Vol.2021-NL-251 No.12
Vol.2021-SLP-139 No.12
2021/12/2

多言語事前学習と敵対的学習による非ネイティブ英語音声認識

森 滉介[1]　篠崎 隆宏[1]

概要：言語学習者のスピーキング能力自動評価には正確な非ネイティブ音声認識が必要である．非ネイティブ話者による発音はネイティブ話者と異なるため，同じ母国語を持つ非ネイティブ話者の音声でモデルを学習させることが望ましい．しかし，利用可能なラベル付き音声データは限られており，モデルの学習データとして用いるには不十分である．そこで本稿では，データ量が豊富な言語の知識を活用する転移学習手法を提案する．この手法では，多言語のコーパスから言語に依存しない知識をモデルに事前学習させ，敵対的学習によってモデルを非ネイティブ英語に適応させる．実験では，日本人英語データセットを用いてモデルの性能を評価した．その結果 WER は，事前学習させる言語を増やすことで減少し，敵対的学習による適応によってさらに改善した．

1. はじめに

言語学習者のスピーキングにおける語彙，文法，内容を自動評価するために，正確な非ネイティブ音声認識が求められている [1–3]．言語学習者の発音は母国語の影響を受けやすいため，非ネイティブ音声認識では，同じ母国語を持つ話者の非ネイティブ音声をモデルに学習させることが望ましい [4]．それに伴って非ネイティブ音声の学習データが必要となるが，非ネイティブ音声の収集とラベル付けはネイティブ音声に比べて困難であり，十分なデータを用いた学習は難しい [5]．

非ネイティブ音声認識の有力な手法として転移学習がある [4, 6]．従来の手法では，学習データが豊富に存在する言語をモデルに事前学習させ，認識対象の非ネイティブ音声に適応する．しかしながら，事前学習させる言語は 2，3 に限られている．事前学習させる言語数を増やすことで言語知識の多様性が高まり，非ネイティブ音声認識に有用な知識をより多く含むことが期待できる．また，事前学習させたモデルを非ネイティブ音声に適応する際，ネイティブ音声を使っていない．ネイティブと非ネイティブのドメインに不変な中間表現を適応時に学習させることで，非ネイティブ音声への適応にネイティブ音声を活用できる．

そこで本稿では，高精度の非ネイティブ英語音声認識を目的として，多言語事前学習と敵対的学習による転移学習手法を提案する．実験では，日本人英語データセットを使用し，事前学習させる言語の増加と敵対的学習による適応がそれぞれ非ネイティブ英語音声の認識性能を向上させることを示す．

[1] 東京工業大学
Tokyo Institute of Technology, Tokyo, Japan
http://www.ts.ip.titech.ac.jp

2. 関連研究

非ネイティブ音声認識における複数言語を用いた転移学習では，複数言語の知識をモデルに学習させて非ネイティブ音声認識に活用する．Duan らは，共通の隠れ層と独立した出力層からなる音響モデルにネイティブの英語と日本語を学習させ，日本人英語の認識性能を向上させた [4]．また，Matassoni らはイタリア語，ドイツ語，英語のネイティブ音声で音響モデルを学習させ，イタリア人ドイツ語，イタリア人英語，ドイツ人英語の非ネイティブ音声に適応させた [6]．そして，得られた非ネイティブモデルが，非ネイティブ音声をスクラッチから学習させたモデルの性能を改善することを示した．

3. 提案手法

非ネイティブ英語の高精度音声認識を目的として，多言語事前学習と敵対的学習による転移学習手法を提案する．提案手法ではまず，大規模な多言語データを用いて音声認識モデルに多言語の知識を事前学習させる．そして，ネイティブ英語と非ネイティブ英語を用いた敵対的学習により，事前学習モデルを非ネイティブ英語に適応する．提案手法の概要を図 1 に示す．

