Transformer を多層にする際の勾配消失問題と解決法について

高瀬翔 [1]　清野舜 [2,3]　小林颯介 [4,3]　鈴木潤 [3,2]
[1] 東京工業大学　[2] 理化学研究所　[3] 東北大学　[4]Preferred Networks
sho.takase@nlp.c.titech.ac.jp　shun.kiyono@riken.jp
sosk@preferred.jp　jun.suzuki@tohoku.ac.jp

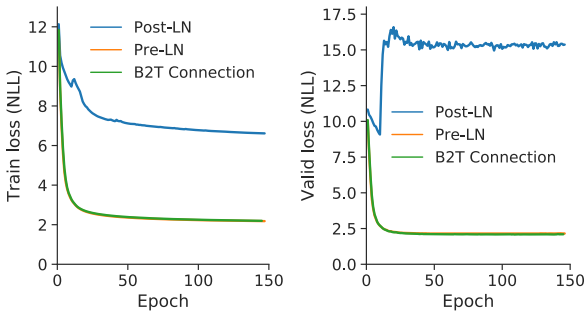
概要

Transformer は機械翻訳や要約のような系列変換タスクをはじめ，様々なタスクに用いられている．他のニューラルネットワークと同様に，Transformer も性能向上のためには層の数を増やす戦略が取られるが，例えば 18 層以上のように多層化する場合には学習の安定性のために Layer Normalization（LN）の位置を変えた構造を使うことが主流になっている．本研究では，オリジナルの構造での多層化における学習の不安定性は LN による勾配消失であることを示し，Residual Connection を追加するだけで多層化が可能になることを示す．また，多層化において主流の LN の位置を変えた構造は性能が低いことを実験的に示し，提案手法により性能に悪影響を与えることなく多層化できることを示す．

1　はじめに

多層ニューラルネットワークの学習において，勾配の爆発や消失を防ぐために Batch Normalization [1] や Residual Connection [2] が考案されてきた．Batch Normalization は適用位置を変化させたときの挙動も議論されており，性能の高い多層のネットワークを構築するためには構造にも気を配る必要がある [3]．

機械翻訳や要約のような系列変換タスクをはじめ，自然言語処理の様々なタスクに用いられている Transformer では Layer Normalization（LN）[4] が採用されている [5]．Transformer についても LN の適用位置に関しての議論がある．最初に提案された構造である Post-LN は多層にすると勾配消失問題で学習が不安定になり，サブレイヤの入力に LN を適用する Pre-LN と呼ばれる構造（各構造を図示した図 2 も参照のこと）の方が学習が安定していると示されている [6, 7]．例えば，18 層 Transformer エンコーダ・デコーダについて，機械翻訳で広く使われている WMT 英-独の訓練および開発データでの損失値を図



(a) Train loss　(b) Valid loss

図 1　18 層 Transformer エンコーダ・デコーダの訓練データ，開発データの損失値．

1 の (a)，(b) に示した．この図から Post-LN は損失値が高いままで，学習を続けても性能が改善されていないことが分かる．このため，多層の Transformer では Pre-LN を採用した研究が多い [8, 9]．

Pre-LN と Post-LN との議論では学習の安定性が俎上に上がることがほとんどであり，性能差に言及した研究は少ない．しかしながら，Liu らは機械翻訳において，学習が成功した場合（例えば 6 層のように多層でない場合）には Post-LN が Pre-LN よりも高い性能を達成したことを報告している [10]．この結果は学習の不安定性に目をつぶれば，Post-LN は Pre-LN よりも優れていることを示唆している．

そこで本研究では，Post-LN の学習の安定性向上に取り組む．特に，本研究では図 1 のような，学習を続けても性能が改善されない現象を学習の不安定性とし，これの改善に取り組む．まず，Post-LN を多層にした際に学習が不安定になる問題は LN による勾配消失が原因であることを示す．加えて，Post-LN と Pre-LN との性能差がどこに起因するかを調査する．上記の観測に基づき，Post-LN に Residual Connection を追加するだけで，学習パラメータの追加や計算コストを増加させることなく，高い性能を維持したまま多層での学習を安定させることが可能になると示す．

