

Federated Learning of N-gram Language Models

Mingqing Chen, Ananda Theertha Suresh, Rajiv Mathews, Adeline Wong,
Cyril Allauzen, Françoise Beaufays, Michael Riley

Google, Inc.

{mingqing, theertha, mathews, adelinew, allauzen, fsb, riley}
@google.com

Abstract

We propose algorithms to train production-quality n-gram language models using federated learning. Federated learning is a distributed computation platform that can be used to train global models for portable devices such as smart phones. Federated learning is especially relevant for applications handling privacy-sensitive data, such as virtual keyboards, because training is performed without the users' data ever leaving their devices. While the principles of federated learning are fairly generic, its methodology assumes that the underlying models are neural networks. However, virtual keyboards are typically powered by n-gram language models for latency reasons.

We propose to train a recurrent neural network language model using the decentralized `FederatedAveraging` algorithm and to approximate this federated model server-side with an n-gram model that can be deployed to devices for fast inference. Our technical contributions include ways of handling large vocabularies, algorithms to correct capitalization errors in user data, and efficient finite state transducer algorithms to convert word language models to word-piece language models and vice versa. The n-gram language models trained with federated learning are compared to n-grams trained with traditional server-based algorithms using A/B tests on tens of millions of users of a virtual keyboard. Results are presented for two languages, American English and Brazilian Portuguese. This work demonstrates that high-quality n-gram language models can be trained directly on client mobile devices without sensitive training data ever leaving the devices.



Figure 1: Glide trails are shown for two spatially-similar words: “Vampire” (in red) and “Value” (in orange). Viable decoding candidates are proposed based on context and language model scores.

1 Introduction

1.1 Virtual keyboard applications

Virtual keyboards for mobile devices provide a host of functionalities from decoding noisy spatial signals from tap and glide typing inputs to providing auto-corrections, word completions, and next-word predictions. These features must fit within tight RAM and CPU budgets, and operate under strict latency constraints. A key press should result in visible feedback within about 20 milliseconds (Ouyang et al., 2017; Alsharif et al., 2015). Weighted finite-state transducers have been used successfully to decode keyboard spatial signals using a combination of spatial and language models (Ouyang et al., 2017; Hellsten et al., 2017). Figure 1 shows the glide trails of two spatially-similar words. Because of the similarity of the two trails, the decoder must rely on the language model to discriminate between viable candidates. For memory and latency reasons, especially on low-end devices, the language models are typically based on n-grams and do not exceed ten megabytes. A language model (LM) is a probabilistic model on words. Given previous words x_1, x_2, \dots, x_{m-1} , an LM assigns a probability to the new words, i.e. $p(x_m|x_{m-1}, \dots, x_1)$. An n-gram LM is a Markovian distribution of order

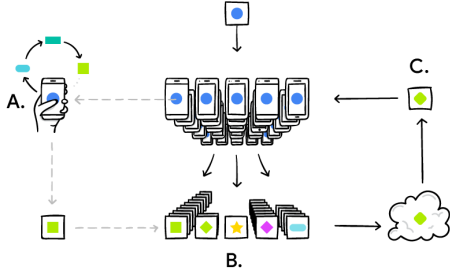


Figure 2: An illustration of the federated learning process from McMahan and Ramage (2017): (A) client devices compute SGD updates on locally-stored data, (B) a server aggregates the client updates to build a new global model, (C) the new model is sent back to clients, and the process is repeated.

$n - 1$, defined by

$$p(x_m | x_{m-1}, \dots, x_1) = p(x_m | x_{m-1}, \dots, x_{m-n+1}),$$

where n is the order of the n-gram. For computation and memory efficiency, keyboard LMs typically have higher-order n-grams over a subset of the vocabulary, e.g. the most frequent 64K words, and the rest of the vocabulary only has unigrams. We consider n-gram LMs that do not exceed 1.5M n-grams and include fewer than 200K unigrams.

N-gram models are traditionally trained by applying a smoothing method to n-gram counts from a training corpus (Chen and Goodman, 1999). The highest quality n-gram models are trained over data that are well-matched to the desired output (Moore and Lewis, 2010). For virtual keyboards, training over users’ typed text would lead to the best results. Of course, such data are very personal and need to be handled with care.

1.2 Federated learning

We propose to leverage *Federated Learning* (FL) (Konečný et al., 2016; Konečný et al., 2016), a technique where machine learning models are trained in a decentralized manner on end-users’ devices, so that raw data never leaves these devices. Only targeted and ephemeral parameter updates are aggregated on a centralized server. Figure 2 provides an illustration of the process. Federated learning for keyboard input was previously explored in Hard et al. (2018), in which a federated recurrent neural network (RNN) was trained for next-word prediction. However, latency constraints prevent the direct use of an RNN for decoding. To overcome this problem, we propose

to derive an n-gram LM from a federated RNN LM model and use that n-gram LM for decoding. Specifically, the approximation algorithm is based on `SampleApprox`, which was recently proposed in Suresh et al. (2019a,b). The proposed approach has several advantages:

Improved model quality: Since the RNN LM is trained directly on domain-matched user data, its predictions are more likely to match actual user behavior. In addition, as shown in Suresh et al. (2019a), an n-gram LM approximated from such an RNN LM is of higher quality than an n-gram LM trained on user data directly.

