論文 / 著書情報
Article / Book Information

論題(和文)	受容野の自動最適化によるモードに適応的なTransformerの開発
Title(English)	Mode-Adaptive Transformer by Automatic Optimization of the Receptive Field
著者(和文)	浅倉 拓也, 井上中順, 横田 理央, 篠田 浩一
Authors(English)	Takuya Asakura, Nakamasa Inoue, Rio Yokota, Koichi Shinoda
出典(和文)	人工知能学会全国大会 (第37回)論文集, , ,
Citation(English)	Proceedings of the Annual Conference of JSAI, , ,
発行日 / Pub. date	2023, 6
Note	

受容野の自動最適化によるモードに適応的な Transformer の開発

Mode-Adaptive Transformer by Automatic Optimization of the Receptive Field

浅倉拓也　井上中順　横田理央　篠田浩一
Takuya Asakura　Nakamasa Inoue　Rio Yokota　Koichi Shinoda

東京工業大学 情報理工学院
School of Computing, Tokyo Institute of Technology

The Vision Transformer (ViT), which uses Attention instead of convolution for feature extraction, has demonstrated high performance in the field of image processing. This result shows that the Transformer can be used for both time-series and images, and is expected to be a versatile model that is independent of the mode of data. However, many of the studies derived from ViT have narrowed the receptive field for feature extraction, and their adaptability to time-series such as speech is compromised. In this paper, we propose a method to adaptively optimize the receptive fields for a given mode of data. We developed a model using the proposed method and conducted experiments on two types of data, images and speech, and found that the proposed method outperforms conventional methods for both. The visualization shows that the proposed method can acquire a suitable receptive field depending on the mode of the given data.

1. はじめに

近年の深層学習において，Transformer [Vaswani 17] が自然言語を始めとして，画像や音声など様々なデータを扱うタスクに対して従来の畳み込みニューラルネットワーク (CNN) や回帰型ニューラルネットワーク (RNN) を上回る精度を達成している．一般的に，Transformer はチャンネル方向の全結合層と Multi Head Attention (MHA) によって構成され，複数のトークンからなるシーケンスを入力データとして受け取る．MHA は各トークンを混合し特徴抽出を行う役割を担うため，Token-Mixing ともよばれる．多くの場合，各トークンには自然言語処理ではベクトル化された各単語を，画像処理では入力画像を分割した各パッチを，音声処理では入力音声を周波数解析して得られるスペクトログラムなどの各要素を割り当てる．Transformer はデータの種別に限らず高い性能を発揮することから，あらゆるデータに対して利用可能な汎用モデルとして期待を集めている．

Token-Mixing に MHA を用いる Transformer には，特に画像データに対して，過学習しやすいという課題があった．そのため，扱うデータの特性に適した Token-Mixing を導入した Transformer の派生モデルが数多く提案されている．一方，これらの応用を考えたとき，目的タスクやモデルの各層によってどのような Token-Mixing が最適かは不明瞭であり，それらを網羅的に検証することは計算コストの面で課題があった．これを解決するため，条件に応じて最適な特性の Token-Mixing を探索可能な手法，Adaptive Fully Connected Layer (AdaFC) が提案されている [浅倉 22]．Token と Channel の平面からなる入力特徴マップが与えられるとき，AdaFC はそれと同サイズの重み行列を用意し，図 1 のように，写像先のトークンの位置ごとに空間的にシフトしながら重み行列を適用する．この重み行列により参照するトークンの数や範囲が決定されるため，これは Token-Mixing における受容野を意味しており，条件に応じて最適化することで任意の特性の Token-Mixing を柔軟に表現することができる．この柔軟性によって，AdaFC は入力データの種別や層の深さに応じて適した特性による特徴抽出を行うことができ，画像分類タスクにおいて人手で設計した Token-Mixing を使用した場合よりも高い精度を発揮した．しかし，AdaFC の重み行列が入力特徴マップから参照するトークンの位置は，写像先のトークンの位置によって異なる場合がある．そのため，全ての写像先に対して一律なパターンによる特徴抽出を行うことはできない．また，AdaFC の有効性は画像データを用いる実験のみにおいて検証されており，音声など異なるモードを持つデータに対する性能は未検証である．