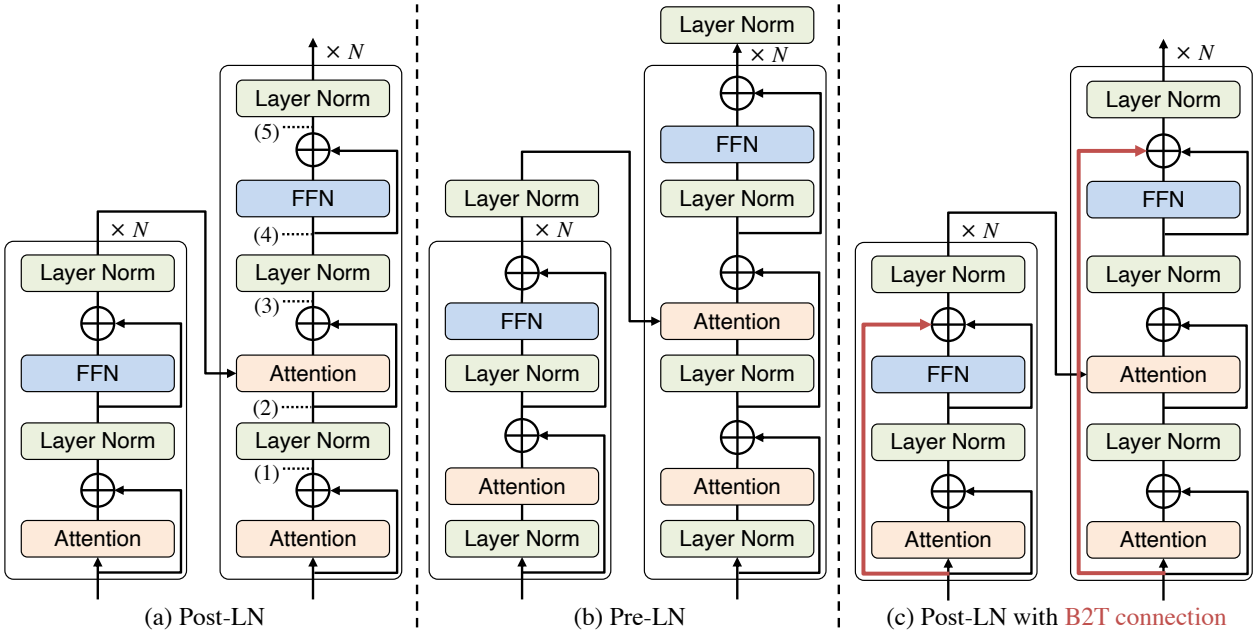


図 2　Transformer を用いたエンコーダ・デコーダの構造．(a) は Post-LN，(b) は Pre-LN，(c) は Post-LN に提案手法を組み合わせたものである．

2　Post-LN と Pre-LN

本節では Post-LN と Pre-LN を概説する．オリジナルの Transformer [5] では各 Residual Connection の後に LN を置く，Post-LN となっている．サブレイヤへの入力を x，フィードフォワードネットやマルチヘッドアテンションのようなサブレイヤを $\mathcal{F}(\cdot)$ とすると，Post-LN は次のようになる：

$$\mathrm{PostLN}(x) = \mathrm{LN}(x + \mathcal{F}(x)). \tag{1}$$

一方，Pre-LN は各サブレイヤの前に LN を置く：

$$\mathrm{PreLN}(x) = x + \mathcal{F}(\mathrm{LN}(x)). \tag{2}$$

図 2 の (a) と (b) はそれぞれ Post-LN と Pre-LN を図示したものである．

3　Transformer における勾配

Liu らが報告しているように，Post-LN では勾配消失が発生する [10]．図 3 は機械翻訳で広く使われている WMT 英-独の訓練データにおける，18 層 Transformer エンコーダ・デコーダの (a) エンコーダ側と (b) デコーダ側の各層の勾配のノルムを示したものである．この図における勾配のノルムは対数目盛りとしてあるため，デコーダ側において Post-LN の勾配は層が浅くなるにつれ指数的に減少していることが分かる．すなわち，Post-LN のデコーダ側では勾配消失が発生しており，この勾配消失が図 1 に