Minimum information transmission: In FL, only the minimal information necessary for model training (the model parameter deltas) is transmitted to centralized servers. The model updates contain much less information than the complete training data.

Additional privacy-preserving techniques: FL can be further combined with privacy-preserving techniques such as secure multi-party computation (Bonawitz et al., 2017) and differential privacy (McMahan et al., 2018; Agarwal et al., 2018; Abadi et al., 2016). By the post-processing theorem, if we train a single differentially private recurrent model and use it to approximate n-gram models, all the distilled models will also be differentially private with the same parameters (Dwork et al., 2014).

For the above reasons, we have not proposed to learn n-gram models directly using `FederatedAveraging` of n-gram counts for all orders.

2 Outline

The paper is organized along the lines of challenges associated with converting RNN LMs to n-gram LMs for virtual keyboards: the feasibility of training neural models with a large vocabulary, inconsistent capitalization in the training data, and data sparsity in morphologically rich languages. We elaborate on each of these challenges below.

Large vocabulary: Keyboard n-gram models are typically based on a carefully hand-curated vocabulary to eliminate misspellings, erroneous capitalizations, and other artifacts. The vocabulary size often numbers in the hundreds of thousands. However, training a neural model directly over the vocabulary is memory intensive as the embedding and softmax layers require space $|\mathcal{V}| \times N_e$, where

$|\mathcal{V}|$ is the vocabulary size and N_e is the embedding dimension. We propose a way to handle large vocabularies for federated models in Section 3.

Incorrect capitalization: In virtual keyboards, users often type with incorrect casing (e.g. “She lives in new york” instead of “She lives in New York”). It would be desirable to decode with the correct capitalization even though the user-typed data may be incorrect. Before the discussion of capitalization, the `SampleApprox` algorithm is reviewed in Section 4. We then modify `SampleApprox` to infer capitalization in Section 5.

Language morphology: Many words are composed of root words and various morpheme components, e.g. “crazy”, “crazily”, and “craziness”. These linguistic features are prominent in morphologically rich languages such as Russian. The presence of a large number of morphological variants increases the vocabulary size and data sparsity ultimately making it more difficult to train neural models. Algorithms to convert between word and word-piece models are discussed in Section 6.

Finally, we compare the performance of word and word-piece models and present the results of A/B experiments on real users of a virtual keyboard in Section 7.

3 Unigram distributions

Among the 200K words in the vocabulary, our virtual keyboard models only use the top 64K words in the higher-order n-grams. We train the neural models only on these most frequent words and train a separate unigram model over the entire vocabulary. We interpolate the two resulting models to obtain the final model for decoding.

3.1 Collection

Unigrams are collected via a modified version of the `FederatedAveraging` algorithm. No models are sent to client devices. Instead of returning gradients to the server, counting statistics are compiled on each device and returned. In our experiments, we aggregate over groups of approximately 500 devices per training round. We count a unigram distribution U from a whitelist vocabulary by $U = \sum_i w_i C_i$, where i is the index over devices, C_i are the raw unigram counts collected from a single device i , and w_i is a weight applied to device i .

To prevent users with large amounts of data

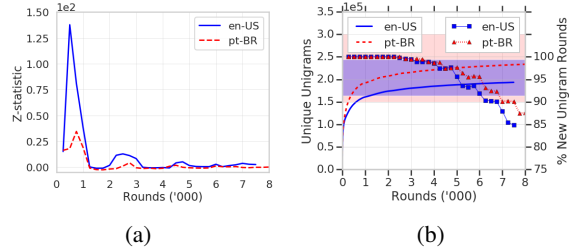


Figure 3: Unigram distribution convergence. Note that by 3000 rounds, the unigram distribution is stable, but the model is still learning new tail unigrams.

from dominating the unigram distribution, we apply a form of L1-clipping:

$$w_i = \frac{\lambda}{\max(\lambda, \sum C_i)}, \quad (1)$$

where λ is a threshold that caps each device’s contribution. When $\lambda = 1$, L1-clipping is equivalent to equal weighting. The limit $\lambda \rightarrow \infty$ is equivalent to collecting the true counts, since $w_i \rightarrow 1$.

3.2 Convergence

Convergence of the unigram distribution is measured using the unbiased chi-squared statistic (for simplicity, referred to as the Z -statistic) defined in [Bhattacharya and Valiant \(2015\)](#), the number of unique unigrams seen, and a moving average of the number of rounds needed to observe new unigrams.

