図 3.15: t-SNE による次元削減の比較。左図は入力変数 (22 変数) を二次元に削減したもの、右図は分類問題への分離の直前の全結合層 (256 変数) を二次元に削減したものである。

3.4 任意の数の飛跡についてのネットワーク

ここでは 3.2 節で紹介した二つのネットワークの内、任意の数の飛跡についてのネットワークに関して述べる。前節と同様にネットワークの構造・学習・評価について 3.4.1 項・3.4.2 項・3.4.3 項でそれぞれ解説する。

任意の数の飛跡についてのネットワークは、崩壊点を生成するためのネットワークである。ここでは、リカレントニューラルネットワークの初期状態として飛跡対 (崩壊点のタネ) を入力し、系列データとして事象中の全ての飛跡を入力する。また出力の作り方は Many to Many とし、事象中のそれぞれの飛跡が初期状態の崩壊点のタネに対して結合しているか否かを評価するネットワークを構築する。

3.4.1 ネットワークの構造

まず、任意の数の飛跡についてのネットワークとして図 3.17 のようなネットワークを考える。

前述したように崩壊点のタネを初期状態として使用している。実際には崩壊点のタネは飛跡二本分の情報 (44 個の変数) であり、更に全結合層を介してより抽象的な崩壊点の情報を初期

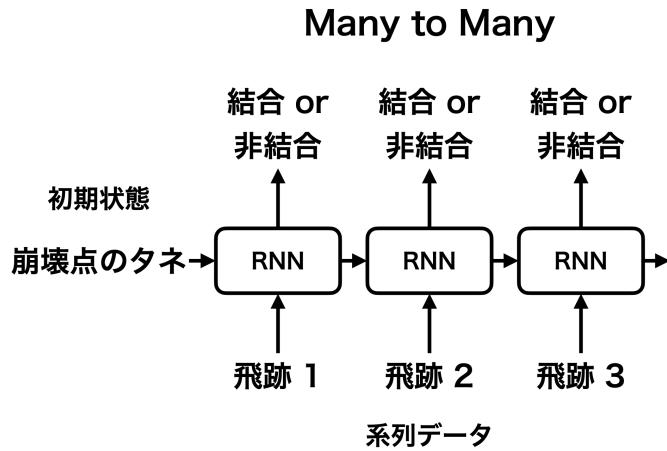


図 3.16: リカレントニューラルネットワークを用いた崩壊点生成。初期状態として崩壊点のタネを入力し、系列データとして事象中の飛跡を入力している。事象中の飛跡はそれ各自本ずつ評価され、出力はそれらの飛跡が初期状態の崩壊点のタネと結合しているか、非結合であるかである。

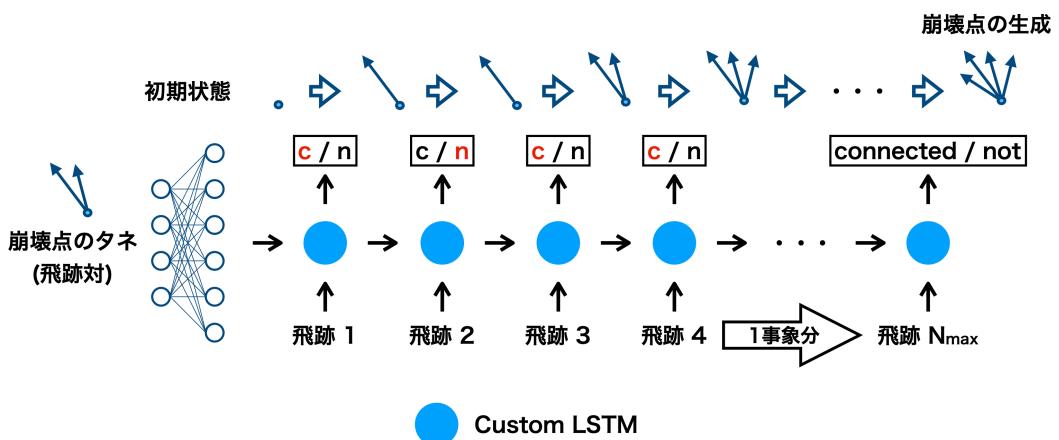


図 3.17: 崩壊点生成のためのリカレントニューラルネットワーク構造。図上部で崩壊点の生成の様子を図示した。飛跡がタネに結合していれば隠れ状態は逐次的に更新されていき、事象中の全ての飛跡を評価した段階で崩壊点が生成される。実際には系列データとして入力される飛跡に関しても、全結合層を用いて Embedding を行い抽象化している。

状態として入力できるようにしている。また系列データとして事象中の全ての飛跡を用いている。出力はこの飛跡のそれが崩壊点のタネと結合しているか否かである。

2.4.3 で解説したようにリカレントニューラルネットワークは系列情報を保持する為、直前の系列に依存するように設計されている。しかし飛跡は本質的に順序を持っていないことに注意する必要がある。図 3.17 では便宜的に飛跡について番号を振り表現しているが、これは人が決めた順序であり、本来、事象中の飛跡は系列データではない。したがって、リカレントニューラルネットワークをそのまま用いることはデータの性質に合わない。この為私は、リ

カレントニューラルネットワークの一つである LSTM を拡張し、新しい独自のリカレントニューラルネットワークの構造を構築した。

図 3.18 はそのような独自のリカレントニューラルネットワークについての 1 系列分のステップの詳細である。

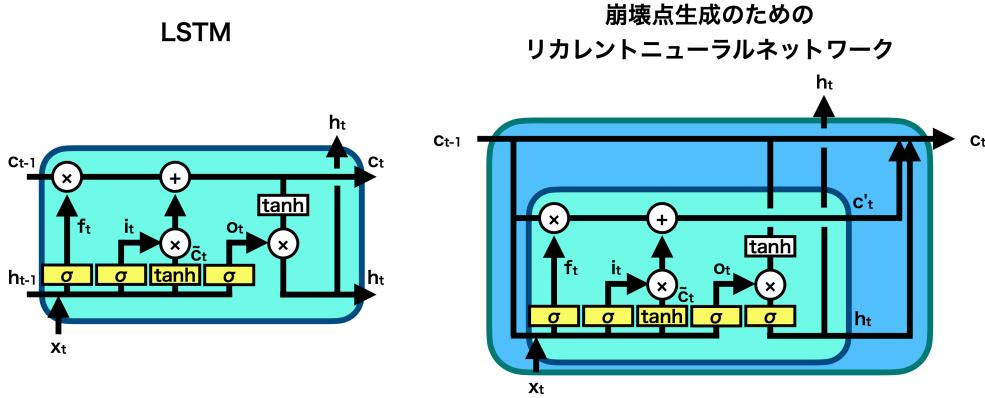


図 3.18: 系列 1 ステップについての独自リカレントニューラルネットワーク構造。図 3.17 の青丸の一つに相当している。

LSTM との大きな構造の違いは、短期記憶である h_t が隠れ状態として入出力されていない点である。入力は隠れ状態の記憶セル c_{t-1} と系列データ x_t の二つである。また出力は結合・非結合を判定する h_t と隠れ状態の記憶セル c_t の二つである。隠れ状態として h_t が使用されていないため、内部の構造も通常の LSTM とは少し異なり、出力 h_t と記憶セル c_t はそれぞれ