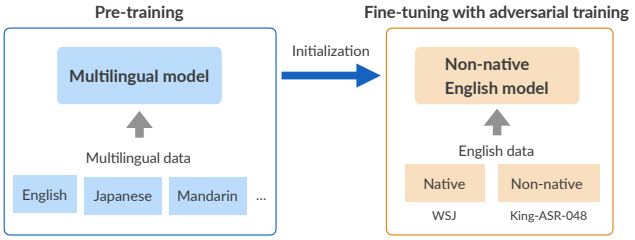

図 1 提案手法の概要図．まず，多言語データを音声認識モデルに事前学習させる．そして，ネイティブ英語と非ネイティブ英語を用いた敵対的学習により，事前学習させた多言語モデルを fine-tuning する．ネイティブ英語データには Wall Street Journal (WSJ) を使用し，非ネイティブ英語データには King-ASR-048 を用いる．

3.1 音声認識モデル

音声認識モデルは，Connectionist Temrporal Classification (CTC) と注意機構のハイブリッド構造による end-to-end ネットワーク [7] で構築する．モデルに end-to-end ネットワークを用いることで，複数の言語を単一モデルで扱いやすくなる．モデルのエンコーダには Transformer [8] または Conformer [9] のエンコーダ部分を使用し，モデルのデコーダには Transformer のデコーダ部分を用いる．

3.2 多言語事前学習

非ネイティブ英語に多言語の知識を適応させるため，モデルに多言語データを事前学習させる．モデルが出力するトークンの集合を学習データに存在する全言語の文字セットを含むように拡張する．さらに，言語の種類を示す言語トークン $l \in \{\texttt{[en]}, \texttt{[ja]}, ...\}$ をテキストの先頭に付与し，入力特徴量 $\boldsymbol{X}$ からテキスト Y と言語トークン l の同時確率分布 $P(l, Y|\boldsymbol{X})$ を学習させる．

（付与前）`THE MARKET HAS TO DO THAT`
（付与後）`[en]THE MARKET HAS TO DO THAT`

これにより，全てのパラメータを言語間で共有する単一の多言語モデルを学習することができる．

3.3 敵対的学習

事前学習させた多言語モデルを敵対的学習によって非ネイティブ英語に適応する．敵対的学習による適応ではドメイン識別器を導入し，エンコーダの出力が「ネイティブ英語から抽出した表現」か「非ネイティブ英語から抽出した表現」かを学習させる．一方で，モデルにはドメイン識別器を騙すドメイン不変な表現の抽出方法を学習させる．モデルとドメイン識別器の間で敵対的学習を実現するため，gradient reversal layer (GRL) [10] をエンコーダの出力表現に適応する．敵対的学習を用いた適応の概念図を図 2 に示す．

ネイティブ英語データのサンプル数 N_s，非ネイティブ英語データのサンプル数 N_t，エンコーダの出力フレーム数 T，ドメインラベル $d \in \{0, 1\}$，ドメイン識別器の出力確率 $D(\boldsymbol{Z})$ を用いて，ドメイン識別ロス $\mathcal{L}_\mathrm{d}$ は，

$$\mathcal{L}_\mathrm{d} = -\frac{1}{(N_\mathrm{s}+N_\mathrm{t})\,T}\sum_{i=1}^{N_\mathrm{s}+N_\mathrm{t}}\sum_{t=1}^{T}\Big\{d_t^i \log D\left(\boldsymbol{Z}_t^i\right) - \left(1-d_t^i\right)\log\left(1-D\left(\boldsymbol{Z}_t^i\right)\right)\Big\}, \quad (1)$$

と表される．敵対的学習によるドメイン適応の最適化問題は，音声認識ロス $\mathcal{L}_\mathrm{ASR}$ とドメイン識別ロス $\mathcal{L}_\mathrm{d}$ を用いて，

$$\min_{\mathrm{ASR}}\max_{\mathrm{D}}\mathcal{L} = \mathcal{L}_\mathrm{ASR} - \lambda\mathcal{L}_\mathrm{d}, \quad (2)$$

