Tuplemax Loss for Language Identification

Li Wan, Prashant Sridhar, Yang Yu, Quan Wang, Ignacio Lopez Moreno
Google Inc., USA
{liwan,psridhar,yyuyy,quanw,elnota}@google.com

Our Mission

Identify the language from variable-length speech, given a user-specified set of candidate languages.

Training

A neural network is trained on batches of fixed-length input segments

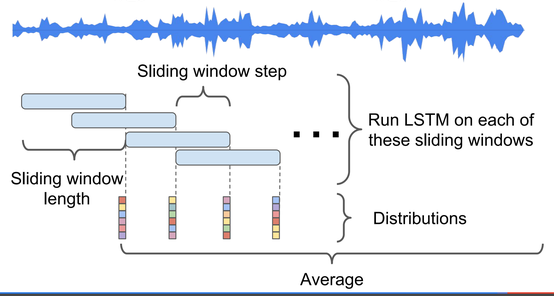
Table 1. LSTM network architecture

Index	Input	Output	Specification
0	40 x 400	80 x 200	Frame concatenation
1	80 x 200	256 x 200	LSTM(1024, 256)
2	256 x 200	256 x 200	LSTM(768, 256)
3	256 x 200	256 x 200	LSTM(512, 256)
4	256 x 200	256 x 200	LSTM(256, -)
5	256 x 200	256	Last frame output
6	256	256	ReLU activation
7	256	79	Linear mapping

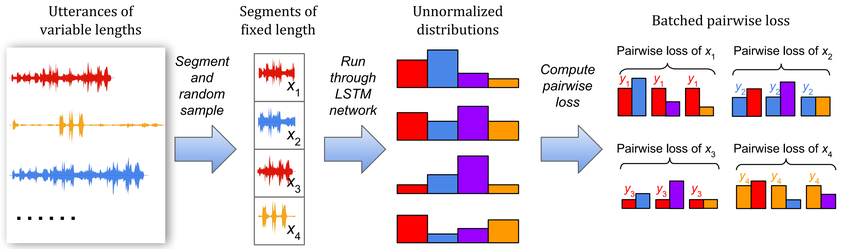
Inference

We perform inference on sliding windows of fixed-length segments, and average the outputs. Assume S is set of candidate languages, x is features, z is output probabilities, y is truth:

$$y^* = \arg\max_{k \in S} \mathbf{E}_t[f(x^t; w)] = \arg\max_{k \in S} \mathbf{E}_t[z_k^t]$$



System Overview



Practical Example

Two distributions over 4 languages. Second is better because it gives us the correct label. It also has smaller pairwise loss.

Ground truth	Distribution	Softmax cross-entropy loss	Pairwise loss
[1, 0, 0, 0]	[0.3, **0.4**, 0.2, 0.1]	-log(0.3)	0.6623
[1, 0, 0, 0]	[**0.3**, 0.25, 0.25, 0.2]	-log(0.3)	0.6604

Loss Functions

Softmax loss:

$$L(y, z) = \log \sum_{k=1}^{N} \exp(z_k) - z_y$$

Pairwise loss:

$$L(y, z) = \mathbf{E}_{k \neq y}[\log(\exp(z_y) + \exp(z_k))] - z_y$$

Tuplemax loss:

$$L(y, z) = \mathbf{E}_{s^n \sim D}[L^n(y, z)] = \sum_{n=2}^{N} p_n L^n(y, z)$$

$$L^n(y, z) = \mathbf{E}_{S_y^n}[\log \sum_{k \in S_y^n} \exp(z_k)] - z_y$$

Experiments

Training set: 79 languages, 1~60M anonymised utterances each.
Evaluation set: 20K utterances per language.
Evaluation setup:

- Tuple size = 2 to mimic real traffic.

Table 2. Classification Error Rate (%) for softmax and tuplemax.

Loss Function	All Pairs	Top 87 Pairs	Checkpoint Type
Softmax	4.50	11.1	Last
Softmax	3.85	9.14	Average
Softmax	2.40	5.50	Best on Test
Tuplemax	**2.33**	**4.55**	Last

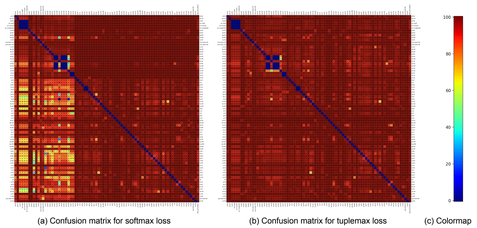


Fig. Confusion matrix for softmax (left) and tuplemax (right).
E.g. E[j, i] = 90% means 90% of utterances with ground truth label j and user preference {j, i} are correctly identified.

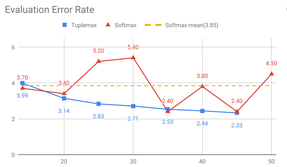


Fig. **Evaluation error**: tuplemax vs. softmax.

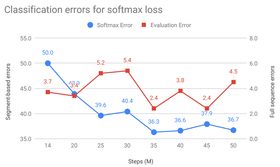


Fig. **Softmax**: training loss vs. evaluation error.

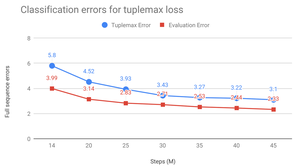


Fig. **Tuplemax**: training loss vs. evaluation error.

Conclusions

1. Tuplemax produces better and more balanced results.
2. Convergence is significantly more stable.

ICASSP 2019
The 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
May 12 - 17, 2019, Brighton, United Kingdom
IEEE Signal Processing Society
IEEE