$$\begin{aligned} c_t &= (1 - h_t)c_{t-1} + h_t c'_t \\ c'_t &= c_{t-1} \sigma(W_f x_t + R_f c_{t-1}) + \tanh(W_c x_t + R_c c_{t-1}) \sigma(W_i x_t + R_i c_{t-1}) \\ h_t &= \sigma(D_h [\tanh(c_{t-1}) \sigma(W_o x_t + R_o c_{t-1})]) \end{aligned} \quad (3.4)$$

となる。第二式の c'_t は更新された記憶セルを示している。第三式は出力ゲートに更に全結合層を掛けた形となっており、二値分類の為の出力を生成している。第一式では、更新された記憶セル c'_t と直前の系列での記憶セル c_{t-1} 、出力 h_t を用いて現在の記憶セル c_t を計算している。二値分類であるので、 h_t は 0 から 1 の値を持つはずである。したがって、第一式は結合 ($h_t \sim 1$) していれば更新された記憶セル c'_t が、非結合 ($h_t \sim 0$) であれば、直前の系列での記憶セル c_{t-1} が現在の記憶セル c_t となることを示している。

初期状態は崩壊点のタネであるので、以上の演算は次のように (図 3.19) 解釈できる。

1. $t-1$ 番目の崩壊点 c_{t-1} と t 番目の飛跡 x_t が結合しているか否かの評価 h_t を行う
2. $t-1$ 番目の崩壊点 c_{t-1} と t 番目の飛跡 x_t を用いて更新された崩壊点 c'_t を計算する
3. $t-1$ 番目の崩壊点 c_{t-1} と t 番目の飛跡 x_t が結合しているならば更新された崩壊点 c'_t を、結合していないならば $t-1$ 番目の崩壊点 c_{t-1} を t 番目の崩壊点 c_t として選択する

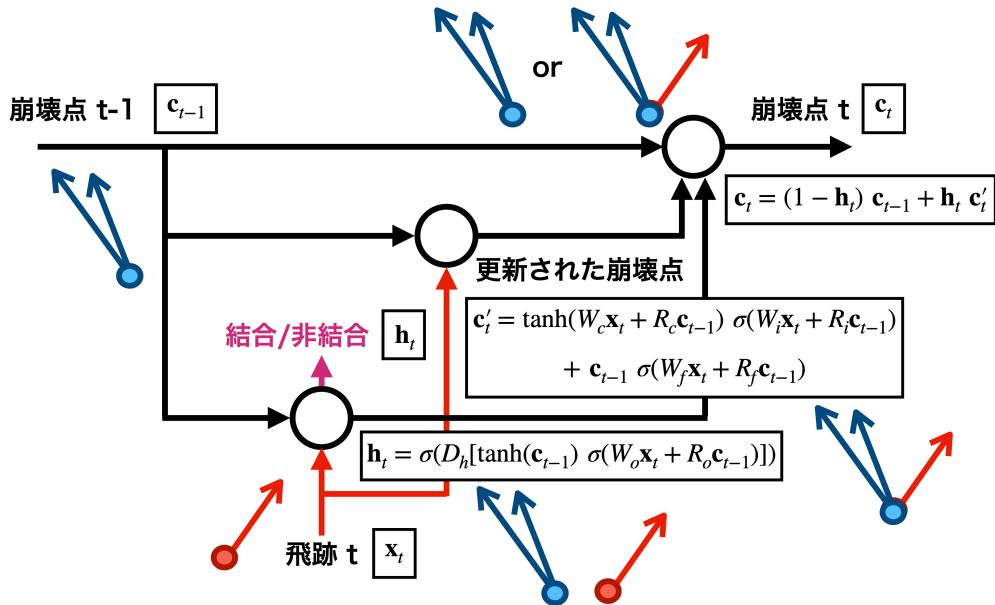


図 3.19: 独自リカレントニューラルネットワーク構造の解釈。それぞれの丸が演算に対応している。左下が演算 1、中央が演算 2、右上が演算 3 である。

短期記憶を破棄する事によって、独自リカレントニューラルネットワークでは、事象中の飛跡の順序(短期記憶)にできる限り依存させず、更に崩壊点のタネに対して飛跡を足していくことによって更新される崩壊点の情報(記憶セル)を表現している。

私たちはシミュレーションデータとして既に一つの事象分の全ての飛跡についての情報を持っている。したがって、作成したリカレントニューラルネットワークをエンコーダー・デコーダーモデルに組み込むことで、事象についての情報(コンテキスト)を活用することができると言えられる。また、エンコーダー・デコーダーモデルの間に Attention を組み込むことも同様に自然な発想である。その様なネットワークを図 3.20 に示す。

図の上部はエンコーダー部である。エンコーダー部では事象中の飛跡から事象全体の情報を(コンテキスト)を抽出している。ここでは先ほどの図 3.17 で紹介したネットワークを双方向リカレントニューラルネットワークとして使用している。ただし、出力 h_t の次元数を減らさないようにするために全結合層 D_h を取り除いている。また図中で表現しているように、双方向から入力されている崩壊点のタネはそれぞれ別の全結合層によって情報を抽象化されている。

図の下部はデコーダー部である。デコーダー部ではエンコーダー部で抽出された情報と崩壊点のタネ、事象中の飛跡を使用して崩壊点のタネにそれぞれの飛跡が結合しているか否かを判別している。エンコーダー部で抽出された情報は Attention によって適切に評価され、デコーダー部の“ある”飛跡がエンコーダー部の任意の飛跡に対して注意を払って、事象中の情報を取得できるようになっている。またエンコーダー・デコーダーモデルへの拡張後も、このネットワークの基本構造がリカレントニューラルネットワークであることに変わりはない為、飛跡の本数を任意に変えることが可能である。