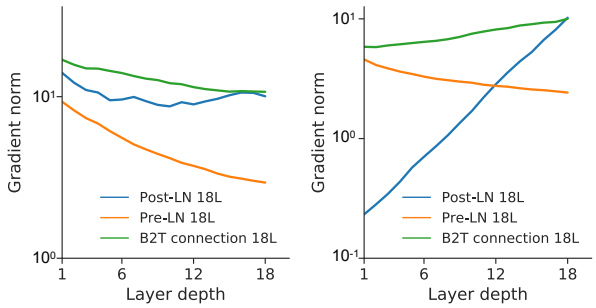


図 3　Transformer エンコーダ・デコーダの勾配のノルム．

示したように多層 Post-LN の学習を不安定にしていると考えられる．

勾配消失の詳細な原因を知るため，図 2(a) の (1) - (5) における勾配のノルムを調査した．図 4 は 18 層目の (1) - (5) における勾配のノルムを示している．この図から，(4) から (3)，(2) から (1) で勾配が大きく減衰していることが分かる．この勾配が大きく減衰している点は LN の位置と合致しており，LN が勾配消失の原因であると推測される．

さらに Post-LN と Pre-LN との勾配のながれの違いを探るために，式 (1) と (2) の微分値を計算する．各微分値は次のようになる：

$$\frac{\partial \mathrm{PostLN}(x)}{\partial x} = \frac{\partial \mathrm{LN}(x + \mathcal{F}(x))}{\partial (x + \mathcal{F}(x))} \left(1 + \frac{\partial \mathcal{F}(x)}{\partial x}\right), \tag{3}$$

$$\frac{\partial \mathrm{PreLN}(x)}{\partial x} = 1 + \frac{\partial \mathcal{F}(\mathrm{LN}(x))}{\partial \mathrm{LN}(x)} \frac{\partial \mathrm{LN}(x)}{\partial x}. \tag{4}$$

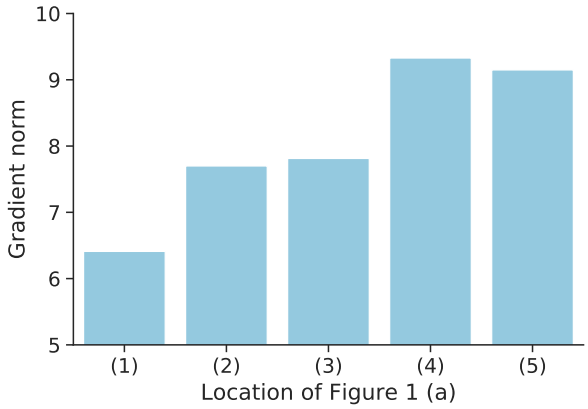


図 4　WMT 英-独で学習した際の 18 層の Post-LN における 18 層目のデコーダの各位置における勾配のノルム．

式 (3) のように，Post-LN の微分は LN の微分およびサブレイヤ $\mathcal{F}$ と Residual Connection の微分の積となる．これに対し，Pre-LN の微分は Residual Connection の微分が LN の微分と独立の項となっている．LN の微分が勾配を大きく減衰させたとしても，この Residual Connection の微分が勾配を維持するため，Pre-LN では勾配消失が発生していないと考えられる．

4　各層による変換

Post-LN では LN によって勾配が減衰し，これによって低層では勾配消失が発生しうる．このため多層の Post-LN の学習は困難であるが，実験の節にあるように，学習に成功した場合は Post-LN は Pre-LN よりも高い性能を達成する．この性能差は各層における変換の程度に依存すると考えられる．

図 5 は Transformer エンコーダ・デコーダについて，WMT データセット内のいくつかの系列を入力した際の，各層の出力間のコサイン類似度を平均し図示したものである．この図において，Pre-LN の左下の類似度は Post-LN の左下の値よりも高いことが分かる．すなわち，Post-LN と比較して，Pre-LN は最初の層と最終層の出力との類似度が高い．式 (2) のとおり，Pre-LN では入力 x が Residual Connection によってサブレイヤの $\mathcal{F}(\cdot)$ を回避しており，その結果，入力 x が最終層の出力に直接足し込まれる．これにより勾配消失は防ぐことができるが，各層からの出力の類似度が高くなってしまう．言い換えれば，Pre-LN は入力を変換する作用が Post-LN よりも小さく，これが Pre-LN が Post-LN よりも性能が低い原因であると考えられる．