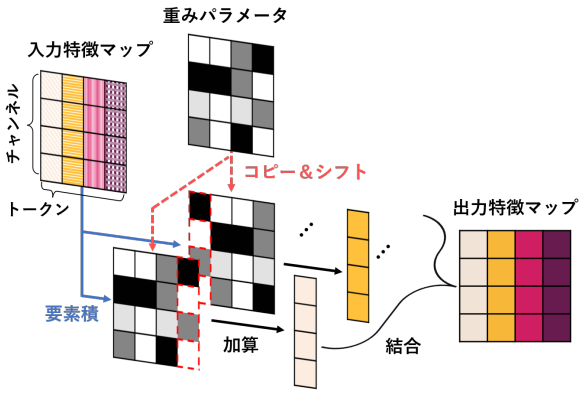


図 1: AdaFC による Token-Mixing．点線で囲った列のように，重み行列の各列が指すトークンの位置が写像先のトークンによって変化する場合がある．

本研究では，上述した AdaFC に関する問題を解決した手法，AdaFCv2 を提案する．AdaFCv2 では入力特徴マップのトークン数よりも大きなサイズの重み行列を用意することで，全ての写像先に対して一律なパターンによる特徴抽出を行う．Token-Mixing に AdaFCv2 を用いた Transformer ベースのモデルを作成し，ImageNet [Russakovsky 15] による画像分類実験と Speech Commands v2 [Warden 18] による音声分類実験を行い，従来の AdaFC と正答率を比較することで性能を評価した．学習後の重み行列を可視化すると，画像データに対しては一般的な画像処理フィルタに似た形状に収束した一方

連絡先: 浅倉拓也，東京工業大学 情報理工学院，
asakura@ks.c.titech.ac.jp

で，音声データに対しては長期的な依存関係を抽出する形状に収束しており，提案手法である AdaFCv2 がデータのモードに応じて適した受容野を獲得可能であることが示されている．

2. 提案手法

2.1 Adaptive Patch Sampling

AdaFCv2 は重みパラメータを用いた任意のトークンのサンプリングと，それらに対する全結合層の適用の二段階で表される．トークンのサンプリングにおいて，本研究では図 2 に示される Adaptive Patch Sampling (APS) を提案する．トークン数 P，チャンネル数 C の入力特徴マップ

$$\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{P-1}] \in \mathbb{R}^{P\times C}, \boldsymbol{x}_p \in \mathbb{R}^C \tag{1}$$

が与えられるとき，APS では学習可能な重みパラメータ

$$\begin{aligned}\boldsymbol{\Gamma} =& [\boldsymbol{\gamma}_{-(P-1)}, \boldsymbol{\gamma}_{-(P-2)}, \cdots, \\ & \boldsymbol{\gamma}_0, \ldots, \boldsymbol{\gamma}_{P-2}, \boldsymbol{\gamma}_{P-1}] \in \mathbb{R}^{(2P-1)\times C}\end{aligned} \tag{2}$$

を用意し，サンプリングに利用する．ここで，$\boldsymbol{\gamma}_p \in \mathbb{R}^C$ は Token-Mixing において p 離れたトークンを参照する際の重みを表す．すなわち，重みパラメータ $\boldsymbol{\Gamma}$ は特徴抽出時にサンプルするトークンの距離を表現している．この $\boldsymbol{\Gamma}$ を用いて，サンプリング後の特徴マップ

$$\boldsymbol{S} = [\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{P-1}] \in \mathbb{R}^{P\times C}, \boldsymbol{s}_p \in \mathbb{R}^C \tag{3}$$

を生成する．$\boldsymbol{s}_p$ を求める際に入力特徴マップ $\boldsymbol{X}$ からサンプル可能なトークンの距離を考えると，$p=0$ のとき 0 から $P-1$，$p=1$ のとき -1 から $P-2$，$p=P-1$ のとき $-(P-1)$ から 0 だけ離れたトークンをサンプルすることができる．よって，$\boldsymbol{S}$ の計算は写像先である $\boldsymbol{s}_p$ に対応する P 個の要素を重みパラメータ $\boldsymbol{\Gamma}$ から取り出して

$$\boldsymbol{s}_{p'} = \sum_{p=0}^{P-1} \boldsymbol{x}_p \odot \boldsymbol{\gamma}_{P-1-p-p'} \tag{4}$$