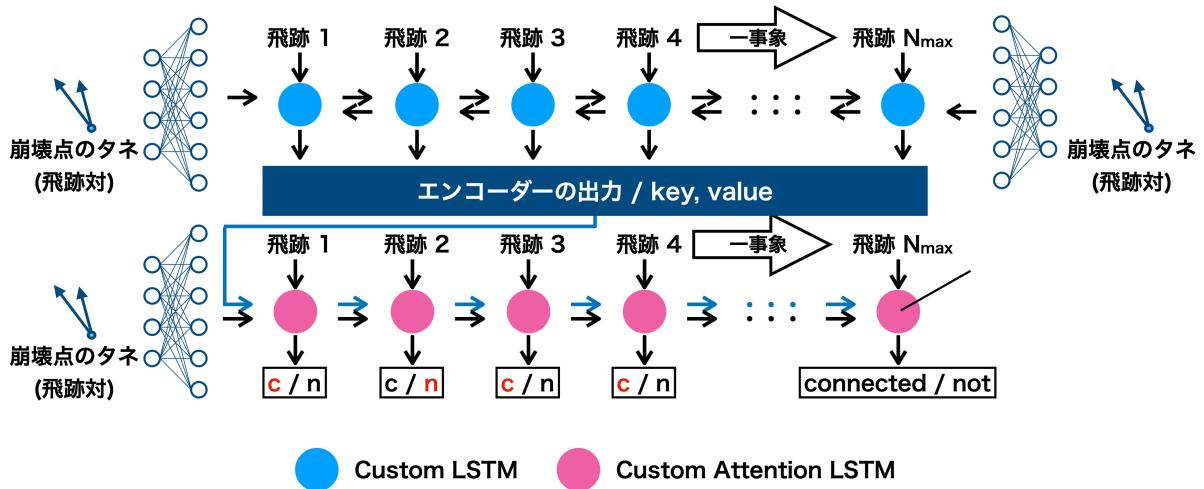


図 3.20: Attention を組み込んだエンコーダー・デコーダーモデルへの拡張。図上部がエンコーダー部、図下部がデコーダー部をそれぞれ示している。初期状態や系列データはエンコーダー部とデコーダー部のそれぞれに入力されている。エンコーダー部では双方向 RNN を用いている。ただし、このステップは独自構造のネットワークになっている。エンコーダー部の出力は図中央の Key・Value に集められ、デコーダー部に入力される。デコーダー部での出力はそれぞれの飛跡がタネと結合しているか、非結合であるかの二値分類である。

図 3.18 で示したネットワーク構造は Attention には対応していないため、新たなネットワークの構築が必要である。そのようなネットワークを図 3.21 に示す。

本研究では、エンコーダー部で抽出された情報 (Key) をリカレントニューラルネットワークの隠れ状態の一つとして入力している。また、この Key はネットワークの各ステップについて共通の値を使用しており、長期記憶・短期記憶と比べて不变記憶のような役割を果たしている。Attention Weight の計算方法として Additive Attention を採用した。 t 番目のコンテキスト γ_t は次のように計算される。

$$\begin{aligned} \gamma_t &= \alpha_t V \\ \alpha_t &= (\alpha_{t,0}, \alpha_{t,1}, \alpha_{t,2}, \dots, \alpha_{t,i}, \dots) \\ &= \left(\frac{\exp(e_{t,0})}{\sum_j \exp(e_{t,j})}, \frac{\exp(e_{t,1})}{\sum_j \exp(e_{t,j})}, \frac{\exp(e_{t,2})}{\sum_j \exp(e_{t,j})}, \dots, \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})}, \dots \right) \quad (3.5) \\ e_t &= u_{\text{energy}}(K U_{\text{key}} + X_t U_{\text{query}}) \end{aligned}$$

ここで、Key を K 、Value を V 、 t 番目の Query である飛跡 x_t を重ねた行列を X_t 、 t 番目の Attention Weight を α_t 、 t 番目の Query についてのエネルギーを e_t とした。また、Additive Attention における重み行列をそれぞれ u_{energy} 、 U_{key} 、 U_{query} と置いた。

添字 i, j はエンコーダー部の系列、添字 t はデコーダー部の系列である。したがって t 番目の Attention Weight α_t はエンコーダー部の飛跡の数だけ要素を持っており、デコーダー部の全ての飛跡について Attention Weight を計算した時、Attention Weight はエンコーダー部の飛跡の数 × デコーダー部の飛跡の数の行列となる。この Attention Weight 行列はネット

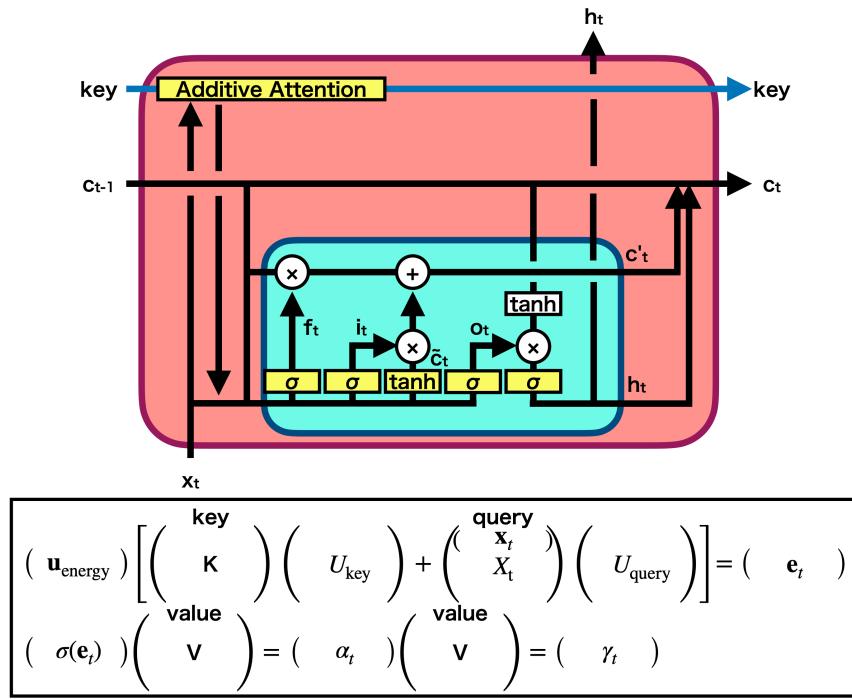


図 3.21: 独自リカレントニューラルネットワークの Attention への拡張。図 3.20 の赤丸の一つに相当している。図上部の青線は不变記憶 Key・Value の流れを表している。Attention に関する演算が加えられている点以外は図 3.18 と同様の構造である。

ワーク内部を把握する上で非常に重要な情報である。図 3.21 では表現していないがオプションとして Attention Weight を出力することで、ネットワーク内部をある程度理解することができる。