で与えられる．提案手法のアルゴリズムを Algorithm 1 に示す．

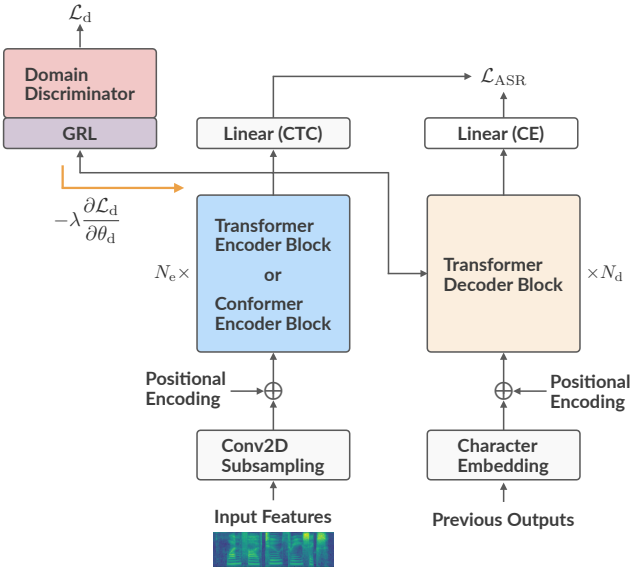

図 2 敵対的学習による事前学習モデルの適応図．ドメイン識別器を導入し，エンコーダの出力表現からモデルの入力特徴量がネイティブ英語か非ネイティブ英語かを識別できるようドメイン識別器を学習する．それに対して，モデルにはドメイン識別器を騙すドメイン不変な表現の抽出方法を学習させる．

Algorithm 1 Adversarial training for a pre-trained model

Require: native speech-text paired data $D_\mathrm{s} = \{X_\mathrm{s}^i, Y_\mathrm{s}^i\}_{i=1}^{N_\mathrm{s}}$, non-native speech-text paired data $D_\mathrm{t} = \{X_\mathrm{t}^i, Y_\mathrm{t}^i\}_{i=1}^{N_\mathrm{t}}$, batch size b, learning rate μ, adaptation parameter λ, pre-trained model weights θ_ASR.

1: Initialize an model with θ_ASR and a domain discriminator with random weights θ_d.
2: **repeat**
3: Sample minibatch $\{X_\mathrm{s}^i, Y_\mathrm{s}^i\}_{i=1}^{b/2}$ and $\{X_\mathrm{t}^i, Y_\mathrm{t}^i\}_{i=1}^{b/2}$ from D_s and D_t
4: Compute $\mathcal{L}_\mathrm{ASR}$ and $\mathcal{L}_\mathrm{d}$
5: $\theta_\mathrm{ASR} \leftarrow \theta_\mathrm{ASR} - \mu\nabla_{\theta_\mathrm{ASR}}\left(\mathcal{L}_\mathrm{ASR} - \lambda\mathcal{L}_\mathrm{d}\right)$
6: $\theta_\mathrm{d} \leftarrow \theta_\mathrm{d} - \mu\nabla_{\theta_\mathrm{d}}\mathcal{L}_\mathrm{d}$
7: **until** θ_ASR and θ_d converge

表 1 事前学習に用いた言語の ISO 639 コード．無装飾の識別子は ISO 639-1 コードを示し，ダガー（†）付きの識別子は ISO 639-3 コードを示す．6 種類の言語数異なる言語数（1, 2, 5, 10, 20, 42）をモデルに事前学習させた．

Pre-trained model	Languages
Monolingual	en
Bilingual	en ja
5-lingual	en ja zh-cn de es
10-lingual	en ja zh-cn de es fr it nl pt ru
20-lingual	en ja zh-cn de es fr it nl pt ru kab yue† ca vi tl ps tr ta ht lo
42-lingual	en ja zh-cn de es fr it nl pt ru kab yue† ca vi tl ps tr ta ht lo as fa bn zu ka jv mn sw am ig ku eu gn ceb† luo† te lt kk tpi† cy zh-tw tt

