DeepType: On-Device Deep Learning for Input Personalization Service with Minimal Privacy Concern

Mengwei Xu¹, Feng Qian², Qiaozhu Mei³
Kang Huang⁴, Xuanzhe Liu¹, Yun Ma¹

¹Peking University, ²University of Minnesota, ³University of Michigan, ⁴Kika Tech
Everyone types a lot everyday

• Per day on earth: 2M Reddit posts, 5M tweets, 100B instant messages, and 200B emails
• A large portion of them are done on mobile devices, which makes:

Input method application (IMA): a killer app
Next-word prediction: a killer feature for productivity
DL-powered next-word prediction

- Next-word prediction techniques has evolved to deep learning (DL)

- **Cheaper and more accurate**
- **More expensive**
  - both training and prediction

Diagram:
- **Dictionary lookup**
  - cheap, inaccurate
- **Traditional ML algs**
  - ngram
- **Deep learning**
  - LSTM
LSTM model for next-word prediction

Predicted word: “priceless”

hidden states

LSTM Cell → LSTM Cell → LSTM Cell → LSTM Cell

embedding lookup (ids → vectors)

vocabulary lookup (words/chars → ids)

“health” → “is” → “p” → “r”
Personalizing prediction models

- Can we further improve the accuracy of DL models?

- The models need to be personalized and adapt to diverse users
  - Training one model for one user using his/her own data

Tomorrow I will go to the **party**

Tomorrow I will go to the **class**
On-cloud personalization is not a good idea

privacy concern
password
accounts
emails
messages
credit card
numbers

scalability issue
Personalizing 1M users takes 36,000 GPU-hrs. Too expensive!

Can we personalize (train) the DL model on mobile devices?
Challenges of on-device personalization

• **Limited data volume**
  
  *Is it enough to make model converge*

• **Limited computational resources**

  *Can we train model w/o compromising user experience*
Challenges of on-device personalization

• Limited data volume

  Is it enough to make model converge

  Key idea 1: use public corpora to pre-train a global model before on-device personalization

• Limited computational resources

  Can we train model w/o compromising user experience

  Key idea 2: compress, customize, and fine-tune the model
DeepType: on-device personalization

- Fresh model
- Cloud training
- Global model
- Public corpora
DeepType: on-device personalization

fresh model → cloud training → global model

public corpora

CLOUD

global model → offline training → personal model

private corpora

DEVICE
DeepType: on-device personalization
DeepType: on-device personalization

- Good privacy: input data never leaves mobile device
- Good flexibility: the model can be updated anytime with small cost
Reducing on-device computations

1. SVD-based model compression (on cloud)
2. Vocabulary compression
3. Fine-tune training
4. Reusing inference results
Reducing on-device computations

1. SVD-based model compression
2. Vocabulary compression (on device)
3. Fine-tune training
4. Reusing inference results

<table>
<thead>
<tr>
<th></th>
<th>Global vocabulary</th>
<th>Personal vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>To cover 95% occurrences</td>
<td>20,000 words</td>
<td>6,000 words</td>
</tr>
</tbody>
</table>

Vocabulary size used by 1M users within 6 months (Jul. 2017 to Dec. 2017). Mean: 6214, median: 5911
Reducing on-device computations

1. SVD-based model compression
2. Vocabulary compression
3. Fine-tune training (on-device)
4. Reusing inference results
Reducing on-device computations

1. SVD-based model compression
2. Vocabulary compression
3. Fine-tune training
4. Reusing inference results (on-device online training)
Implementation and Evaluation

• Extension to TensorFlow

• Dataset: half-year input data from 1M real users
  • IRB-approved, fully anonymized
  • Over 10 billion messages in English

• Metrics:
  • Input efficiency (accuracy)
  • On-device overhead (latency & energy)

<table>
<thead>
<tr>
<th>User input</th>
<th>User wants</th>
<th>Model output (top 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I”</td>
<td>“will”</td>
<td>[“am”, “have”, “don’t”]</td>
</tr>
<tr>
<td>“I”, “w”</td>
<td>“will”</td>
<td>[“was”, “would”, “wish”]</td>
</tr>
<tr>
<td>“I”, “wi”</td>
<td>“will”</td>
<td>[“wish”, “will”, “with”]</td>
</tr>
</tbody>
</table>

Top-3-efficiency = \(1 - \frac{2}{4}\)