得られた t 番目のコンテキスト γ_t は、出力 \mathbf{h}_t や更新された崩壊点 \mathbf{c}'_t の計算に使用される。

$$\begin{aligned} \mathbf{c}_t &= (1 - \mathbf{h}_t) \mathbf{c}_{t-1} + \mathbf{h}_t \mathbf{c}'_t \\ \mathbf{c}'_t &= \mathbf{c}_{t-1} \sigma(W_f \mathbf{x}_t + R_f \mathbf{c}_{t-1} + C_f \gamma_t) \\ &\quad + \tanh(W_c \mathbf{x}_t + R_c \mathbf{c}_{t-1} + C_c \gamma_t) \sigma(W_i \mathbf{x}_t + R_i \mathbf{c}_{t-1} + C_i \gamma_t) \\ \mathbf{h}_t &= \sigma(D_h [\tanh(\mathbf{c}_{t-1}) \sigma(W_o \mathbf{x}_t + R_o \mathbf{c}_{t-1} + C_o \gamma_t)]) \end{aligned} \tag{3.6}$$

式中ではコンテキスト γ_t に関する各ゲートそれぞれの重み行列を C_f, C_c, C_i, C_o と置いた。 t 番目のコンテキスト γ_t についての演算を加えている点以外は図 3.18 でのネットワークの演算と全く同様である。

3.4.2 ネットワークの学習と戦略

訓練データとして必要な情報は、初期状態としての飛跡対（崩壊点のタネ）と事象中の全ての飛跡である。また、正解ラベルはそれぞれの飛跡が崩壊点のタネと結合しているか否かの MC 情報である。推論時は、初期状態の崩壊点のタネとして非結合な飛跡対が入力される場合を考えられるが、本研究ではネットワークの学習時は崩壊点のタネとして結合している飛跡対（PV・SVCC・SVBB・TVCC）のみを使用する。ここで、準崩壊点 SVBC は崩壊点生成にお

いて雑音となりうる可能性があるため含んでいない。

リカレントニューラルネットワークでは、推論時は系列長を変えることができるが、学習時は重み更新の計算のため系列長を揃える必要がある。^{*4} ここでの系列長は事象中の飛跡の本数であった。このため、不足している飛跡の本数をゼロ埋め (Zero padding) し、ゼロ埋めした飛跡が学習に影響しないよう損失関数においてマスクしている。本研究では最も飛跡数の多い事象との兼ね合いから系列長を 60 本とし、それ以下の本数の事象については 60 本となるようにゼロ埋めしている。

最終的な入力変数の大きさを表 3.6 に示す。

崩壊点のタネ	(samples, 44)
事象中の全ての飛跡	(samples, 60, 23)

表 3.6: 任意の数の飛跡についてのネットワークの入力変数の大きさ

ここで、第一引数はデータのサンプル数である。訓練データは全ての崩壊点のタネについて一つのサンプルが生成されるため、全ての事象 ($c\bar{c} - 05$, $b\bar{b} - 06$) を使用すると非常に時間がかかる。よって、本研究では全ての事象を使用して訓練データを作成した後、1 エポック毎にランダムに 50000 サンプルを訓練に 10000 サンプルを検証に使用している。事象中の全ての飛跡については飛跡の本数である 60 本とそれぞれの飛跡について 22 個の変数を持っている。また、ゼロ埋めした飛跡との区別のためのマスク変数 (0, 1) を一つ加えている。

崩壊点生成において、飛跡は順序を持って足されていく。短期的な順序に依存しないような独自のネットワークを構築しているが、そのような人によって決められた飛跡の順序にネットワークが依存してしまうことは、できる限り避けねばならないため、1 エポック毎に飛跡の系列順をランダムにシャッフルしている。

損失関数は二値交差エントロピー誤差を使用した。ただし、前述したようにゼロ埋めした飛跡については損失関数や正答率の計算ではマスクした。

$$L = -M_{ZP} t_C \log(y_C) - M_{ZP} (1 - t_C) \log(1 - y_C) \quad (3.7)$$

M_{ZP} はゼロ埋めのためのマスク変数 (0, 1) である。

ハイパーパラメータとして、最適化手法は Adam、学習率は 0.001、エポック数は 100、バッチサイズは 32 を用いた。また、ネットワークの学習可能な重みのパラメータ数を表??に示す。

3.4.3 ネットワークの評価

ネットワークの評価は飛跡対についてのネットワークと同様に、1. ネットワーク間の比較と 2. ネットワーク内の理解の二つの観点で行う。任意の数の飛跡についてのネットワークでは、1. ネットワーク間の比較については標準的な LSTM と本研究で構築した独自のネット

^{*4} 実際にはバッチサイズ毎に

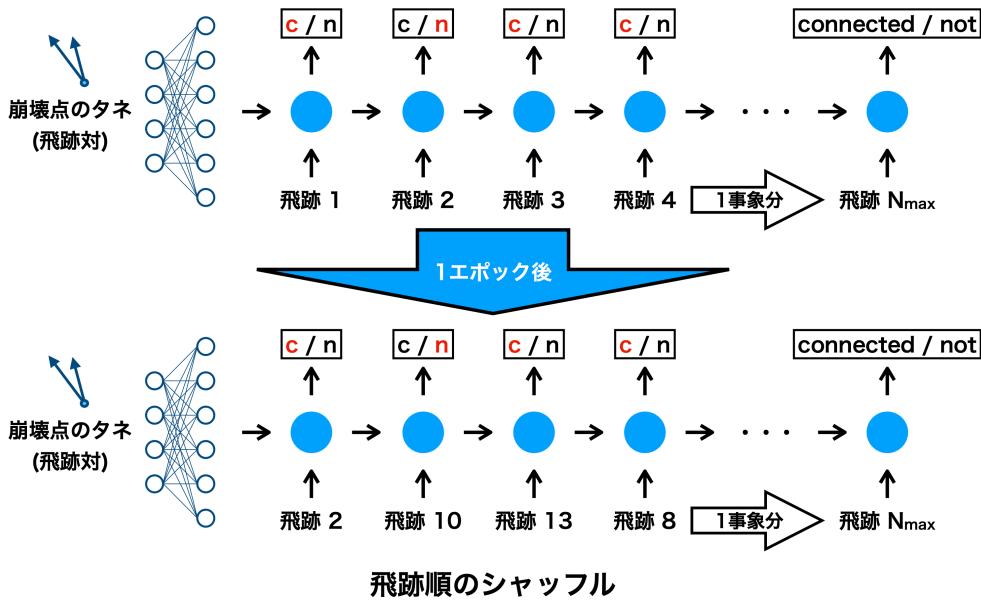


図 3.22: 飛跡順のシャッフル。1 エポック毎に系列データの飛跡順序をシャッフルし、崩壊点のタネへ足し合わせる順番への依存を取り除いた。この順序のシャッフルは学習や検証に使用する全てのデータについて同一である。