Figure 3(a) shows the overall distributional convergence based on the Z -statistic. At round k , unigram counts after $k/2$ and k rounds are compared. Figure 3(b) plots the number of whitelist vocabulary words seen and a moving average of the number of rounds containing new unigrams. New unigrams are determined by comparing a round k with all rounds through $k-1$ and noting if any new words are seen. The shaded bands range from the LM’s unigram capacity to the size of the whitelist vocabulary.

3.3 Experiments

Since the whitelist vocabulary is uncased, capitalization normalization is applied based on an approach similar to Section 5. We then replace the unigram part of an n-gram model with this distribution to produce the final LM.

In A/B experiments, unigram models with different L1-clipping thresholds are compared against a baseline unigram model gathered from

| Model | acc@1 [%] | OOV rate [%] |
|----------------|------------------|------------------|
| baseline | 8.14 | 18.08 |
| $\lambda = 1$ | $+0.19 \pm 0.21$ | -1.33 ± 0.75 |
| $\lambda = 1K$ | $+0.11 \pm 0.24$ | -1.06 ± 0.66 |
| $\lambda = 5K$ | -0.08 ± 0.26 | -0.78 ± 0.93 |

Table 1: Relative change with L1-clipped unigrams on live traffic of en_US users on the virtual keyboard. Quoted 95% confidence intervals are derived using the jackknife method with user buckets.

centralized log data. Results are presented in Table 1. Accuracy is unchanged and OOV rate is improved at $\lambda = 1$ and $\lambda = 1K$.

Before we discuss methods to address inconsistent capitalization and data sparsity in morphologically rich languages, we review `SampleApprox`.

4 Review of `SampleApprox`

`SampleApprox`, proposed in Suresh et al. (2019a,b), can be used to approximate a RNN as a weighted finite automaton such as an n-gram model. A *weighted finite automaton* (WFA) $A = (\Sigma, Q, E, i, F)$ over \mathbb{R}_+ (probabilities) is given by a finite alphabet Σ (vocabulary words), a finite set of states Q (n-gram contexts), an initial state $i \in Q$ (sentence start state), a set of final states $F \in Q$ (sentence end states), and a set of labeled transitions E and associated weights that represent the conditional probability of labels (from Σ) given the state (list of n-grams and their probabilities). WFA models allow a special *backoff* label φ for succinct representation as follows. Let $L[q]$ be the set of labels on transitions from state q . For $x \in L[q]$, let $w_q[x]$, be the weight of the transition of x at state q and $d_q[x]$ be the destination state. For a label x and a state q ,

$$\begin{aligned} p(x|q) &= w_q[x] && \text{if } x \in L[q], \\ &= w_q[\varphi] \cdot p(x|d_q[\varphi]) && \text{otherwise.} \end{aligned}$$

In other words, φ is followed if $x \notin L[q]$. The definition above is consistent with that of backoff n-gram models (Chen and Goodman, 1999). Let $B(q)$ denote the set of states from which q can be reached by a path of backoff labels and let $q[x]$ be the first state at which label x can be read by following a backoff path from q .

Given an unweighted finite automaton A and a neural model, `SampleApprox` finds the proba-

bility model on A that minimizes the Kullback-Leibler (KL) divergence between the neural model and the WFA. The algorithm has two steps: a counting step and a KL minimization step. For the counting step, let $\bar{x}(1), \bar{x}(2), \dots, \bar{x}(k)$ be k independent samples from the neural model. For a sequence \bar{x} , let x_i denote the i^{th} label and $\bar{x}^i = x_1, x_2, \dots, x_i$ denote the first i labels. For every $q \in Q$ and $x \in \Sigma$, the algorithm computes $C(x, q)$ given by

$$\sum_{q' \in B(q)} \sum_{j=1}^m \sum_{i \geq 0} 1_{q(\bar{x}^i(j))=q', q=q'[x]} \cdot p_{\text{nn}}(x|\bar{x}^i(j)).$$

We illustrate this counting with an example. Suppose we are interested in the count of the bi-gram New York. Given a bi-gram LM, `SampleApprox` generates m sentences and computes

$$C(\text{York}, \text{New}) = \sum_{j: x_i(j)=\text{New}} p_{\text{nn}}(\text{York}|\bar{x}^i(j)).$$

In other words, it finds all sentences that have the word New, observes how frequently York appears subsequently, and computes the conditional probability. After counting, it uses a difference of convex (DC) programming based algorithm to find the KL minimum solution. If ℓ is the average number of words per sentence, the computational complexity of counting is $\tilde{O}(k \cdot \ell \cdot |\Sigma|)$ ¹ and the computational complexity of the KL minimization is $\tilde{O}(|E| + |Q|)$ per iteration of DC programming.

