

Federated Learning for Emoji Prediction in a Mobile Keyboard

Swaroop Ramaswamy Rajiv Mathews Kanishka Rao Françoise Beaufays

Google LLC

Mountain View, CA, U.S.A.

{swaroopram, mathews, kanishkarao, fsb}@google.com

Abstract

We show that a word-level recurrent neural network can predict emoji from text typed on a mobile keyboard. We demonstrate the usefulness of transfer learning for predicting emoji by pretraining the model using a language modeling task. We also propose mechanisms to trigger emoji and tune the diversity of candidates. The model is trained using a distributed on-device learning framework called federated learning. The federated model is shown to achieve better performance than a server-trained model. This work demonstrates the feasibility of using federated learning to train production-quality models for natural language understanding tasks while keeping users' data on their devices.

1 Introduction

Emoji have become an important mode of expression on smartphones as users increasingly use them to communicate on social media and chat applications. Easily accessible emoji suggestions have therefore become an important feature for mobile keyboards.

Gboard is a mobile keyboard with more than 1 billion installs and support for over 600 language varieties. With this work, we provide a mechanism by which Gboard offers emoji as predictions based on the text previously typed, as shown in Figure 1.

Mobile devices are constrained by both memory and CPU. Low-latency is also required, since users typically expect a keyboard response within 20 ms of an input event (Hellsten et al., 2017).

A unidirectional recurrent neural network architecture (RNN) is used in this work. Since forward RNNs only include dependencies backwards in time, the model state can be cached at each timestep during inference to reduce latency.

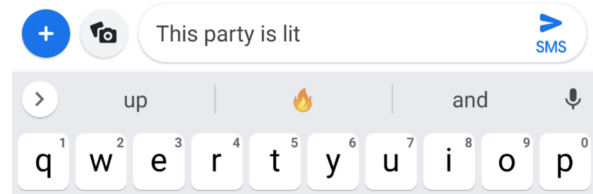


Figure 1: Emoji predictions in Gboard. Based on the context “This party is lit”, Gboard predicts both emoji and words.

2 Federated Learning

Federated Learning (FL) (Bonawitz et al., 2019) is a new computation paradigm in which data is kept on users' devices and never collected centrally. Instead, minimal and focused model updates are transmitted to the server. This allows us to train models while keeping users' data on their devices. FL can be combined with other privacy-preserving techniques like secure multi-party computation (Bonawitz et al., 2017) and differential privacy (McMahan et al., 2018; Agarwal et al., 2018; Abadi et al., 2016). FL has been shown to be robust to unbalanced and non-IID data.

We use the FederatedAveraging algorithm presented in McMahan et al. (2017) to aggregate client updates after each round of local, on-device training to produce a new global model. At training round t , a global model with parameters w_t , is sent to K devices selected from the device population. Each device has a local dataset P_k which is split into batches of size B . Stochastic gradient descent (SGD) is used on the clients to compute new model parameters w_k^{t+1} . These client weights are then averaged across devices, on the server, to compute the new model parameters w_{t+1} .

3 Method

3.1 Network architecture

The Long-Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) architecture has been shown to achieve state-of-the-art performance of a number of sentiment prediction and language modeling tasks (Radford et al., 2017).

We use an LSTM variant called the Coupled Input and Forget Gate (CIFG) (Greff et al., 2017). As with Gated Recurrent Units (Cho et al., 2014), the CIFG uses a single gate to control both the input and recurrent cell self-connections. The input gate (i) and the forget gate (f) are related by $f = 1 - i$. This coupling reduces the number of parameters per cell by 25%, compared to an LSTM.

We use an input word vocabulary size of 10,000, an input embedding size of 96, and a two-layer CIFG with 256 units per layer. The logits are passed through a softmax layer to predict probabilities over 100 emoji.

3.2 Pretraining

Howard and Ruder (2018) demonstrated that pre-training parameters on a language modeling task can improve performance on other tasks.