ワークの比較を行う。更に、終状態の違いや崩壊点のタネの違いについて、各データ属性に特化したネットワークとそれら全てのデータを用いて学習した標準的なネットワークについての比較を行う。2. ネットワーク内の理解については任意の数の飛跡についてのネットワークは内部に Attention を持っているため、Attention Weight を確認することにより、エンコーダー部からどのような情報を抽出しているかについて調査する。

1. ネットワーク間の比較

まず、以下の三つのネットワークの比較によって独自のネットワーク構造がどの程度効果的であるかの確認を行う。

- 標準的な LSTM : 図 3.17 のようなネットワークの個々のステップを標準的な LSTM に置き換えたネットワーク
- 独自の LSTM : 図 3.17 のようなネットワーク
- 独自の Attention LSTM : 図 3.20 のようなネットワーク

これらのネットワークに関して損失と正答率のエポック毎の変化を図 3.23 に示す。ここでは全ての評価基準についてゼロ埋めした飛跡を取り除いている。また、ネットワークの構造やデータ特性を考慮し、正答率や真陽性率 (True Positive Rate, TPR)、真陰性率 (True Negative Rate, TNR) から初期状態となる飛跡対を取り除いている。したがって正答率、TPR、TNR はそれぞれ次のような定義となる。

層の名称	出力の形状	パラメータ数	接続先
Pair Input	(None, 44)	0	
Encoder Input	(None, 60, 23)	0	
Decoder Input	(None, None, 23)	0	
Encoder Forward Dense 1	(None, 256)	11520	Pair Input[0][0]
Encoder Backward Dense 1	(None, 256)	11520	Pair Input[0][0]
Encoder Forward Activation 1	(None, 256)	0	Encoder Forward Dense 1[0][0]
Encoder Backward Activation 1	(None, 256)	0	Encoder Backward Dense 1[0][0]
Encoder Forward Dense 2	(None, 256)	11520	Encoder Forward Activation 1[0][0]
Encoder Backward Dense 2	(None, 256)	11520	Encoder Backward Activation 1[0][0]
Encoder Forward Activation 2	(None, 256)	0	Encoder Forward Dense 2[0][0]
Encoder Backward Activation 2	(None, 256)	0	Encoder Backward Dense 2[0][0]
Encoder Embedding Dense	(None, 60, 256)	6144	Encoder Input[0][0]
Bidirectional Encoder VLSTM	(None, 60, 512)	1050624	Encoder Embedding Dense[0][0] Encoder Forward Activation 2[0][0] Encoder Forward Activation 2[0][0] Encoder Backward Activation 2[0][0] Encoder Backward Activation 2[0][0]
Reshape Bidirectional Encoder	(None, 27136)	0	Bidirectional Encoder VLSTM[0][0]
Decoder Dense 1	(None, 256)	11520	Pair Input[0][0]
Decoder Activation 1	(None, 256)	0	Decoder Forward Dense 1[0][0]
Decoder Dense 2	(None, 256)	11520	Decoder Forward Activation 1[0][0]
Decoder Activation 2	(None, 256)	0	Decoder Forward Dense 2[0][0]
Decoder Embedding Dense	(None, None, 256)	6144	Encoder Input[0][0]
Decoder Attention VLSTM	(None, None, 1)	1246976	Decoder Embedding Dense[0][0] Reshape Bidirectional Encoder[0][0] Decoder Activation 2[0][0]

表 3.7: 任意の数の飛跡についてのネットワークにおける訓練可能なパラメータ

$$\text{Accuracy} = \text{Mean}(M_{\text{ZP}} M_{\text{IS}} \text{Equal}(\text{true}, \text{pred}))$$

$$\text{True Positive Rate} = \text{Mean}(M_{\text{ZP}} M_{\text{IS}} \text{Equal}(\text{true}_C, \text{pred}_C)) \quad (3.8)$$

$$\text{True Negative Rate} = \text{Mean}(M_{\text{ZP}} M_{\text{IS}} \text{Equal}(\text{true}_{NC}, \text{pred}_{NC}))$$

	正の予測	負の予測
正の正解	真陽性 (True Positive, TP)	偽陰性 (False Negative, FN)
負の正解	偽陽性 (False Positive, FP)	真陰性 (True Negative, TN)

標準的な LSTM は系列の順序を重視しているため学習が安定していないことが分かる。また、独自のネットワーク同士の比較においても、エンコーダー部からのコンテキストを受け取ることのできる独自の Attention LSTM の性能が高くなっている。以上のことから、標準的な LSTM を単純に使うことは不適であり、独自のネットワーク構造による大幅な性能の改善が可能であると確認できる。