5 Capitalization

As mentioned in Section 2, users often type with incorrect capitalization. One way of handling incorrect capitalization is to store an on-device capitalization normalizer (Beaufays and Strope, 2013) to correctly capitalize sentences before using them to train the neural model. However, capitalization normalizers have large memory footprints and are not suitable for on-device applications. To overcome this, the neural model is first trained on uncased user data. `SampleApprox` is then modified to approximate cased n-gram models from uncased neural models.

As before, let $\bar{x}(1), \bar{x}(2), \dots, \bar{x}(k)$ be k independent (uncased) samples from the neural model. We capitalize them correctly at the server using Beaufays and Strope (2013). Let

¹ $a_n = \tilde{O}(b_n)$, means $a_n \leq b_n \cdot \text{poly} \log(n), \forall n \geq n_0$.

$\bar{y}(1), \bar{y}(2), \dots, \bar{y}(k)$ represent the corresponding k correctly capitalized samples. Let p_{cap} be another probability model on non-user data that approximates the ratio of uncased to cased probabilities given a context. Given a label y , let $u(y)$ be the uncased symbol. For example, if y is York, then $u(y)$ is york. With the above definitions, we modify the counting step of `SampleApprox` as follows:

$$\sum_{q' \in B(q)} \sum_{j=1}^m \sum_{i \geq 0} 1_{q(\bar{y}^i(j))=q', q=q'[y]} \cdot \tilde{p}(y|\bar{y}^i(j)),$$

where $\tilde{p}(y|\bar{y}^i(j))$ is given by

$$p_{\text{nn}}(u(y)|u(\bar{y}^i(j))) \cdot \frac{p_{\text{cap}}(y|\bar{y}^i(j))}{\sum_{y':u(y')=u(y)} p_{\text{cap}}(y'|\bar{y}^i(j))}.$$

We refer to this modified algorithm as `CapSampleApprox`. We note that word-piece to word approximation incurs an additional computation cost of $\tilde{O}((|E| + |Q| + |\Delta|)\ell)$, where Δ is the number of words, E and Q are the set of arcs and set of states in the word n-gram model, and ℓ is the maximum number of word-pieces per word.

6 Morphologically rich languages

To train neural models on morphologically rich languages, subword segments such as byte-pair encodings or word-pieces (Shibata et al., 1999; Schuster and Nakajima, 2012; Kudo, 2018) are typically used. This approach assigns conditional probabilities to subword segments, conditioned on prior subword segments. It has proved successful in the context of speech recognition (Chiu et al., 2018) and machine translation (Wu et al., 2016). Following these successes, we propose to train RNN LMs with word-pieces for morphologically rich languages.

We apply the word-piece approach of Kudo (2018), which computes a word-piece unigram LM using a word-piece inventory $\mathcal{V}_{\mathcal{P}}$. Each word-piece $x_i \in \mathcal{V}_{\mathcal{P}}$ is associated with a unigram probability $p(x_i)$. For a given word y and its possible segmentation candidates, the word is encoded with the segmentation that assigns the highest probability.

Throughout this paper we apply 4K, 16K, and 30K as the word-piece inventory sizes. These values lie within a range that provides good trade-off between the LSTM embedding size and the richness of the language morphology. We apply 100%

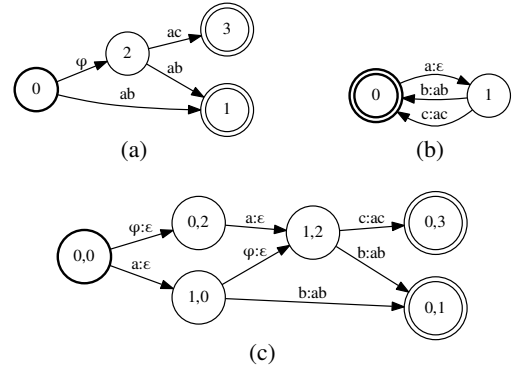


Figure 4: The (a) WFA A and WFSTs (b) T and (c) B for the word vocabulary $\{ab, ac\}$ and word-piece vocabulary $\{a, b, c\}$. Initial states are represented by bold circles and final states by double circles.

character coverage to include all the symbols that appeared in the unigram distribution (Section 3), including the common English letters, accented letters e.g. \acute{e} , \hat{o} , and digits. Accented letters are important for languages like Portuguese. For fast decoding, the n-gram models still need to be at the word-level, since word-piece n-gram models increase the depth of the beam-search during decoding. We convert the word n-gram topology to an equivalent word-piece WFA topology and use `SampleApprox` to approximate the neural word-piece model on the word-piece WFA topology. We then convert the resulting word-piece WFA LM to the equivalent n-gram LM. The remainder of this section outlines efficient algorithms for converting between word and word-piece WFA models.