として計算される．ここで，$\odot$ は行列同士の要素積である．また，式 4 は全てのトークンに渡って並列化可能である．最後に，サンプリングによって得られた $\boldsymbol{S}$ に全結合層による線形変換を適用する．

上記の処理は全て微分可能であるため，重みパラメータ $\boldsymbol{\Gamma}$ は一般的なニューラルネットワークと同じく勾配法により最適化可能である．重みパラメータ $\boldsymbol{\Gamma}$ が最適化されるに従って，Token-Mixing における受容野がデータ種別や層の深さなどの条件に適したものに収束すると期待される．また，APS では任意の受容野の Token-Mixing を表現可能なため，AS-MLP [Lian 21] におけるトークンのシフトや，[Chen 22b] によって検証された様々なサンプリングのパターンを表現することもできる．特に，$\boldsymbol{\Gamma}$ において $\boldsymbol{\gamma}_0$ のみが $\mathbf{1}$，それ以外の全ての要素が 0 である場合，$\boldsymbol{S} = \boldsymbol{X}$ である．

2.2 モデル構成

精度の比較を容易にするため，本研究では従来の AdaFC で使用されたものと同様の構造を持つモデルを用い，Token-Mixing に AdaFCv2 を使用する．このモデルは 4 つのステージからなり，ステージ s では図 3 に示すブロックを N_s 回繰り返す．1 ステージ目に与えられた画像や音声などのデータは，カーネルサイズが 7×7 でストライドが 4 の畳み込みによりベクトル化される．2 ステージ目以降では，学習の効率化のためステージの始めにカーネルサイズ 3×3，ストライド 2 の畳み込みを適用し特徴マップを圧縮する．

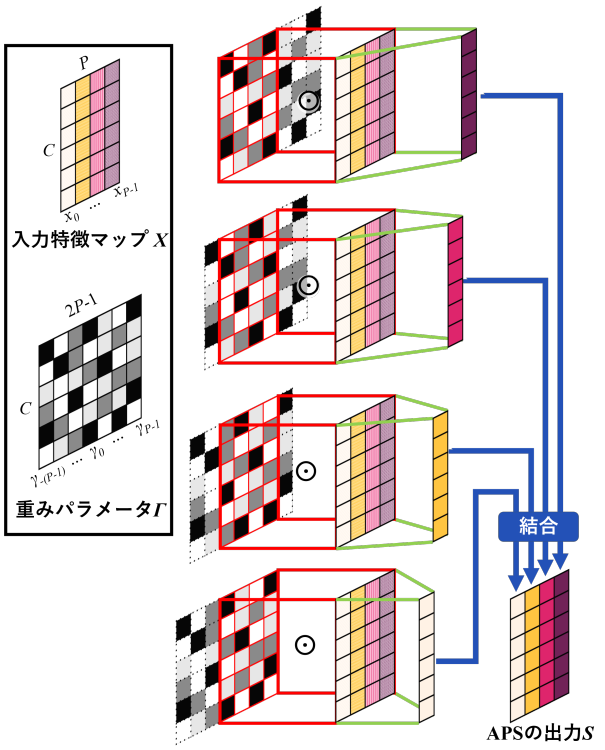


図 2: APS によるトークンのサンプリング処理．

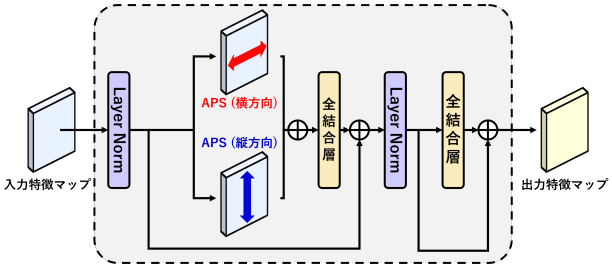


図 3: Token-Mixing に AdaFCv2 を用いたブロック．

画像や音声のスペクトログラムは二次元データとして表されるため，各ブロックにおいて APS は縦方向および横方向の両者に対し提供される．また，APS における重みパラメータ $\boldsymbol{\Gamma}$ は層ごとに用意して最適化する．従来の AdaFC では縦方向と横方向の特徴抽出を直列に繋げていたが，本研究の AdaFCv2 ではそれぞれの方向に対する APS の出力特徴マップを加算したのち全結合層を適用する．