4. 実験

4.1 データセット

4.1.1 多言語

モデルの事前学習には公開コーパスを用いた．使用したコーパスは，AISHELL-1，Aurora-4，BABEL，CHiME-4，Common Voice，Corpus of Spontaneous Japanese，Fisher-CallHome Spanish，Fisher-SwitchBoard，HKUST Mandarin CTS，Voxforge，Wall Street Journal (WSJ) である．事前学習させる言語数と非ネイティブ英語音声の認識性能の関係を調べるため，6 種類の異なる言語数（1, 2, 5, 10, 20, 42）をモデルに事前学習させた．10 言語モデルの事前学習には Watanabe らが用いた言語 [11] を使用し，42 言語モデルの事前学習には Hou らが用いた言語 [12] を使用した．事前学習に用いた言語を ISO 639 コードで表 1 に示す．敵対的学習におけるネイティブ英語データには WSJ を用いた．表 3 に敵対的学習に用いた WSJ のサブセットを示す．

4.1.2 非ネイティブ英語

非ネイティブ英語のデータセットには King-ASR-048 を用いた．King-ASR-048 は，Android 端末と iOS 端末を同時に用いて収録した 2 チャンネルの日本人英語データセットである．音声データのサンプリング周波数は 16 kHz であり，1 チャンネルあたりの全収録時間は約 17.9 時間である．実験では，Android 端末で収録したチャンネルの音声データを使用した．King-ASR-048 の学習セット，検証セット，テストセットにおける発話数と長さを表 2 に示す．また，表 3 に敵対的学習に使用した King-ASR-048 のサブセットを示す．

表 2 King-ASR-048 の学習セット，検証セット，テストセットにおける発話数と長さ．

Subset	# Utterances	Length [h]
Training	8,784	14.3
Development	1,099	2.03
Test	1,100	1.68

表 3 敵対的学習における学習, 検証, テストセットの構成．ネイティブ英語データには WSJ を使用し, 非ネイティブ英語データには King-ASR-048 を用いた．使用した WSJ と King-ASR-048 のサブセットをチェックマークで示す．

Dataset	Subset		
	Training	Development	Test
WSJ	✓		
King-ASR-048	✓	✓	✓

4.2 特徴量抽出

音声データの特徴量抽出には Kaldi ツールキット [13] を用いた．25 ms の窓を 10 ms ずつ移動させ，音声データから 83 次元の特徴量（80 次元の log-mel フィルタバンクと 3 次元のピッチ特徴量）を抽出した．多言語データに対しては，抽出した特徴量のフレーム数が 2,500 以上または 10 未満の学習データとテキストの文字数が 250 以上の学習データを Hou らの実験条件 [12] に合わせて取り除いた．

4.3 詳細設定

モデルの構築，学習，デコードには，ESPnet ツールキット [14] を使用した．モデル構築におけるエンコーダブロック数 N_{e} は 12，デコーダブロック数 N_{d} は 6 である．また，エンコーダブロックとデコーダブロックのそれぞれにおいて，Feed Forward ネットワークの次元 d_{ff} を 2048，自己注意ヘッドの次元 d_{att} を 256，ヘッド数 H を 4 とした．Conformer の Convolution モジュールにおけるカーネルサイズは 31 である．

ドメイン識別器は，全結合層，1 次元バッチ正規化（Batch Normalization; BN），活性化関数 ReLU を用いて構築した．エンコーダの出力表現を 256 次元の全結合層で処理し，ReLU と 1 次元 BN による変換を施した．そして，1 次元の全結合層とシグモイド関数によって確率を算出した．

事前学習と fine-tuning のオプティマイザには Adam [15]（$\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1\mathrm{e}{-9}$）を用いた．事前学習における学習率 μ は，Vaswani らのスケジューリング [16] によって変化させた．スケジューリングにおける学習率 μ は，学習率パラメータ k，自己注意ヘッドの次元 d_{att}，ステップ数 $step$，ウォームアップステップ数 $wustep$ を用いて，

$$\mu = k \cdot d_{\mathrm{att}}^{-0.5} \cdot \min\left(step^{-0.5}, step \cdot wustep^{-1.5}\right), \quad (3)$$

と表される．学習率パラメータ k とウォームアップステッ

表 4 学習とデコードに関するハイパーパラメータ設定．学習とデコードには ESPnet ツールキット [14] を使用した．

Train	
CTC loss weight	0.3
Label smoothing weight	0.1
Dropout rate	0.1
Gradient accumulation	1
Gradient clipping	5
Decode	
CTC decoding weight	0.5
Incertion penalty	0
Beam size	10