A natural way to represent the transduction from word-piece sequences to word sequences is with a finite-state transducer. Given the properties of our word-piece representation, that transducer can be made sequential (i.e., input deterministic).

A *sequential weighted finite-state transducer* (WFST) is a deterministic WFA where each transition has an output label in addition to its (input) label and weight. We will denote by $o_q[x]$ the output label of the transition at state q with input label x , $o_q[x] \in \Delta \cup \{\epsilon\}$, where Δ denotes the output alphabet of the transducer and ϵ the empty string/sequence.

Let M be the minimal sequential (unweighted) finite-state transducer (FST) lexicon from word-piece sequences in Σ^* to word sequences in Δ^* , where Σ denotes our word-piece inventory, Δ denotes our vocabulary, and $*$ is Kleene closure. A word-piece topology B equivalent to the word

Algorithm 1 Approximating a Neural Model as an N-Gram with a Supplemental Topology.

```
Train  $R_W^u, R_P^u$  with FederatedAveraginga
Train  $A_W$  from supplemental corpus  $\mathbb{C}$ 
 $A_{W_e}, A_{W_i}, A_{W_m}, A_{W_r} \leftarrow \text{Gen}(R_W^u, A_W, \emptyset, \text{NN2WFA}_w)$ 
 $A_{P_e}, A_{P_i}, A_{P_m}, A_{P_r} \leftarrow \text{Gen}(R_P^u, A_W, A_{W_i}, \text{NN2WFA}_p)$ 

function Gen( $R^u, A_W, A_{W_i}$ , function NN2WFA)
   $A_e \leftarrow \text{NN2WFA}(R^u, A_W)$ 
  if NN2WFA == NN2WFAw then
     $A_i \leftarrow \text{NN2WFA}(R^u, A_W, \text{self\_infer}=\text{true})$ 
  else
     $A_i \leftarrow \text{NN2WFA}(R^u, A_{W_i})$ 
  end if
   $A_m \leftarrow \text{Interpolate}(A_e, A_i)$ 
```

^a T denotes an unweighted topology and A denotes the weighted n-gram model. Superscript u represents uncased models.

```
 $A_r \leftarrow \text{NN2WFA}(R^u, A_m)$ 
return  $A_e, A_i, A_m, A_r$ 
end function
function NN2WFAw( $R_W^u, A_W, \text{self\_infer}=\text{false}$ )
  if self_infer then
    return CapSampleApprox( $R_W^u, \emptyset, A_W$ )
  else
    return CapSampleApprox( $R_W^u, A_W, A_W$ )
  end if
end function
function NN2WFAp( $R_P^u, A_W$ )
   $T_W^u \leftarrow \text{ConvertToLowercaseTopology}(A_W)$ 
   $T_P^u \leftarrow \text{ConvertToWordPieceTopology}(T_W^u)$ 
   $A_P^u \leftarrow \text{SampleApprox}(R_P^u, T_P^u)$ 
   $A_W^u \leftarrow \text{ConvertToWordTopology}(A_P^u)$ 
  return CapSampleApprox( $A_W^u, A_W, A_W$ )
end function
```

topology A can be obtained by composing the word-piece-to-word transducer M with A :

$$B = M \circ A.$$

Since A has backoff transitions, the generic composition algorithm of (Allauzen et al., 2011) is used with a custom composition filter that ensures the result, B , is deterministic with a well-formed backoff structure, and hence is suitable for the counting step of SampleApprox. We give an explicit description of the construction of B , from which readers familiar with Allauzen et al. (2011) can infer the form of the custom composition filter.

The states in B are pairs (q_1, q_2) , with $q_1 \in Q_M$ and $q_2 \in Q_A$, initial state $i_B = (i_M, i_A)$, and final state $f_B = (f_M, f_A)$. Given a state $(q_1, q_2) \in Q_B$, the outgoing transitions and their destination states are defined as follows. If $x \in L[q_1]$, then an x -labeled transition is created if one of two conditions holds:

1. if $o_{q_1}[x] \in L[q_2]$, then

$$d_{(q_1, q_2)}[x] = (d_{q_1}[x], d_{q_2}[o_{q_1}[x]]) \text{ and} \\ o_{(q_1, q_2)}[x] = o_{q_1}[x];$$

2. if $o_{q_1}[x] = \epsilon$ and $R[d_{q_1}[x]] \cap L[q_2] \neq \emptyset$, then

$$d_{(q_1, q_2)}[x] = (d_{q_1}[x], d_{q_2}[o_{q_1}[x]]) \text{ and} \\ o_{(q_1, q_2)}[x] = \epsilon$$

where $R[q]$ denotes the set of output non- ϵ labels that can be emitted after following an output- ϵ path from q . Finally if $\varphi \in L[q_1]$, a backoff transition is created:

$$d_{(q_1, q_2)}[\varphi] = (q_1, d_{q_2}[\varphi]) \text{ and } o_{q_1, q_2}[\varphi] = \epsilon.$$

The counting step of SampleApprox is applied to B , and transfers the computed counts from B to A by relying on the following key property of M . For every word y in Δ , there exists a unique state $q_y \in Q_M$ and unique word-piece x_y in Σ such that $o_{q_y}[x_y] = y$. This allows us to transfer the counts from B to A as follows:

$$w_q[y] = w_{(q_y, q)}[x_y]$$

The KL minimization step of SampleApprox to A is applied subsequently.