次に独自の Attention LSTM についての比較を行う。ここでは訓練データをデータ属性毎

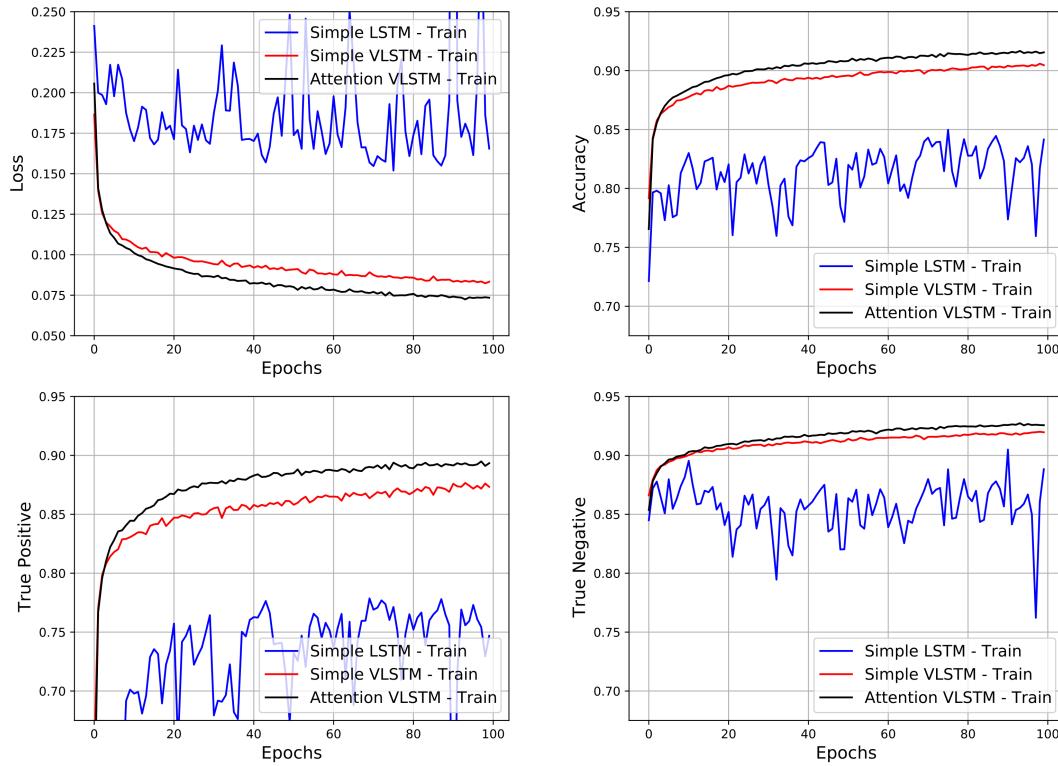


図 3.23: 標準的な LSTM と独自のネットワークの比較。左上が損失、右上が正答率、左下が TPR、右下が TNR である。それぞれ青線は標準的な LSTM、赤線は独自の LSTM、青線は独自の Attention LSTM についての結果を表している。

に分離した。ここでデータ属性とは終状態 ($c\bar{c} \cdot b\bar{b}$)、崩壊点のタネ (PV · SV) の 4 つである。それぞれデータ属性のみで構成された訓練データで学習した特化型のネットワークを用意した。それらのネットワークと全てのデータを使用した標準的なネットワークについての比較の結果を図 3.24 と図 3.25 に示す。

- ALL : 全てのデータを使用した標準的なネットワーク
- CC, BB : 終状態が $c\bar{c}$ 、 $b\bar{b}$ のデータのみを使用した特化型のネットワーク
- PV, SV : 崩壊点のタネが PV、SV のデータのみを使用した特化型のネットワーク

比較に関しては ROC 曲線を使用している。評価に際してのデータは特化型のネットワークの学習に使用した訓練データと同じ属性のものを用いた。終状態 $c\bar{c}$ や終状態 $b\bar{b}$ に関しては基本的な性能の差がほとんどないと分かる。したがって、個々の終状態のデータに関してはその特性に大きな差異は存在しないと判断できる。各崩壊点のタネに特化したネットワークに関しては両者とも性能の向上が見られた。特に、PV のタネに特化したネットワークは結合と非結合との分離が非常に良くできている。ROC 曲線では SV に特化したネットワークが標準的なネットワークより大幅に性能が改善されていると分かる。

2. ネットワーク内の理解

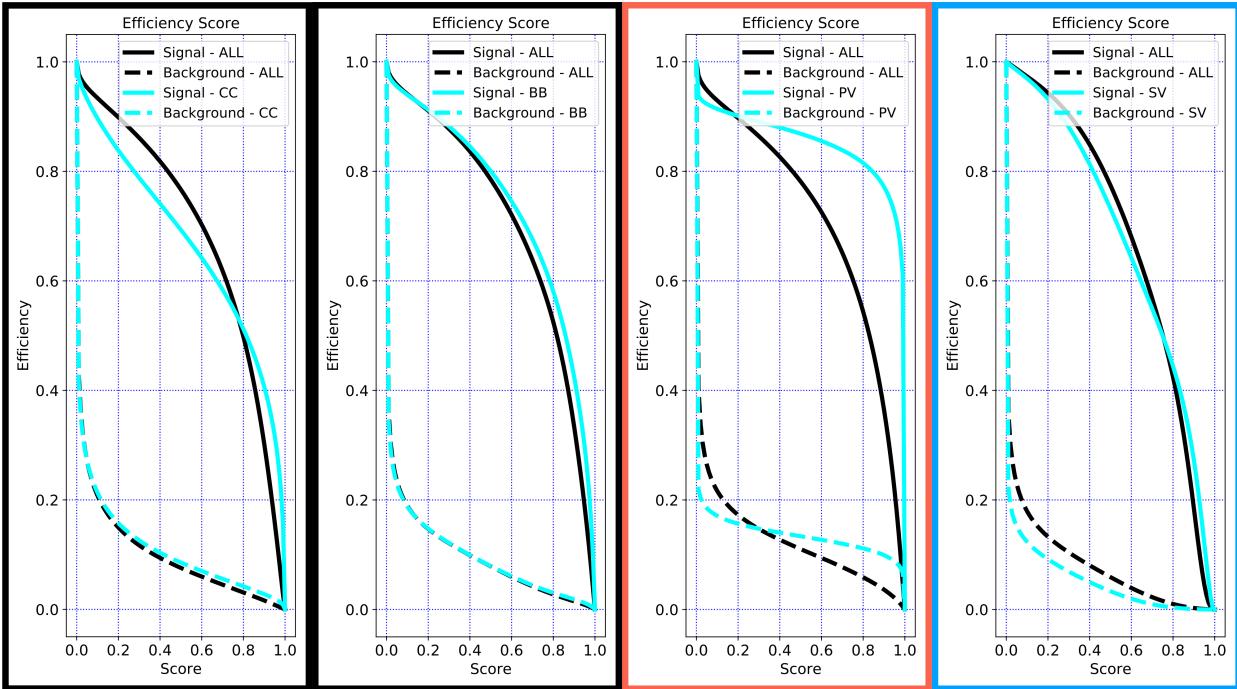


図 3.24: 各データ属性の効率とスコアの関係。縦軸は効率、横軸はネットワークのスコアについての閾値である。実線は信号効率、破線はバックグラウンド効率である。黒線は全データを使用した標準的なネットワーク、青線は個々の特化型のネットワークをそれぞれ表している。

