XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech

Dat Quoc Nguyen
Head of NLP department, VinAI
https://datquocnguyen.github.io

Joint work with
Linh The Nguyen & Thinh Pham

21/04/2023
Motivation

- Language model BERT [1]—Bidirectional Encoder Representations from Transformers [2]
  - BERT and its variants, pre-trained on large-scale corpora, help improve the state-of-the-art performances of various NLP research and application tasks
  - Represent words by embedding vectors which encode the contexts where the words appear, i.e. contextualized word embeddings

https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
Motivation

• Several Text-to-Speech (TTS) works incorporate contextualized word embeddings generated by the pre-trained BERT into their standard encoder [3, 4, 5]
  • An input phoneme sequence is fed into the standard encoder to produce phoneme representations
  • The corresponding input raw text is fed into BERT to obtain contextualized word embeddings
  • Concatenate the representations of the input phonemes with the corresponding BERT-based word embeddings to construct the input vectors of the TTS decoder (for decoding)

Phoneme sequence:   'eɪ ɪˈmætɪʃən ˈmʌplɪ ə ˈmʌltɪliŋɡwəl ˈmɒdl
Raw text:            a multilingual model
Motivation

• Several Text-to-Speech (TTS) works incorporate contextualized word embeddings generated by the pre-trained BERT into their standard encoder [3, 4, 5]
  • Provide extra contextual information for phoneme representation

• Pre-trained models for phoneme representations, including PnG BERT [6], Mixed-Phoneme BERT [7] and Phoneme-level BERT [8], help improve advanced TTS systems
  • PnG BERT takes both phonemes and graphemes (i.e. subword tokens) as the input
  • Mixed-Phoneme BERT takes both phonemes and sup-phoneme tokens as the input
  • Phoneme-level BERT only takes phonemes as the input
Motivation

- Pre-trained models for phoneme representations, including PnG BERT [6], Mixed-Phoneme BERT [7] and Phoneme-level BERT [8], help improve advanced TTS systems
  - Limited to English only
  - Not publicly available

⇒ A need of developing pre-trained models for phoneme representations in languages other than English
Motivation

- **Our contributions:**
  - Present the first large-scale pre-trained multilingual model for phoneme representations, named XPhoneBERT
  - XPhoneBERT helps significantly improve the performance of the strong TTS baseline VITS [9]
  - We will publicly release XPhoneBERT
Pre-training XPhoneBERT

- Multilingual pre-training data
  - *Raw dataset collection:* Collect texts for the 90+ languages and locales supported by CharsiuG2P [10], from multilingual datasets *wiki40b* and *wikipedia*
    - Perform word and sentence segmentation as well as duplicate removal and text normalization
  - *Text-to-phoneme conversion*
  - *Phoneme segmentation*
Pre-training XPhoneBERT

- Multilingual pre-training data
  - Raw dataset collection
  - Text-to-phoneme conversion: Employ the grapheme-to-phoneme conversion toolkit CharsiuG2P to convert sentences into their phonemic description
  - Phoneme segmentation: Employ the segments toolkit for phoneme segmentation for a better map between phonemes and speech
  - Demonstration example:

a multilingual model → CharsiuG2P → 'eiˌməltiˈliːmədəl' → segments → 'eɪˌmʌltɪliːmədəl'
Pre-training XPhoneBERT

- Multilingual pre-training data
- A pre-training corpus of 330M phoneme-level sentences across 94 languages and locales
Pre-training XPhoneBERT

- Pre-training approach
  - Employ the BERT-base model architecture
  - Use a whitespace tokenizer, thus resulting in a vocabulary of 1960 phoneme types, and a model of 87.6M parameters
  - Use the masked language modeling objective and follow the RoBERTa pre-training approach [11]
    - Use a dynamic masking strategy and without the next sentence prediction objective
  - Train for 20 epochs in about 18 days
Downstream TTS evaluation

• Evaluate the effectiveness of XPhoneBERT on the TTS task for English and Vietnamese
  • English: Training, validation and test sets of 12,500, 100 and 500 audio clips
  • Vietnamese: Training, validation and test sets of 12,000, 100 and 200 clips

• Employ the strong TTS model VITS [9] as the baseline
  • VITS is an end-to-end model that contains a Transformer encoder to encode the input phoneme sequence
  • Extend VITS with XPhoneBERT by replacing the VITS’s encoder with XPhoneBERT
Downstream TTS evaluation

- Two training settings using 100% and 5% of the TTS training data

- Evaluation metrics
  - Subjective evaluation: For each language, following [9], we randomly select 50 ground truth test audios and their text transcription to measure the Mean Opinion Score (MOS)
  - Objective evaluation: We compute two metrics of the mel-cepstrum distance (MCD) and the F0 root mean square error (RMSE; cent)
**Downstream TTS evaluation**

- XPhoneBERT helps improve the performance of VITS on all three evaluation metrics for both English and Vietnamese in both experimental settings:
  - 100% of the training set for training: MOS significantly increases from 4.00 to 4.14 (+0.14) for English and from 3.74 to 3.89 (+0.15) for Vietnamese
  - 5% of the training set for training: MOS increases from 2.88 to 3.22 (+0.34) for English and especially from 1.59 to 3.35 (+1.76) for Vietnamese