As an alternative, the unweighted word automaton A could be used to perform the counting step directly. Each sample $\bar{x}(j)$ could be mapped to a corresponding word sequence $\bar{y}(j)$, mapping out-of-vocabulary word-piece sequences to an unknown token. However, the counting steps would have become much more computationally expensive, since $p_{\text{nn}}(y|\bar{y}^i(j))$ would have to be evaluated for all i, j and for all words y in the vocabulary, where p_{nn} is now a word-piece RNN.

7 Experiments

7.1 Neural language model

LSTM models (Hochreiter and Schmidhuber, 1997) have been successfully used in a variety of sequence processing tasks. LSTM models usually have a large number of parameters and are not suitable for on-device learning. In this work, we use various techniques to reduce the memory footprint and to improve model performance.

We use a variant of LSTM with a Coupled Input and Forget Gate (CIFG) (Greff et al., 2017) for the federated neural language model. CIFG couples

| Model | N_l | N_h | N_e | S_e | S_{total} |
|--------------------|-------|-------|-------|-------|-------------|
| W _{30K} | 1 | 670 | 96 | 2.91M | 3.40M |
| P _{4K-S} | 1 | 670 | 96 | 0.38M | 0.85M |
| P _{4K-L} | 2 | 1080 | 140 | 0.56M | 2.70M |
| P _{4K-G} | 2 | 1080 | 280 | 1.12M | 2.71M |
| P _{16K-S} | 1 | 670 | 96 | 1.54M | 2.00M |
| P _{16K-L} | 1 | 670 | 160 | 2.56M | 3.33M |
| P _{30K} | 1 | 670 | 96 | 2.91M | 3.40M |

Table 2: Parameters for neural language models. W and P refer to word and word-piece models, respectively. N_l , N_h , N_e , S_e and S_{total} refer to the number of LSTM layers, the number of hidden states in LSTM, the embedding dimension size, the number of parameters in the embedding layer and in total, respectively. The suffixes ‘‘S’’ and ‘‘L’’ indicate small and large models. ‘‘G’’ represents GLSTM. The suffixes 4K, 16K and 30K represent the vocabulary sizes.

the forget and input decisions together, which reduces the number of LSTM parameters by 25%. We also use group-LSTM (GLSTM) (Kuchaiev and Ginsburg, 2017) to reduce the number of trainable variables of an LSTM matrix by the number of feature groups, k . We set $k = 5$ in experiments. Table 2 lists the parameter settings of the word (W) and word-piece (P) models used in this study. Due to the memory limitations of on-device training, all models use fewer than 3.5M parameters. For each vocabulary size, we first start with a base architecture consisting of one LSTM layer, a 96-dimensional embedding, and 670 hidden state units. We then attempt to increase the representational power of the LSTM cell by increasing the number of hidden units and using multi-layer LSTM cells (Sutskever et al., 2014). Residual LSTM (Kim et al., 2017) and layer normalization (Lei Ba et al., 2016) are used throughout experiments, as these techniques were observed to improve convergence. To avoid the restriction that $N_h = N_e$ in the output, we apply a projection step at the output gate of the LSTM (Sak et al., 2014). This step reduces the dimension of the LSTM hidden state from N_h to N_e . We also share the embedding matrix between the input embedding and output softmax layer, which reduces the memory requirement by $|\mathcal{V}| \times N_e$. We note that other recurrent neural models such as *gated recurrent units* (Chung et al., 2014) can also be used instead of CIFG LSTMs.

The federated RNN LMs are trained on two language settings of the virtual keyboard: Amer-

ican English (en_US) and Brazilian Portuguese (pt_BR). Following McMahan et al. (2017), 500 reporting clients are used to compute the gradient updates for each round. A server-side learning rate of 1.0, a client-side learning rate of 0.5, and Nesterov momentum of 0.9 are used. Both the word and word-piece models are trained over the same time range and with the same hyperparameters. Prior to federated training of the RNN LM, the word-piece inventory is constructed from the unigram distribution collected via the federated approach introduced in Section 3.