プ数 $wustep$ は, Hou らの学習設定 [12] に合わせて $k = 4.5$, $wustep = 25000$ とした．一方で，fine-tuning における学習率は {5e−5, 1e−4, 2e−4, 5e−4} の中からグリッドサーチにて探索した．事前学習におけるバッチサイズは 1280, fine-tuning におけるバッチサイズは 8 である．事前学習と fine-tuning におけるその他のハイパーパラメータ設定を表 4 に示す．

敵対的学習にける適応パラメータ λ は，Ganin らのスケジューリング [10] によって変化させた．適応パラメータ λ の変化式は，ステップ数 $step$ と全ステップ数 T を用いて，

$$\lambda = \frac{2}{1 + \exp\left(-10 \cdot \frac{step}{T}\right)} - 1, \tag{4}$$

で与えられる．

デコードでは，CTC の出力確率とデコーダの出力確率を同時に用いた [7]．デコードに関するハイパーパラメータを表 4 に示す．デコードには外部の言語モデルを使用しない．

4.4 比較モデル

提案手法の有効性を検証するため，スクラッチからの学習や fine-tuning に基づく学習によるベースラインを設計し，提案手法を含む以下のモデルと単語誤り率（Word Error Rate; WER）を比較した．

- **Scratch**: モデルのパラメータを乱数で初期化し，非ネイティブ英語をスクラッチから学習させる．
- **ScratchAT**: モデルのパラメータを乱数で初期化し，非ネイティブ英語とネイティブ英語を用いてスクラッチから敵対的学習させる．
- **FT**: 事前学習モデルでモデルのパラメータを初期化し，非ネイティブ英語で fine-tuning する．
- **FT+native**: 事前学習モデルでモデルのパラメータを初期化し，非ネイティブ英語とネイティブ英語の両方を用いて fine-tuning する．
- **FTAT (proposed)**: 事前学習モデルでモデルのパラメータを初期化し，非ネイティブ英語とネイティブ英語を敵対的学習を用いて fine-tuning する．

表 5 Scratch モデル, ScratchAT モデル, FT モデルの WER [%]. FT モデルにおける言語数は事前学習させた言語数を示す．

Model	Method	WER [%]	
		dev	test
Transformer	Scratch	89.0	90.0
	ScratchAT	70.2	71.6
	FT		
	Monolingual	35.9	28.1
	Bilingual	34.6	27.3
	5-lingual	35.3	27.6
	10-lingual	**34.6**	**26.4**
	20-lingual	35.4	28.2
	42-lingual	36.2	28.7
Conformer	Scratch	71.0	72.8
	ScratchAT	61.2	64.8
	FT		
	Monolingual	34.1	26.6
	Bilingual	32.9	24.4
	5-lingual	32.5	23.7
	10-lingual	**32.5**	**23.4**
	20-lingual	33.1	25.1
	42-lingual	32.8	25.7

表 6 FT モデル，FT+native モデル，FTAT モデルの WER [%]. 全モデルのパラメータは 10 言語モデルで初期化した．

Model	Method	WER [%]	
		dev	test
Transformer	FT	34.6	26.4
	FT+native	34.6	25.8
	FTAT (proposed)	**33.5**	**25.3**
Conformer	FT	32.5	23.4
	FT+native	32.2	22.8
	FTAT (proposed)	**32.0**	**22.4**

5. 結果

表 5 に Scratch モデル，ScratchAT モデル，FT モデルの WER を示す．Transformer と Conformer の両モデルにおいて，Scratch モデルと ScratchAT モデルの WER を FT モデルは大幅に改善した．FT モデルにおける WER を比較すると，事前学習させる言語数を 1 から 10 まで増加させた場合，10 言語を事前学習させた FT モデルが最高の認識性能を示した．これにより，適応させる知識の多言語拡張が非ネイティブ英語音声認識に有効であることが確認できた．

一方で，事前学習させる言語数を 10 より多くすると認識性能は低下した．これは，言語知識の多様性と量のトレードオフが影響している可能性がある [17]．モデルのパラメータ数を固定した状態において，事前学習させる言語