**Obtained results on the English test set**

<table>
<thead>
<tr>
<th>Model</th>
<th>MOS (↑)</th>
<th>MCD (↓)</th>
<th>RMSE_{F0} (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td>4.39 ± 0.08</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Baseline VITS</td>
<td>4.00 ± 0.08</td>
<td>7.04</td>
<td>377</td>
</tr>
<tr>
<td>VITS w/ XPB</td>
<td>4.14 ± 0.07</td>
<td>6.63</td>
<td>348</td>
</tr>
<tr>
<td>5% Baseline VITS</td>
<td>2.88 ± 0.11</td>
<td>7.40</td>
<td>407</td>
</tr>
<tr>
<td>5% VITS w/ XPB</td>
<td>3.22 ± 0.11</td>
<td>7.15</td>
<td>383</td>
</tr>
</tbody>
</table>

**Obtained results on the Vietnamese test set**

<table>
<thead>
<tr>
<th>Model</th>
<th>MOS (↑)</th>
<th>MCD (↓)</th>
<th>RMSE_{F0} (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td>4.26 ± 0.06</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Baseline VITS</td>
<td>3.74 ± 0.08</td>
<td>5.41</td>
<td>249</td>
</tr>
<tr>
<td>VITS w/ XPB</td>
<td>3.89 ± 0.08</td>
<td>5.12</td>
<td>234</td>
</tr>
<tr>
<td>5% Baseline VITS</td>
<td>1.59 ± 0.05</td>
<td>6.20</td>
<td>291</td>
</tr>
<tr>
<td>5% VITS w/ XPB</td>
<td>3.35 ± 0.10</td>
<td>5.39</td>
<td>248</td>
</tr>
</tbody>
</table>

“XPB” abbreviates XPhoneBERT
Downstream TTS evaluation

- MOS is not “always” correlated with MCD and RMSE
  - For Vietnamese, VITS under the first setting (100%) obtains higher MOS but slightly poorer MCD and RMSE than VITS extended with XPhoneBERT under the second setting (5%)
  - XPhoneBERT helps synthesize fairly high-quality speech with limited training data (5%)

<table>
<thead>
<tr>
<th>Model</th>
<th>MOS (↑)</th>
<th>MCD (↓)</th>
<th>RMSE$_{F_0}$ (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td>4.39 ± 0.08</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Baseline VITS</td>
<td>4.00 ± 0.08</td>
<td>7.04</td>
<td>377</td>
</tr>
<tr>
<td>VITS w/ XPB</td>
<td>4.14 ± 0.07</td>
<td>6.63</td>
<td>348</td>
</tr>
<tr>
<td>Baseline VITS</td>
<td>2.88 ± 0.11</td>
<td>7.40</td>
<td>407</td>
</tr>
<tr>
<td>VITS w/ XPB</td>
<td>3.22 ± 0.11</td>
<td>7.15</td>
<td>383</td>
</tr>
</tbody>
</table>

Obtained results on the Vietnamese test set

<table>
<thead>
<tr>
<th>Model</th>
<th>MOS (↑)</th>
<th>MCD (↓)</th>
<th>RMSE$_{F_0}$ (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td>4.26 ± 0.06</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Baseline VITS</td>
<td>3.74 ± 0.08</td>
<td>5.41</td>
<td>249</td>
</tr>
<tr>
<td>VITS w/ XPB</td>
<td>3.89 ± 0.08</td>
<td>5.12</td>
<td>234</td>
</tr>
<tr>
<td>Baseline VITS</td>
<td>1.59 ± 0.05</td>
<td>6.20</td>
<td>291</td>
</tr>
<tr>
<td>VITS w/ XPB</td>
<td>3.35 ± 0.10</td>
<td>5.39</td>
<td>248</td>
</tr>
</tbody>
</table>

“XPB” abbreviates XPhoneBERT
Downstream TTS evaluation

• Spectrograms visualization by different models, illustrating that XPhoneBERT helps improve the spectral details of the baseline VITS’s output
  • “Ít ai biết được rằng nơi này trước kia từng là một mỏ đá với không ai để ý tới” (Little is known that this place was once a limestone quarry that no one paid any attention to)
  • (a): Ground truth
  • (b): VITS with XPhoneBERT (100%)
  • (c): VITS with XPhoneBERT (5%)
  • (d): Original VITS (100%)
  • (e): Original VITS (5%)
Takeaways

• XPhoneBERT is the first large-scale multilingual language model pre-trained for phoneme representations.

• Using XPhoneBERT as an input phoneme encoder improves the quality of the speech synthesized by a strong neural TTS baseline:
  • XPhoneBERT also helps produce fairly high-quality speech when the training data is limited.

• We will publicly release XPhoneBERT, which can be used with popular libraries `fairseq` and `transformers`. 
Thank you!

@VinAI

https://www.vinai.io/
References

1. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding
2. Attention is All you Need
3. Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis
4. Improving the Prosody of RNN-based English Text-To-Speech Synthesis by Incorporating a BERT model
5. Improving Prosody Modelling with Cross-Utterance BERT Embeddings for End-to-end Speech Synthesis
6. PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS
7. Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech
8. Phoneme-Level BERT for Enhanced Prosody of Text-to-Speech with Grapheme Predictions
9. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech
10. ByT5 model for massively multilingual grapheme-to-phoneme conversion
11. RoBERTa: A Robustly Optimized BERT Pretraining Approach