表 1: 各モデルにおけるステージごとのブロックの繰り返し数．

モデル名	N_1	N_2	N_3	N_4
モデル-S	1	1	3	1
モデル-M	2	2	6	2
モデル-L	3	3	12	3

3. 実験内容

Token-Mixing として AdaFCv2 を採用したモデルを用いて，画像分類と音声分類のモードが異なる二種類のタスクに対して実験を行う．実験ではステージ s におけるブロックの繰り返し数 N_s を変化させ，サイズが異なる 3 つのモデルを作成した．モデルごとの N_s を表 1 に示す．また，ステージ s におけるチャンネル数 C_s は一律に $[C_1, C_2, C_3, C_4] = [64, 128, 320, 512]$ とした．最適化アルゴリズムには AdamW を使用した．APS の重みパラメータ $\boldsymbol{\Gamma}$ は平均 0，分散 1 の正規分布に従って初期化し実験を行った．

3.1 画像分類実験

画像分類実験として，ImageNet-1K による 1000 クラス分類タスクを学習させた．各画像は $224 \times 224 \times 3$ の RGB 画像としてモデルに与えた．バッチサイズは 512 とした．正則化アルゴリズムおよびハイパーパラメータは [浅倉 22] と同様のものを使用し，300Epoch に渡って学習させた．学習率は WarmUp 時に 1×10^{-6} から 1×10^{-3} まで 5 Epoch かけて増加させ，それ以降は 1×10^{-3} から 1×10^{-6} まで Cosine 関数に従って減少させた．

3.2 音声分類実験

音声分類実験として，Speech Commands v2 (SPCv2) による 35 クラス分類タスクを学習させた．各音声は 128 次元の対数メルフィルタバンク特徴量 (log Mel filterbank features, fbank) に変換する．窓幅 25ms の Hanning Window を 10ms ごとに適用して fbank を計算することで，$128 \times 128 \times 1$ の特徴量としてモデルに与える．バッチサイズは 128 とした．正則化として 0.05 の Weight Decay，$\alpha = 1$ の MixUP および SpecAugment を適用した．学習率は WarmUp 時に 1×10^{-5} から 1×10^{-3} まで 3 Epoch に渡って増加させ，それ以降は 1×10^{-3} から 1×10^{-5} まで Cosine 関数に従って減少させた．この実験では，モデル-L についてのみ実験を行った．

4. 結果と考察

4.1 画像分類タスクでの精度

各モデルのパラメータ数と画像分類タスクにおける学習後の正答率を表 2 に示す．従来手法である AdaFC と比較した場合，提案した AdaFCv2 の使用によりモデル-L における精度の向上が見られた．正答率が 81.1%から 81.8%に改善されたことで，従来は下回っていた CycleMLP-B2 の正答率 81.6%を超える精度を達成した．モデル-S およびモデル-L での実験では精度の変化は確認されなかった．これらのことから，AdaFCv2 が画像データに対して従来手法以上の精度を有していることが示されている．

4.2 音声分類タスクでの精度

音声分類タスクにおける実験結果を表 3 に示す．SPCv2 に対し，モデル-L は 98.03%の正答率を達成した．これは従来の CNN や Transformer を上回る精度である．また，事前学習済みの HTS-AT や AST-S と比較すると，モデル-L は事前学習無しでこれらと同等の精度を達成している．この結果から，AdaFCv2 は画像データだけでなく，音声データに対しても有効であることが示されている．