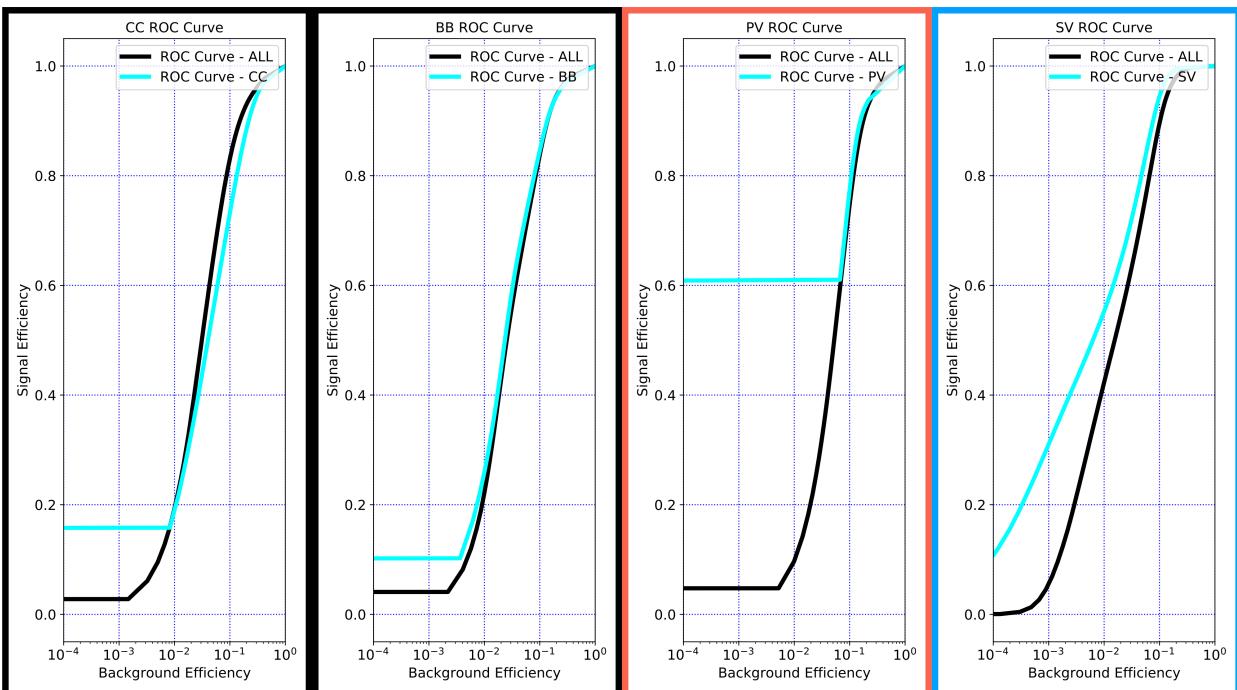


図 3.25: 各データ属性の ROC 曲線。縦軸は信号効率、横軸はバックグラウンド効率である。黒線は全データを使用した標準的なネットワーク、青線は個々の特化型のネットワークをそれぞれ表している。

飛跡対についてのネットワークとは異なり、任意の数の飛跡についてのネットワークは Attention を持っている。したがって、ネットワークの内部を Attention Weight を確認することによって、ネットワーク内部を把握することができる。具体的には、エンコーダー部の飛跡がデコーダー部のどのような飛跡に注意し情報を受け取っているかを確認することが可能である。そのような Attention Weight について、標準的なネットワークと PV のタネを用いて作成したサンプルを図 3.26 に示す。

図上部は横軸がデコーダー部の飛跡番号、縦軸がエンコーダー部の飛跡番号となっている。ここでは Attention Weight はカラースケールで表されている。デコーダー部の飛跡による Attention Weight の出力を縦方向に積んだような表現方法である。

図下部はエンコーダー部の飛跡を上に、デコーダー部の飛跡を下に並べた図となっている。ここでは Attention Weight は透過度で表現されている。また、デコーダー部の飛跡の内、崩壊点のタネと結合しているものを赤、結合していないものを青で描画している。デコーダー部の飛跡の数が少ないのでゼロ埋めを取り除いている為である。

これらの図から結合している飛跡はエンコーダー部の飛跡に注意し、結合していない飛跡はエンコーダー部のゼロ埋めに注意している傾向があると分かる。

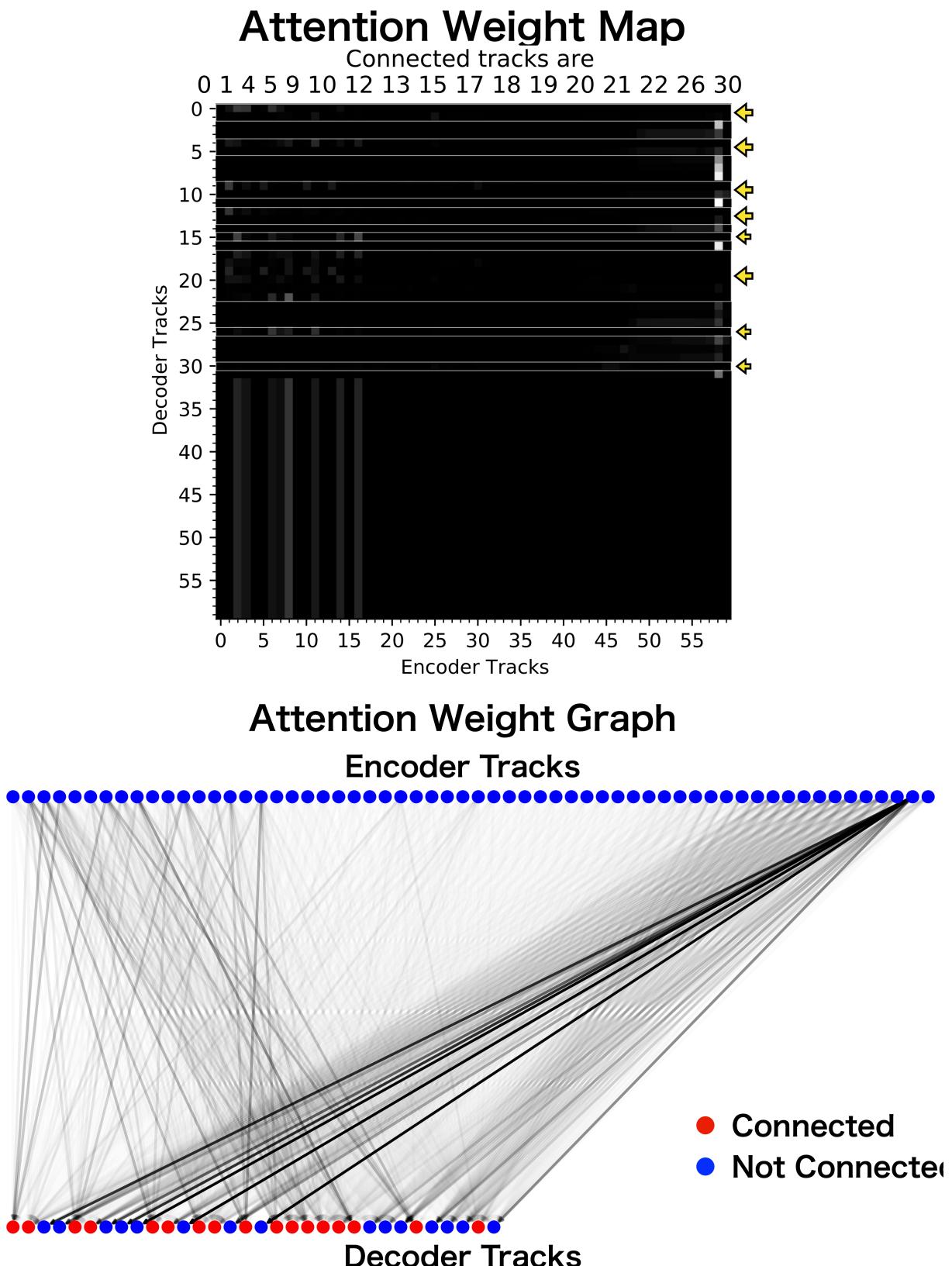


図 3.26: 図上部は Attention Weight を行列として表現した図、図下部は Attention Weight をより直感的に示した図である。図上部では縦軸がデコーダーの飛跡番号、横軸がエンコーダーの飛跡番号となっている。また、カラースケールは Attention Weight である。図下部では上段にデコーダーの飛跡、下段にデコーダーの飛跡を描画している。デコーダーの飛跡ではタネと結合している飛跡については赤丸を使用している。