How many chars user has to input to get the correct prediction

Length of output word “will”
DeepType improves model accuracy

<table>
<thead>
<tr>
<th>pre-train dataset (global model)</th>
<th>personalization (private model)</th>
<th>top-3-efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter corpora</td>
<td>✓</td>
<td>0.616</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.513</td>
</tr>
<tr>
<td>Wikipedia corpora</td>
<td>✓</td>
<td>0.508</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.325</td>
</tr>
<tr>
<td>private corpora</td>
<td>✓</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.568</td>
</tr>
<tr>
<td>no pre-train</td>
<td>✓</td>
<td>0.331</td>
</tr>
</tbody>
</table>

DeepType no personalization
DeepType improves model accuracy

<table>
<thead>
<tr>
<th>pre-train dataset (global model)</th>
<th>personalization (private model)</th>
<th>top-3-efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter corpora</td>
<td>✓</td>
<td>0.616</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.513</td>
</tr>
<tr>
<td>Wikipedia corpora</td>
<td>✓</td>
<td>0.508</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.325</td>
</tr>
<tr>
<td>private corpora</td>
<td>✓</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.568</td>
</tr>
<tr>
<td>no pre-train</td>
<td>✓</td>
<td>0.331</td>
</tr>
</tbody>
</table>
DeepType improves model accuracy

<table>
<thead>
<tr>
<th>pre-train dataset</th>
<th>personalization</th>
<th>top-3-efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>(global model)</td>
<td>(private model)</td>
<td></td>
</tr>
<tr>
<td>Twitter corpora</td>
<td>✓</td>
<td>0.616</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.513</td>
</tr>
<tr>
<td>Wikipedia corpora</td>
<td>✓</td>
<td>0.508</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.325</td>
</tr>
<tr>
<td>private corpora</td>
<td>✓</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>0.568</td>
</tr>
<tr>
<td>no pre-train</td>
<td>✓</td>
<td>0.331</td>
</tr>
</tbody>
</table>

DeepType
Ideal but impractical. Bad user privacy
DeepType improves model accuracy

<table>
<thead>
<tr>
<th>pre-train dataset (global model)</th>
<th>personalization (private model)</th>
<th>top-3-efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter corpora</td>
<td>✓</td>
<td>0.616</td>
</tr>
<tr>
<td></td>
<td>✘</td>
<td>0.513</td>
</tr>
<tr>
<td>Wikipedia corpora</td>
<td>✓</td>
<td>0.508</td>
</tr>
<tr>
<td></td>
<td>✘</td>
<td>0.325</td>
</tr>
<tr>
<td>private corpora</td>
<td>✓</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>✘</td>
<td>0.568</td>
</tr>
<tr>
<td>no pre-train</td>
<td>✓</td>
<td>0.331</td>
</tr>
</tbody>
</table>
DeepType reduces on-device overhead

- **91.6%** reduction of training time
  - Less than 1.5 hours to personalize the model on half-year input history
- **90.3%** reduction of energy consumption

---

**Training time on different Android devices**

- Nexus 6: 3.5 hours
- Pixel: 3.0 hours
- Note 8: 2.5 hours
- Xiaomi 6: 2.0 hours

---

**Normalized energy consumption with/without optimization**

- Without optimization: 1.0
- With optimization: 0.5

---

**Training energy w/ and w/o optimization**
DeepType reduces on-device overhead

- 91.6% reduction of training time
  - Less than 1.5 hours to personalize the model on half-year input history
- 90.3% reduction of energy consumption

Device is in **favored state**

1. Device is one
2. Device screen is turned off
3. Device is being charged and has high remaining battery

More than 50% users spend around 2.7 hours on favored states per day -> enough for offline training!
DeepType reduces on-device overhead

- **91.6% reduction of training time**
  - Less than 1.5 hours to personalize the model on half-year input history
- **90.3% reduction of energy consumption**

- On-device online training typically takes only 20ms~60ms
  - Unnoticeable to users
DeepType improves the user experience

- A field study: 34 voluntary subjects in Indiana University, 3 weeks.
  - Embed DeepType into a commercial keyboard app

  ![Kika Logo]

- Quantitative analysis
  - Prediction: 25ms, training (online): 86ms << inter-keystroke: 264ms

- Qualitative analysis (feedbacks):
  - 78% users report improved accuracy
  - 93.7% users report good responsiveness
  - 100% users report no battery impacts
Summary

• On-cloud personalization vs. on-device personalization
  • Privacy and scalability matter

• DeepType: on-device personalization framework
  • Cloud pre-train, device fine-tune -> ensure both privacy and accuracy
  • Model compression and customized -> reduce computation overhead

Thank you for attention!