A common evaluation metric for both word and word-piece models is desirable during federated training. Such a metric can be used to monitor the training status and select models to be used for the CapSampleApprox algorithm. Neither cross-entropy nor accuracy serves this need due to the mismatch in vocabularies used. Word-level accuracy is hard to compute for the word-piece model, since it requires hundreds of inference calls to traverse all combinations of a word from the word-piece vocabulary. In this study, we apply sentence log likelihood (SLL) in the evaluation. Given a sentence $\bar{x}^m = \{x_1, x_2, \dots, x_m\}$ composed of m units (either words or word-pieces), SLL is evaluated as $\sum_{i=1}^m \log(p_{nn}(x_i|\bar{x}^{i-1}))$. One issue that arises is the handling of out-of-vocabulary (OOV) words. The OOV probability of the word model is about 8%. The comparable probability of an OOV word (according to \mathcal{V}) for word-piece models is the product of the corresponding word-piece conditional probabilities, which is much smaller than 8%. To mitigate this issue, we define SLL excluding OOV as:

$$SLL^e = \sum_{i:x_i \neq \text{OOV}}^m \log(p_{nn}(x_i|\bar{x}^{i-1})),$$

where the OOV in the equation includes word-pieces that are components of OOV words. In the following, SLL^e is used as model selection metric.

7.2 Approximated n-gram model

Algorithm 1 illustrates the workflow we use to generate different n-gram models for evaluation. Recall that CapSampleApprox takes a RNN LM, an n-gram topology, and a reweighting FST for capitalization normalization. The n-gram topology is empty under self-inference mode. Suresh et al. (2019a) showed that inferring topology from the RNN LM does not perform as well as

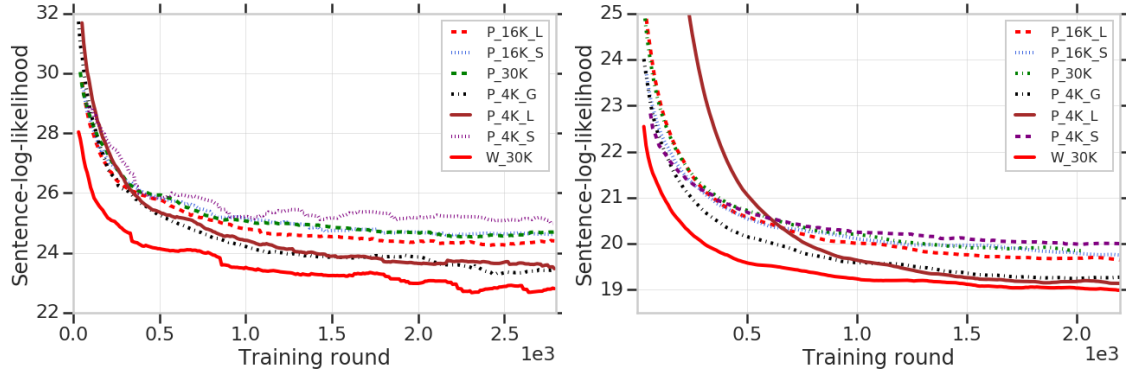


Figure 5: Sentence log likelihood excluding OOV token for en_US (left) and pt_BR (right).

| Model | en_US | pt_BR |
|-----------|--------------------|-------------------|
| Baseline | 10.03% | 8.55% |
| A_{W_e} | $10.52 \pm 0.03\%$ | $9.66 \pm 0.02\%$ |
| A_{W_i} | $10.47 \pm 0.02\%$ | $9.67 \pm 0.02\%$ |
| A_{W_m} | $10.27 \pm 0.03\%$ | $9.40 \pm 0.02\%$ |
| A_{W_r} | $10.49 \pm 0.03\%$ | $9.65 \pm 0.02\%$ |

Table 3: Result of top-1 prediction accuracy on the live traffic of the virtual keyboard for en_US and pt_BR populations. Quoted 95% confidence intervals for federated models are derived using the jackknife method.

| Model | top-1 |
|-----------|--------------------|
| Baseline | 10.03% |
| A_{P_e} | $10.49 \pm 0.03\%$ |
| A_{P_i} | $10.46 \pm 0.03\%$ |
| A_{P_m} | $10.48 \pm 0.04\%$ |
| A_{P_r} | $10.53 \pm 0.03\%$ |

Table 4: Result of top-1 prediction accuracy on the live traffic of the virtual keyboard for en_US derived using word-piece models.

using the true n-gram topology obtained from the training corpus. Hence, we supplement the neural-inferred topology with the topology obtained by a large external large corpus denoted by A_W . We use `CapSampleApprox` on four topologies and compare the resulting models: an n-gram model obtained from an external corpus’s topology A_e , an n-gram model obtained from a neural inferred topology A_i , an n-gram model obtained by interpolating (merging) the two models above A_m , and an n-gram model obtained by approximating on the interpolated topology A_r . We repeat this experiment for both word and word-piece RNN LMs and use subscripts W and P , respectively. We evaluate all eight produced n-gram models directly on the traffic of a production virtual keyboard, where prediction accuracy is evaluated over user-typed words.