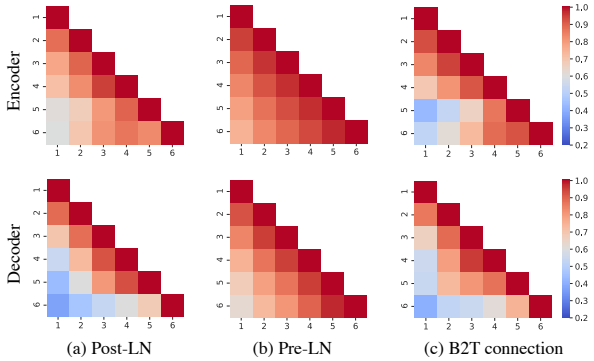


図 5　各層からの出力間のコサイン類似度．

5　提案手法：B2T Connection

本節では Post-LN の高い性能を維持したまま多層化する手法を提案する．多層化による勾配消失を防ぐためには式 (4) にあるような，微分した際に勾配を維持する項が必要である．これを満たすため，各層における最後の LN 以外の LN を回避する Residual Connection を追加する[1)]．図 2(c) の赤い矢印で示したように，提案手法は各層への入力をフィードフォワードネットの結果へと結合する．本研究ではこれを **Bottom-to-Top (B2T)** Connection と呼ぶ．詳細には次式で表現される：

$$x_{\text{inp}} + x_{\text{ffn}} + \text{FFN}(x_{\text{ffn}}), \tag{5}$$

ここで x_{inp} は層への入力，x_{ffn} はフィードフォワードネットへの入力，FFN(·) はフィードフォワードネットである．つまり，x_{inp} はセルフおよびエンコーダ・デコーダ間のマルチヘッドアテンション後の LN を回避し，勾配の維持に貢献する．実際，図 3(b) は B2T Connection が 18 層のエンコーダ・デコーダにおいて勾配消失を防いでいることを示しており，図 1 に示すように学習も安定している．また，図 5(c) は B2T Connection は出力間の類似度の傾向が Post-LN と似ており，層ごとの変換を Pre-LN よりも行えることを示している．

6　機械翻訳での実験

系列変換タスクとして，機械翻訳タスクでの実験を行う．要約，自動音声認識タスクについては付録に記した．

1)　各層の最後の LN も含め全ての LN を回避する Residual Connection も試してみたが，性能が大きく低下した．全ての LN を回避した場合には入力が出力に直接つながることになり，節 4 で議論したように各層での変換作用が小さくなり，Post-LN の利点が消失するからであると考えられる．

6.1 データセット

機械翻訳タスクはオリジナルの Transformer をはじめ，Transformer エンコーダ・デコーダの性能を調べるために広く使われているタスクである [5, 11, 6, 7, 10]．本研究では広く使われている，WMT 英-独の 450 万文対を含む訓練データを用いる [5, 11]．既存研究と同様，語彙の構築には BPE [12] を用いる．性能評価には newstest2013-2016 を用いる．

6.2 比較手法

実験では **Post-LN**，**Pre-LN**，Post-LN に提案手法である B2T Connection を組み合わせたもの（**B2T Connection**）の比較を行う．層の数は広く使われている設定および多層の設定として 6 層と 18 層を採用する．中間層の次元数については Vaswani ら [5] の base 設定と同一の値とする．上記に加えて，多層 Post-LN の学習を安定化する，下記の既存手法と比較を行う．なお，既存手法は学習を安定化させるために追加のパラメータおよび計算を要する．

DLCL　Wang らは下層の出力の重み付き和を次の層への入力とする，Dynamic Linear Combination of Layers（DLCL）を提案した [6]．各層内に Residual Connection を追加する提案手法とは異なり，DLCL は各層間を接続する経路を設ける．各出力への重みはパラメータであり学習によって適切な値を得る．

Admin　多層の Transformer の学習を安定化させるため，Liu らは Adaptive Model Initialization（Admin）を提案した [10]．これは，各層の出力の分散を元に初期化したパラメータを導入することで，学習初期での安定性を高めている．追加したパラメータを初期化するためには各層の出力が必要であるため，実際に学習を開始する前に，複数回の前向き計算を要する．すなわち，本手法はパラメータを追加していることに加え，必要とする計算量も増加している．

6.3 結果

表 1 に newstest2013-2016 における各手法の SacreBLEU [14][2] で計算した BLEU スコアとその平均値を示した．表 1 の上部は 6 層，下部は 18 層の結果を示している．表 1 の上部から，Pre-LN の BLEU スコアは他の手法と比べて低いことが分かる．すなわち，構造として Post-LN を用いた方が Pre-LN よりも高い性能を達成している．

2） SacreBLEU の signature は BLEU+nrefs:1+case:mixed+eff:no+tok:13a+smooth:exp+version:2.0.0.