We pretrain all layers except the output projection layer, using a language model trained to predict the next word in a sentence. For the output projection, we reuse the input embeddings. This type of sharing of input and output embeddings has been shown to improve performance of language models (Press and Wolf, 2017). Pretraining is done with federated learning using techniques similar to those described by Hard et al. (2018). The language model achieves an Accuracy@1 of 13.7%, on the same vocabulary. Pretraining with a language model task leads to much faster convergence for the emoji model, as seen in Figure 2.

3.3 Triggering

In addition to predicting the correct emoji, a triggering mechanism must determine when to show emoji predictions to users. For instance, a user is likely to type 🎉 after typing “Congrats” or “Congrats to you” but not after “Congrats to”.

One way to handle this would be to use a single language model that can predict both words and emojis. However, we want to separate the task of predicting relevant emoji from that of deciding how much we wanted emoji to trigger, since the latter is more of a product decision, rather than a technical

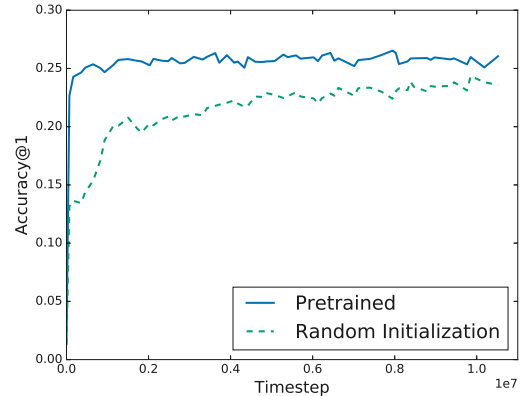


Figure 2: Accuracy@1 vs. training step with and without pretraining, using server-based evaluations.

challenge. For instance, if we want to allow users to control how often emoji predictions are offered, it’s easier to do with a separate model.

Another way to handle triggering is to use a separate binary classification model that predicts the likelihood of the user typing any emoji after a given phrase. However, using a separate model for triggering leads to additional overhead in terms of memory and latency. Instead, we adjust the softmax layer of the model to predict over N emoji and an additional unknown token $\langle \text{UNK} \rangle$ class. The $\langle \text{UNK} \rangle$ class is set as the target output for inputs without emoji labels. At inference, we show the predictions from the model only if the probability of the $\langle \text{UNK} \rangle$ class is less than a certain threshold.

During training, sentences without emoji are truncated to a random length in the range $[1, \text{length of sentence}]$. Truncation allows the model to learn to not predict emoji after conjunctions, prepositions etc. which typically occur in the middle of sentences.

3.4 Diversification

The distribution of emoji usage frequency is very light-tailed as seen in Figure 3. As a result, the top predictions from the model are almost always the most frequent emoji regardless of the input context. To overcome this, the probability of each emoji (\hat{P}) is scaled by the empirical probability of that emoji (P) in the training data as follows.

$$S_i = \frac{\hat{P}(\text{emoji} = i | \text{text})}{P(\text{emoji} = i)^\alpha} \quad (1)$$

where α is a scaling factor, determined empirically through experiments on live traffic. Setting

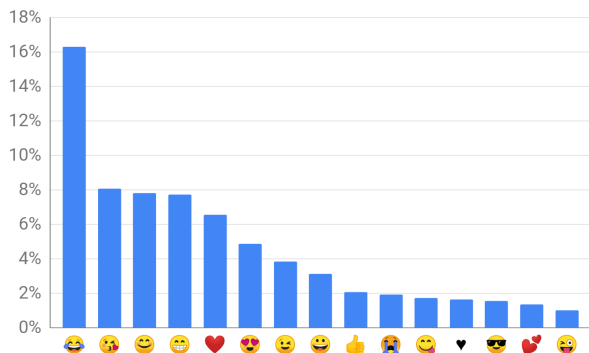


Figure 3: Distribution of 15 most frequently used emoji in English (US).

Context	$\alpha = 0.0$	$\alpha = 0.7$
Sorry I ended up falling asleep	😄	😞
Good morning sunshine	😂	☀️
Coz I miss you xx	😘	😭
I'm so sorry sweetie	😞	💔
Hey girl you take it easy	😄	😏
not sure what happened to that	😄	😞

Table 1: Examples of emoji predictions with and without diversification

α to 0 removes diversification. Table 1 provides examples with and without diversification.