7.3 Results

Figure 5 shows the SLL^e metric for all the experiments listed in Table 2. In general, larger models generate better results than smaller baseline models. For the baseline architectures with same RNN size, having a larger vocabulary leads to some gains. For the larger architectures that have similar

total numbers of parameters, 4K word-piece models are shown to be superior to 16K and 30K. For 4K word-piece models, GLSTM is in general on-par with its P_{4K-L} counterpart. The word model is better than all the word-piece models in both languages in SLL^e . We were surprised by this result, and hypothesize that it is due to the SLL^e metric discounting word-piece models’ ability to model the semantics of OOV words. The solid lines are the best models we pick for A/B experiment evaluation for the virtual keyboard (P_{4K-L} and W_{30K}).

Table 3 shows the A/B evaluation result on both en_US and pt_BR populations. The baseline model is an n-gram model trained directly from centralized logs. All of the federated trained models perform better than the baseline model. We repeated the A/B evaluation with word-piece models on en_US and the results are in Table 4. The performance of word-piece models is similar to that of word models. Among the federated models for en_US, A_{P_r} has the best result. This meets our expectation that the supplemental corpus helps improve the performance of the topology inferred from the RNN LM.

8 Conclusion

We have proposed methods to train production-quality n-gram language models using federated learning, which allows training models without user-typed text ever leaving devices. The proposed methods are shown to perform better than traditional server-based algorithms in A/B experiments on real users of a virtual keyboard.

Acknowledgments

The authors would like to thank colleagues in Google Research for providing the federated learning framework and for many helpful discussions.

References

- Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM.
- Naman Agarwal, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and Brendan McMahan. 2018. [cpsgd: Communication-efficient and differentially-private distributed sgd](#). In *Neural Information Processing Systems*.
- Cyril Allauzen, Michael Riley, and Johan Schalkwyk. 2011. A filter-based algorithm for efficient composition of finite-state transducers. *International Journal of Foundations of Computer Science*, 22(8):1781–1795.
- Ouais Alsharif, Tom Ouyang, Françoise Beaufays, Shumin Zhai, Thomas Breuel, and Johan Schalkwyk. 2015. Long short term memory neural network for keyboard gesture decoding. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2076–2080.
- Françoise Beaufays and Brian Strobe. 2013. Language model capitalization. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6749–6752. IEEE.
- Bhaswar Bhattacharya and Gregory Valiant. 2015. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems 28*. NIPS.
- Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. [Practical secure aggregation for privacy-preserving machine learning](#). In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17*, pages 1175–1191, New York, NY, USA. ACM.
- Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.
- Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjali Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.
- Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2017. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.
- Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. [Federated learning for mobile keyboard prediction](#). *CoRR*, abs/1811.03604.
- Lars Hellsten, Brian Roark, Prasoon Goyal, Cyril Allauzen, Françoise Beaufays, Tom Ouyang, Michael Riley, and David Rybach. 2017. [Transliterated mobile keyboard input via weighted finite-state transducers](#). In *Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing (FSMNL)*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. 2017. [Residual LSTM: design of a deep recurrent architecture for distant speech recognition](#). In *InterSpeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 1591–1595.
- Jakub Konečný, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.
- Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. 2016. [Federated learning: Strategies for improving communication efficiency](#). In *NIPS Workshop on Private Multi-Party Machine Learning*.

- Oleksii Kuchaiev and Boris Ginsburg. 2017. [Factorization tricks for LSTM networks](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.
- Taku Kudo. 2018. [Subword regularization: Improving neural network translation models with multiple subword candidates](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. [Communication-efficient learning of deep networks from decentralized data](#). In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pages 1273–1282.
- Brendan McMahan and Daniel Ramage. 2017. Federated learning: Collaborative machine learning without centralized training data. <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>.
- Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. [Learning differentially private recurrent language models](#). In *International Conference on Learning Representations (ICLR)*.
- Robert C. Moore and William Lewis. 2010. [Intelligent selection of language model training data](#). In *Proceedings of the ACL 2010 Conference Short Papers, ACLShort '10*, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tom Ouyang, David Rybach, Françoise Beaufays, and Michael Riley. 2017. [Mobile keyboard input decoding with finite-state transducers](#). *CoRR*, abs/1704.03987.
- Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*.
- Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.
- Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical report, Technical Report DOI-TR-161, Department of Informatics, Kyushu University.
- Ananda Theertha Suresh, Michael Riley, Brian Roark, and Vlad Schogol. 2019a. Approximating probabilistic models as weighted finite automata. *CoRR*, abs/1905.08701.
- Ananda Theertha Suresh, Brian Roark, Michael Riley, and Vlad Schogol. 2019b. Distilling weighted finite automata from arbitrary probabilistic models. In *Proceedings of the 14th International Conference on Finite State Methods and Natural Language Processing (FSMNL 2019)*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.