表 1　WMT newstest2013-2016 における各手法の BLEU スコアとその平均値．

手法	2013	2014	2015	2016	平均
層の数が 6 の場合					
Post-LN	26.11	27.13	29.70	34.40	29.34
Pre-LN	25.63	26.27	29.07	33.84	28.70
DLCL [6]	26.11	**27.37**	**29.71**	34.26	**29.36**
Admin [10]	26.26	27.14	29.61	34.12	29.28
B2T connection	**26.31**	26.84	29.48	**34.73**	29.34
層の数が 18 の場合					
Post-LN	学習失敗				N/A
Pre-LN	26.28	27.36	29.74	34.16	29.38
DLCL [6]	**26.59**	27.97	30.24	33.98	29.70
Admin [10]	26.48	27.99	**30.35**	33.88	29.67
B2T connection	26.53	**28.41**	30.21	**34.29**	**29.86**

表 1 の下部では，18 層，すなわち，多層にした際に素朴な Post-LN は学習が失敗していることを示している．具体的には図 1 に示したように，学習を続けても性能が改善しなかった．これに対し，提案手法（B2T Connection）は学習に成功しており，また，Pre-LN よりも高い性能を達成している．これらの結果から，Post-LN は学習が成功すれば Pre-LN よりも高い性能を達成すること，提案手法は Post-LN の利点を維持しつつ多層の学習の安定性を向上させていることが分かる．

従来手法との比較では提案手法が同等以上の性能を達成しており，追加での学習パラメータや計算コストを要求しない点も鑑みると提案手法が優れていると言える．付録に要約タスクでの比較も示した．

7 おわりに

本研究では Post-LN の学習の安定性向上に取り組んだ．Post-LN を多層にした際は LN による勾配消失のために学習が不安定になることを示し，Pre-LN と Post-LN の性能差が各層での変換の差に起因する可能性を示した．Post-LN の利点を維持したまま多層での学習を安定させる手法として，B2T Connection という，層内の LN を回避する Residual Connection を提案した．系列変換タスクでの実験を通して以下の 3 点を明らかにした．1. Post-LN は学習が成功すれば Pre-LN よりも高い性能を達成可能である．2. 提案手法により多層 Post-LN の学習が安定する．3. 提案手法は Post-LN の利点を維持したまま学習を安定させることで，多層にした際に Pre-LN よりも高い性能を達成できる．

謝辞

本研究は JSPS 科研費 JP21K17800 および JST, ACT-X，JPMJAX200I の助成を受けたものです．また，本研究の一部（基礎研究）は JST ムーンショット JPMJMS2011 の助成を受けたものです．

参考文献

[1] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In **Proceedings of ICML**, Vol. 37, pp. 448–456, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In **CVPR**, pp. 770–778, 2016.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, **ECCV**, pp. 630–645, 2016.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS**, pp. 5998–6008. 2017.

[6] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In **Proceedings of ACL**, pp. 1810–1822, 2019.

[7] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In **Proceedings of ICML**, pp. 10524–10533, 2020.

[8] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In **Proceedings of ICLR**, 2019.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In **NeurIPS**, pp. 1877–1901, 2020.

[10] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In **Proceedings of EMNLP**, pp. 5747–5763, 2020.

[11] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In **Proceedings of WMT**, pp. 1–9, 2018.

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **Proceedings of ACL**, pp. 1715–1725, 2016.

[13] Sho Takase and Shun Kiyono. Rethinking perturbations in encoder-decoders for fast training. In **Proceedings of NAACL-HLT**, pp. 5767–5780, 2021.

[14] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of WMT**, pp. 186–191, 2018.

[15] Alexander M. Rush, Sumit Chopra, and Jason Weston. A Neural Attention Model for Abstractive Sentence Summarization. In **Proceedings of EMNLP**, pp. 379–389, 2015.

[16] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. Annotated Gigaword. In **Proceedings of AKBC-WEKEX**, pp. 95–100, 2012.

[17] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In **NeurIPS**, pp. 9054–9065, 2019.

[18] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In **Proceedings of WMT**, pp. 1–61, 2019.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In **ICASSP**, pp. 5206–5210, 2015.

[20] Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. Fairseq S2T: Fast speech-to-text modeling with fairseq. In **Proceedings of AACL-IJCNLP**, pp. 33–39, 2020.

[21] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of EMNLP**, pp. 66–71, 2018.