4 Server-based Training

Server-based training of models is done on data logged from Gboard users who have opted to periodically share anonymized snippets of text typed in selected apps. All personally identifiable information is stripped from these logs. The logs are filtered further to only include sentences that are labeled as English with high confidence by a language detection model (Botha et al., 2017; Zhang et al., 2018). The subset of logs used for training contain approximately 370 million snippets, approximately 11 million of which contain emoji. Hyperparameters for server-based training are optimized using a black-box optimization technique (Golovin et al., 2017).

5 Federated Training

The data used for federated training is stored in local caches on client devices. For a device to participate in training, it must have at least 2 GB of RAM, must be located in United States or Canada, and must be using English (US) as the primary lan-

guage. In addition, only devices that are connected to un-metered networks, idle, and charging are eligible for participation at any given time. On average, each client has approximately 400 sentences. The model is trained for one epoch on each client, in each round. The model typically converges after 2000 training rounds.

In federated training, there is no explicit split of data into train and eval samples. Instead, a separate evaluation task runs on a different subset of client devices in parallel to the training task. The eval task uses model checkpoints generated by the federated training task during a 24-hour period and aggregates the metrics across evaluation rounds.

6 Evaluation

Model quality is evaluated using Accuracy@1, defined as the ratio of accurate top-1 emoji predictions to the total number of examples containing emoji. Area Under ROC Curve (AUC) is used to evaluate the quality of the triggering mechanism. Computing the AUC involves numerical integration and is not straightforward to do in the FL setting. Therefore, we report AUC only on logs data that is collected on the server. All evaluation metrics are computed prior to diversification.

7 Federated Experiments

In FL, the contents of the client caches are constantly changing as old entries are cleared and replaced by new activity. Since these experiments were conducted non-concurrently, the client cache contents are different and therefore numbers cannot be compared across experiments. We conduct experiments to study the effect of client batch size (B), devices per round (K) and server optimizer configuration on model quality. We then take the best model and compare it with a server trained model. The results are summarized in Table 2.

Because of the sparsity of sentences containing emoji in the client caches, the model quality is improved to a large degree by using large client batch sizes. This is not entirely surprising, since gradient updates are more accurate with larger batch sizes (Smith et al., 2018). This is particularly true when the target classes are heavily imbalanced.

The accuracy of the model also increases with the number of devices per round but there are diminishing returns beyond $K = 500$.

We experimented with various optimizers for the server update after each round of federated train-

<i>Experiment</i>	<i>Accuracy@1</i>	<i>AUC</i>
$B = 1$	0.008	0.513
$B = 10$	0.037	0.500
$B = 50$	0.240	0.837
$B = 200$	0.253	0.863
$K = 20$	0.239	0.846
$K = 50$	0.242	0.852
$K = 200$	0.253	0.867
$K = 500$	0.255	0.863
SGD, $\eta_s = 1.0$	0.236	0.850
SGD, $\eta_s = 2.0$	0.245	0.856
Momentum, $\eta_s = 1.0$	0.247	0.856
Best federated	0.256	0.863
Best server trained	0.239	0.898

Table 2: The results from federated experiments. All numbers reported are after 2000 training rounds. η_s refers to the learning rate used on the server for applying the update aggregated across users in each round.

ing and found that using momentum of 0.9 with Nesterov accelerated gradients (Sutskever et al., 2013) gives significant benefits over using SGD, both in terms of speed of convergence and model performance.

The best federated model, which runs in production, uses $B = 1000, K = 1000$, and is trained with momentum. We assign a weight of 0 to 99% of the <UNK> examples at training time so as to balance the triggering and emoji prediction losses. We ran federated evaluation tasks of the best server-trained model on the client caches in order to fairly compare the two training approaches. The federated model achieved better Accuracy@1 in the federated evaluation, as shown in Figure 4. However, the AUC achieved by the federated model is lower than that of the server trained model.