4.3 可視化

AdaFCv2 がどのような受容野を獲得したか確認するため，学習後のモデル-L から 1 層，2 層，12 層，24 (最終) 層の重みパラメータ $\boldsymbol{\Gamma}$ を取り出し，チャンネルで平均をとり可視化する．画像分類実験における結果を図 4 に，音声分類実験における結果を図 5 に示す．各画像において，中央の値が大きいほど近傍のトークンを，端の値が大きいほど離れたトークンをサンプルしていることを意味している．従来の AdaFC と同じく，画像に対する受容野は浅い層で狭く，深い層で広く学習されている．加えて，AdaFCv2 では全ての層においてラプラシアンのような特性が獲得されていることが確認された．また，画像の縦方向と横方向で学習結果の違いは見られなかった．

音声に対しては画像を学習した場合と全く異なる結果が確認された．1 層目では γ_0 にピークが立っている他，2 層目の時点で離れたトークンをサンプルしている．また，画像の場合と異なり，音声データを学習した場合は周波数軸と時間軸で異なる結果が得られた．これらのことから，AdaFCv2 ではデータの種別による違いだけでなく，そのデータにおける各次元が意味する情報の違いにも対応しながら，それらに適した特性による特徴抽出を行うことができる．

表 2: ImageNet 分類タスクにおける正答率.

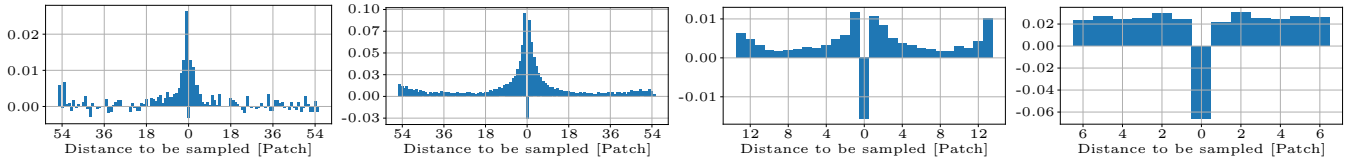
モデル	パラメータ数	正答率 (%)
DeiT-S [Touvron 21]	22M	79.8
gMLP-Ti [Liu 21]	6M	72.3
gMLP-S [Liu 21]	20M	79.6
AS-MLP-T [Lian 21]	28M	81.3
CycleMLP-B1 [Chen 22b]	15M	79.1
CycleMLP-B2 [Chen 22b]	27M	81.6
モデル-S [浅倉 22]	9M	76.4
モデル-M [浅倉 22]	13M	79.4
モデル-L [浅倉 22]	24M	81.1
モデル-S (AdaFCv2)	8M	76.4
モデル-M (AdaFCv2)	13M	79.4
モデル-L (AdaFCv2)	24M	**81.8**

表 3: Speech Commands v2 分類タスクにおける正答率.

モデル	事前学習	正答率 (%)
Branchformer [Peng 22]	-	97.3
MatchboxNet [Touvron 21]	-	97.37
Conformer [Gulati 20]	-	97.5
KWT-2 [Berg 21]	-	97.74
HTS-AT [Chen 22a]	音声データ	98.0
AST-S [Gong 21]	画像データ	**98.11**
モデル-L	-	**98.03**