表 2　見出し文生成 [15] における各手法の ROUGE-1，2，L の F-1 スコア（それぞれ R-1，R-2，R-L とする）．

手法	R-1	R-2	R-L
層の数が 6 の場合			
Post-LN	**38.57**	**19.37**	**35.79**
Pre-LN	38.27	19.29	35.39
DLCL [6]	38.13	18.49	35.00
Admin [10]	37.96	18.93	35.05
B2T Connection	38.43	**19.37**	35.72
層の数が 18 の場合			
Post-LN	学習失敗		
Pre-LN	38.97	19.94	35.99
DLCL [6]	38.25	19.44	35.57
Admin [10]	39.10	20.08	36.30
B2T Connection	**39.61**	**20.28**	**36.66**

表 3　LibriSpeech における各手法の単語誤り率．

Method	dev-clean	dev-other	test-clean	test-other
エンコーダ側の層数が 6 の場合				
Post-LN	3.78	**8.76**	4.19	**8.74**
Pre-LN	3.89	9.69	4.22	9.65
B2T Connection	**3.69**	8.97	**3.86**	8.94
エンコーダ側の層数が 12 の場合				
Post-LN	学習失敗			
Pre-LN	**3.21**	7.91	3.49	8.22
B2T Connection	3.26	**7.74**	**3.48**	**7.68**

A　要約・自動音声認識での実験

A.1　データセット

生成型要約タスクは機械翻訳に並んで代表的な系列変換タスクある．本研究ではニュースの 1 文目を入力とし，見出し文を生成する見出し文生成に取り組む．Rush ら [15] によって Annotated English Gigaword [16] から構築されたデータセットを用いる．本データセットは 380 万文対の訓練データと 1951 文対のテストデータからなる．また，既存研究 [13] にならい，REALNEWS [17] と NewsCrawl [18] から構築した 1300 万文対の追加の訓練データを用いる．機械翻訳と同様，BPE で語彙の構築を行う．

加えて，言語以外のモダリティとして自動音声認識での実験を行う．英語の音声認識で広く使われているデータセットとして LibriSpeech [19] を用いる．既存研究 [20] にならい，利用可能な全ての訓練データを学習に用い，開発，テストセットについて Clean と Other の 2 種を用いる．デコーダ側の語彙は SentencePiece [21] で構築する．

A.2　比較手法

要約においては機械翻訳タスクでの実験と同様，**Post-LN**，**Pre-LN**，**B2T connection**，**DLCL**，**Admin** との比較を行う．層の数は 6 層および 18 層とする．

自動音声認識では **Post-LN**，**Pre-LN**，**B2T connection** の比較を行う．なお，従来研究 [20] はエンコーダ側を多層にすることで性能が向上することを示しているため，本実験ではエンコーダ側のみ多層にし，デコーダ側の層数は 6 層で固定とする．エンコーダ側の層数は 6 層および 12 層とする．

A.3　結果

表 2 に見出し文生成のテストデータにおける ROUGE-1，2，L の F-1 スコアを示す．層の数が 6 層の場合には Post-LN および提案手法は Pre-LN よりも高いスコアを達成している．18 層の場合には Post-LN は学習に失敗してしまっているが，提案手法は学習に成功しており，Pre-LN よりも高い性能を達成している．この結果から，見出し文生成においても，学習に成功すれば Post-LN は Pre-LN よりも高い性能を達成可能なこと，提案手法は Post-LN の利点を活かしたまま多層にした際の学習の安定性を向上させることが分かる．

既存研究である DLCL [6]，Admin [10] との比較では，6 層，18 層のどちらの場合でも提案手法が高い性能を達成している．機械翻訳タスクでの結果とあわせると，提案手法はこれら従来手法と同等以上の性能を達成すると言える．なお，提案手法には追加での学習パラメータや計算コストを要求しない点を再度強調したい．

表 3 に各手法の単語誤り率を示す．エンコーダの層数が 6 層の場合には，機械翻訳や要約タスクと同様，Post-LN および提案手法が良い性能を達成している．エンコーダの層数が 12 層の場合には素朴な Post-LN は学習に失敗してしまっているが，提案手法は学習に成功しており，dev-clean を除いて Pre-LN よりも高い性能を達成している．この結果も機械翻訳や要約タスクと同様，提案手法は多層にした際の学習の安定性を向上させること，Post-LN のように高い性能を達成可能であることを示している．