が多いほどモデルが持つ言語知識の多様性は増加するが，モデルが保持できる各言語の知識量は減少する．事前学習言語数を 10 から増加させたことにより，10 言語の中ですでに学習したネイティブ英語音声の認識に有効な知識が減少した可能性がある．

表 6 に FT モデル，FT+native モデル，FTAT モデルの WER を示す．モデルパラメータの初期化には，10 言語を事前学習させたモデルを用いた．FT+native モデルは FT モデルの WER を改善した．さらに，提案手法である FTAT モデルは FT+native モデルを上回る認識性能を示した．これは，非ネイティブ英語とネイティブ英語のドメインに不変な表現を敵対的学習によって学習させることで，非ネイティブ英語への適応にネイティブ英語を有効活用できることを示唆している．

6. おわりに

本稿では，高精度の非ネイティブ英語音声認識を目的として，多言語事前学習と敵対的学習による転移学習手法を提案した．実験では，日本人英語データセットを使用し，WER によってモデルの認識性能を評価した．その結果，事前学習させる言語を増やすことで WER が改善した．さらに，モデルの適応に敵対的学習を導入することにより，WER をさらに改善できることが分かった．

謝辞 本研究は，JSPS 科研費 20H00095 の助成を受けたものである．

参考文献

[1] Qian, Y., Wang, X., Evanini, K. and Suendermann-Oeft, D.: Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment, *Interspeech*, pp. 3122–3126 (2016).
[2] Qian, Y., Evanini, K., Wang, X., Lee, C. M. and Mulholland, M.: Bidirectional LSTM-RNN for Improving Automated Assessment of Non-Native Children’s Speech, *Interspeech*, pp. 1417–1421 (2017).
[3] Knill, K., Gales, M., Kyriakopoulos, K., Malinin, A., Ragni, A., Wang, Y. and Caines, A.: Impact of ASR Performance on Free Speaking Language Assessment, *Interspeech*, pp. 1641–1645 (2018).
[4] Duan, R., Kawahara, T., Dantsuji, M. and Nanjo, H.: Cross-Lingual Transfer Learning of Non-Native Acoustic Modeling for Pronunciation Error Detection and Diagnosis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 391–401 (2020).
[5] Chen, N. F., Wee, D., Tong, R., Ma, B. and Li, H.: Large-scale characterization of non-native Mandarin Chinese spoken by speakers of European origin: Analysis on iCALL, *Speech Communication*, Vol. 84, pp. 46–56 (2016).
[6] Matassoni, M., Gretter, R., Falavigna, D. and Giuliani, D.: Non-Native Children Speech Recognition Through Transfer Learning, *ICASSP*, pp. 6229–6233 (2018).
[7] Kim, S., Hori, T. and Watanabe, S.: Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning, *ICASSP*, pp. 4835–4839 (2017).
[8] Karita, S., Soplin, N. E. Y., Watanabe, S., Delcroix, M., Ogawa, A. and Nakatani, T.: Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration, *Interspeech*, pp. 1408–1412 (2019).
[9] Gulati, A., Qin, J., Chiu, 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, *Interspeech*, pp. 5036–5040 (2020).
[10] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M. and Lempitsky, V.: Domain-Adversarial Training of Neural Networks, *Journal of Machine Learning Research*, Vol. 17, No. 59, pp. 1–35 (2016).
[11] Watanabe, S., Hori, T. and Hershey, J. R.: Language independent end-to-end architecture for joint language identification and speech recognition, *ASRU*, pp. 265–271 (2017).
[12] Hou, W., Dong, Y., Zhuang, B., Yang, L., Shi, J. and Shinozaki, T.: Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning, *Interspeech*, pp. 1037–1041 (2020).
[13] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K.: The Kaldi speech recognition toolkit, *ASRU* (2011).
[14] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. and Ochiai, T.: ESPnet: End-to-End Speech Processing Toolkit, *Interspeech*, pp. 2207–2211 (2018).
[15] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *ICLR* (2015).
[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is All you Need, *NeurIPS*, pp. 5998–6008 (2017).
[17] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale, *ACL*, pp. 8440–8451 (2020).