AUC is only computed on the logs collected on the server. These logs are restricted to short snippets of text typed in selected apps, therefore the data is not believed to be as representative of the text typed by users as data that resides on the client caches. The lower AUC of the federated model is likely because of this bias.

8 Live experiment

At inference time, we use a quantized TensorFlow Lite (TFLite) model format. The average inference latency is around 1 ms.

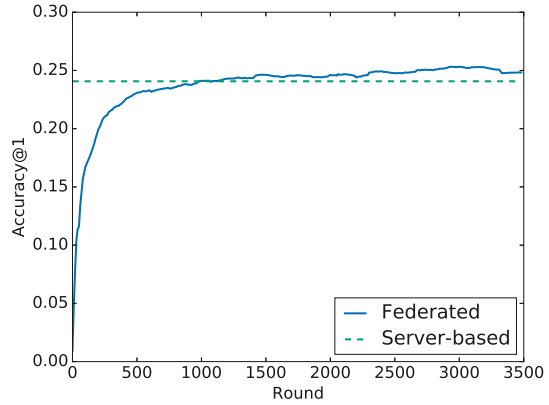


Figure 4: Evaluation Accuracy@1 vs. Round for federated and server trained models.

<i>Metric</i>	<i>Relative change [%]</i>	
	<i>Server trained</i>	<i>Federated</i>
Prediction CTR	3.61 ± 1.00	3.66 ± 0.95
Emoji Shares	3.63 ± 0.99	5.54 ± 1.19
Emoji DAU	9.57 ± 0.39	11.22 ± 0.48

Table 3: Relative changes to metrics as a result of the server trained and federated emoji prediction models, measured in experiments on live user traffic. The baseline does not have any emoji predictions. Quoted 95% confidence interval errors for all results are derived using the jackknife method with user buckets.

We ran a live-traffic experiment for users in USA and Canada typing in English (US). We observed that both the federated and the server trained model lead to significant increases in the overall click-through rate (CTR) of predictions, total emoji shares, and daily active users (DAU) of emoji (see Table 3). We also observed that the federated model did better than the server trained model on all of the metrics.

Given that emoji are triggered rarely, the increase in CTR is quite large, for both the models.

9 Conclusions

In this paper, we train an emoji prediction model using a CIFG-LSTM network. We demonstrate that this model can be trained using FL to achieve better performance than a server trained model. This work builds on previous practical applications of federated learning in Yang et al. (2018); Hard et al. (2018); Bonawitz et al. (2019). We show that FL works even with sparse data and poorly balanced classes.

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*.
- Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konecny, Stefano Mazzocchi, H Brendan McMahan, et al. 2019. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*.
- 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.
- Jan A. Botha, Emily Pitler, Ji Ma, Anton Bakalov, Alex Salcianu, David I Weiss, Ryan T. McDonald, and Slav Petrov. 2017. Natural language processing with small feed-forward networks. In *EMNLP*.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using RNN encoder–decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Elliot Karro, and D. Sculley, editors. 2017. *Google Vizier: A Service for Black-Box Optimization*.
- Klaus Greff, Rupesh Kumar Srivastava, Jan Koutnı́k, Bas R. Steunebrink, and Jürgen Schmidhuber. 2017. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2222–2232.
- 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 Comput.*, 9(8):1735–1780.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 328–339.
- 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, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. [Learning differentially private recurrent language models](#). In *International Conference on Learning Representations (ICLR)*.
- Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *EACL*.
- Alec Radford, Rafal Józefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444.
- Samuel L. Smith, Pieter-Jan Kindermans, and Quoc V. Le. 2018. [Don't decay the learning rate, increase the batch size](#). In *International Conference on Learning Representations*.
- Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. [On the importance of initialization and momentum in deep learning](#). In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1139–1147, Atlanta, Georgia, USA.
- TFLite. Tensorflow lite, TensorFlow's solution for running machine learning models on mobile and embedded devices. <https://www.tensorflow.org/lite>. Accessed: 2019-01-16.
- Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestion. *arXiv preprint arXiv:1812.02903*.
- Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David I Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *EMNLP*.