第4章

深層学習を用いた崩壊点検出

4.1 崩壊点検出アルゴリズム

4.2 崩壊点検出の最適化と評価

4.3 崩壊点検出の性能

第 5 章

現行の手法との比較

5.1 崩壊点検出単体での比較

5.2 フレーバータギングを含めた比較

第 6 章

まとめと今後の展望

謝辞

付録 A

ソースコード

A.1 飛跡対についてのネットワーク

A.2 任意の数の飛跡についてのネットワーク

A.3 崩壊点の再構成

参考文献

- [1] T. Kramer, Track parameters in LCIO, (2006).
- [2] P. Bambade *et al.*, The international linear collider: A global project, 2019, 1903.01629.
- [3] T. I. Collaboration, International large detector: Interim design report, 2020, 2003.01116.
- [4] T. Behnke *et al.*, The International Linear Collider Technical Design Report - Volume 1: Executive Summary, (2013), 1306.6327, See also <http://www.linearcollider.org/ILC/TDR> . The full list of contributing institutes is inside the Report. %%CITATION = arXiv : 1306.6327; %%
- [5] Ilc photo gallery, <http://ilcgallery.com/>.
- [6] **KEK International Working Group**, K. Desch *et al.*, Recommendations on ILC Project Implementation, (2019).
- [7] H. Baer *et al.*, The International Linear Collider Technical Design Report - Volume 2: Physics, (2013), 1306.6352, See also <http://www.linearcollider.org/ILC/TDR> . The full list of signatories is inside the Report. %%CITATION = arXiv : 1306.6352; %%
- [8] ilcsoft, <https://github.com/iLCSOFT>.
- [9] lcfiplus/lcfiplus: Flavor tagging code for ilc detectors, <https://github.com/lcfiplus/LCFIPlus>.
- [10] M. Frank, F. Gaede, C. Grefe, and P. Mato, DD4hep: A detector description toolkit for high energy physics experiments, Journal of Physics: Conference Series **513**, 022010 (2014).
- [11] M. Frank, F. Gaede, N. Nikiforou, M. Petric, and A. Sailer, DDG4 a simulation framework based on the DD4hep detector description toolkit, Journal of Physics: Conference Series **664**, 072017 (2015).
- [12] F. Gaede, Marlin and lccd—software tools for the ilc, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **559**, 177 (2006), Proceedings of the X International Workshop on Advanced Computing and Analysis Techniques in Physics Research.
- [13] S. Catani, Y. Dokshitzer, M. Olsson, G. Turnock, and B. Webber, New clustering

- algorithm for multijet cross sections in e+e – annihilation, Physics Letters B **269**, 432 (1991).
- [14] 斎. 康毅, ゼロから作る Deep Learning: Python で学ぶディープラーニングの理論と実装 ゼロから作る Deep Learning (オライリー・ジャパン, 2016).
- [15] 斎. 康毅, ゼロから作る Deep Learning 2: 自然言語処理編ゼロから作る Deep Learning (オライリー・ジャパン, 2018).
- [16] S. Raschka, クイープ, and 真. 福島, Python 機械学習プログラミング : 達人データサイエンティストによる理論と実践 Impress top gear (インプレス, 2016).
- [17] J. McCarthy, M. Minsky, N. Rochester, and C. E. Shannon, A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955., AI Magazine **27**, 12 (2006).
- [18] V. VAPNIK, Pattern recognition using generalized portrait method, Automation and Remote Control **24**, 774 (1963).
- [19] B. E. BOSEN, A training algorithm for optimal margin classifiers, Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, United States (1992-7) (1992).
- [20] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review **65**, 386 (1958).
- [21] G. CYBENKO, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals and Systems **2**, 303 (1989).
- [22] Tensorflow, <https://www.tensorflow.org/?hl=ja>.
- [23] Home - keras documentation, <https://keras.io/ja/>.
- [24] Pytorch, <https://pytorch.org/>.
- [25] Caffe — deep learning framework, <http://caffe.berkeleyvision.org/>.
- [26] V. Nair and G. E. Hinton, Rectified linear units improve restricted boltzmann machines., in *ICML*, edited by J. F. Ürnkranz and T. Joachims, pp. 807–814, Omnipress, 2010.
- [27] B. Widrow and M. E. Hoff, Adaptive switching circuits, 1960 IRE WESCON Convention Record , 96 (1960), Reprinted in *Neurocomputing* MIT Press, 1988 .
- [28] T. Tieleman and G. Hinton, Lecture 6.5—rmsProp: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural Networks for Machine Learning, 2012.
- [29] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, 2017, 1412.6980.
- [30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning representations by back-propagating errors, Nature **323**, 533 (1986).
- [31] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with

- neural networks, *Science* **313**, 504 (2006).
- [32] I. J. Goodfellow *et al.*, Generative adversarial networks, 2014, 1406.2661.
 - [33] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation* **9**, 1735 (1997), <https://doi.org/10.1162/neco.1997.9.8.1735>.
 - [34] K. Cho *et al.*, Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014, 1406.1078.
 - [35] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, 2016, 1409.0473.
 - [36] M.-T. Luong, H. Pham, and C. D. Manning, Effective approaches to attention-based neural machine translation, 2015, 1508.04025.
 - [37] A. Vaswani *et al.*, Attention is all you need, 2017, 1706.03762.
 - [38] Goto-k/vertexfinderwithdl, <https://github.com/Goto-K/VertexFinderwithDL>.
 - [39] W. Kilian, T. Ohl, and J. Reuter, Whizard—simulating multi-particle processes at lhc and ilc, *The European Physical Journal C* **71** (2011).
 - [40] T. Suehara and T. Tanabe, Lcfiplus: A framework for jet analysis in linear collider studies, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **808**, 109 – 116 (2016).
 - [41] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015, 1502.03167.
 - [42] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research* **9**, 2579 (2008).