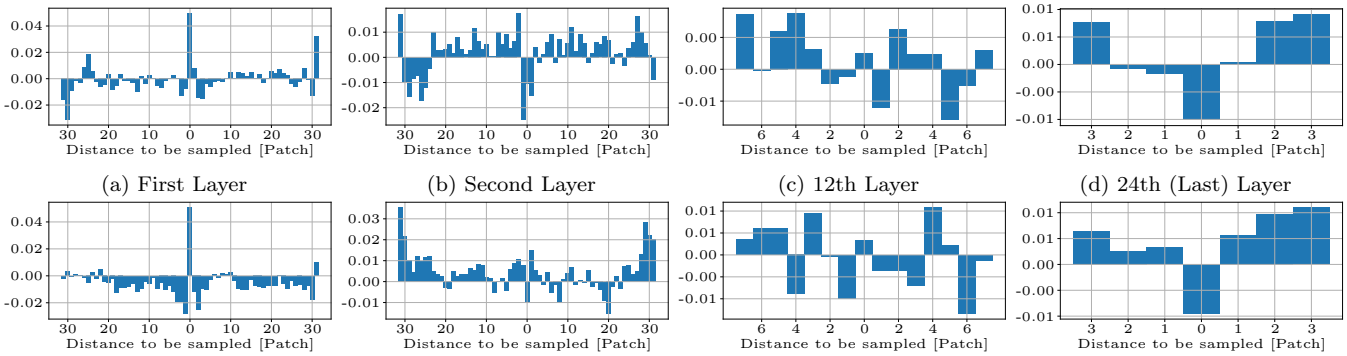
5. まとめと今後の課題

本研究では従来の AdaFC における重み行列が一律な位置のトークンをサンプルできない場合がある問題を取り上げ，それを改善した AdaFCv2 を提案した．AdaFCv2 ではトークン数 P の入力特徴マップに対して $2P-1$ の要素を持つ重みパラメータを提供することで，全てのケースで一律なパターンによる特徴抽出を可能とした．AdaFCv2 を Token-Mixing として採用したモデルを定義し，画像と音声の 2 種類のデータについて実験を行った結果，モデルの構造を変えることなく両者に対して従来の AdaFC より高い精度を達成することができた．



(a) First Layer (b) Second Layer (c) 12th Layer (d) 24th (Last) Layer

図 4: 画像を学習した重みパラメータ $\boldsymbol{\Gamma}$ の可視化．ここでは横方向に対する $\boldsymbol{\Gamma}$ のみを可視化している．



(a) First Layer (b) Second Layer (c) 12th Layer (d) 24th (Last) Layer

(e) First Layer (f) Second Layer (g) 12th Layer (h) 24th (Last) Layer

図 5: 音声を学習した重みパラメータ $\boldsymbol{\Gamma}$ の可視化．(a) から (d) は時間軸で，(e) から (h) は周波数軸で学習した結果である．

学習結果の可視化は画像と音声で異なる結果を示しており，提案した AdaFCv2 がデータのモードに対して適した受容野を獲得できることが示された．

課題として，本研究では画像単体または音声単体のみを扱う実験に留まったため，自然言語処理やマルチモーダルタスクに対する性能も評価していきたい．また，手書き文字と医療画像のような，モードが同じでドメインが異なるデータを学習した場合にどのような結果が得られるのかについても検証していきたい．

参考文献

[Berg 21] Berg, A., O'Connor, M., and Cruz, M. T.: Keyword Transformer: A Self-Attention Model for Keyword Spotting, in *Proc. Interspeech 2021* (2021)

[Chen 22a] Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., and Dubnov, S.: HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection, in *ICASSP* (2022)

[Chen 22b] Chen, S., Xie, E., GE, C., Chen, R., Liang, D., and Luo, P.: CycleMLP: A MLP-like Architecture for Dense Prediction, in *ICLR* (2022)

[Gong 21] Gong, Y., Chung, Y.-A., and Glass, J.: AST: Audio Spectrogram Transformer, in *Proc. Interspeech 2021* (2021)

[Gulati 20] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition, in *Proc. Interspeech 2020* (2020)

[Lian 21] Lian, D., Yu, Z., Sun, X., and Gao, S.: AS-MLP: An Axial Shifted MLP Architecture for Vision, *CoRR*, Vol. abs/2107.08391, (2021)

[Liu 21] Liu, H., Dai, Z., So, D., and Le, Q. V.: Pay Attention to MLPs, in *NeurIPS*, Curran Associates, Inc. (2021)

[Peng 22] Peng, Y., Dalmia, S., Lane, I., and Watanabe, S.: Branchformer: Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding, in *ICML* (2022)

[Russakovsky 15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252 (2015)

[Touvron 21] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H.: Training data-efficient image transformers & distillation through attention, in *ICML* (2021)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I.: Attention is All you Need, in *NeurIPS* (2017)

[Warden 18] Warden, P.: SpeechCommands: A Dataset for Limited-Vocabulary Speech Recognition, *CoRR*, Vol. abs/1804.03209, (2018)

[浅倉 22] 浅倉 拓也, 宇都 有昭, 篠田 浩一：Transformer における Token-Mixing の探索, 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 3J4OS3b04–3J4OS3b04 (2022)