Compact Kernel Models for Acoustic Modeling via Random Feature Selection

Avner May*, Michael Collins*, Daniel Hsu*, Brian Kingsbury†
*Columbia University Computer Science Department, †IBM TJ Watson Research Center

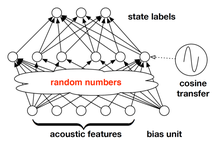
Main Idea

- Kernel approximation methods often require very many random features for good performance.
- We propose a simple feature selection method for reducing the number of required features, and show it is effective in acoustic modeling.
- We perform comparisons to DNNs, and achieve comparable WER results on Broadcast News (50 hour) dataset.

Background and Motivation

Kernel Approximation

- $K(x_i, x_j) \approx z(x_i)^T z(x_j)$
- Random Fourier Features [1]:
  $z_i(x) = \sqrt{2/D}\cos(w_i^T x + b_i)$
  RBF: $w_i \sim Normal, b_i \sim Unif[0,2\pi]$



Acoustic Modeling

- $p(y \mid x\,; \theta) = \exp\left(\theta_y^T z(x)\right) / \sum_{\hat{y}} \exp\left(\theta_{\hat{y}}^T z(x)\right)$
- Maximizing regularized log-likelihood is convex.
- Similar to single hidden-layer neural net, with cosine activation function, and random first layer weights.

[1] A. Rahimi and B. Recht, "Random Features for Large-Scale Kernel Machines," in Proceedings of NIPS, 2007

Datasets Used

Dataset	# Train	# Heldout	# Test	# Classes
Cantonese	7.7M	1M	7M	1000
Bengali	7.5M	0.9M	7.2M	1000
Broadcast News: 50 hour	16.2M	1.8M	N/A	5000

Kernels Used

Kernel	Formula for $K(x, y)$	Dist. for W
Laplacian	$\exp\left(-\lambda\|\|x-y\|\|_1^2\right)$	Cauchy
Gaussian	$\exp\left(-\frac{1}{2\sigma^2}\|\|x-y\|\|_2^2\right)$	Normal
"K-Sparse Gaussian"	$\sum_{F\subseteq[d]\,:\,\|F\|=k} \exp\left(-\frac{1}{2\sigma^2}\|\|x_F - y_F\|\|_2^2\right)$	"Sparse Normal"

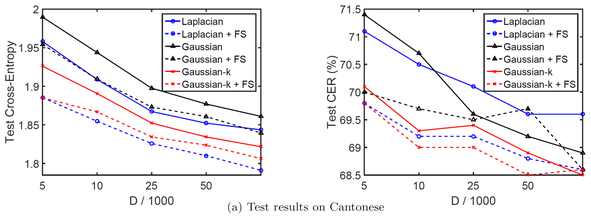
Feature Selection Algorithm

input Target number of random features D,
selection schedule $0 = s_0 < s_1 < \cdots < s_T = D$.
initialize feature pool $P := \emptyset$.
for $t = 1, 2, \ldots, T$ **do**

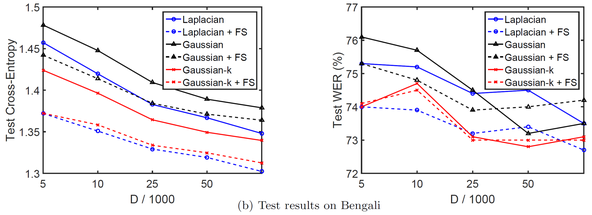
- Generate $D - s_{t-1}$ new random features, and add them to P.
- Learn weights $W \in \mathbb{R}^{P \times C}$ over the D features in P using a single pass of SGD over random subset of data.
- Select s_t features $j \in P$ for which $\sum_{c=1}^{C} W_{j,c}^2$ are largest; discard the remaining $D - s_t$.

end for
return Final collection of D random features P

Cantonese and Bengali Results

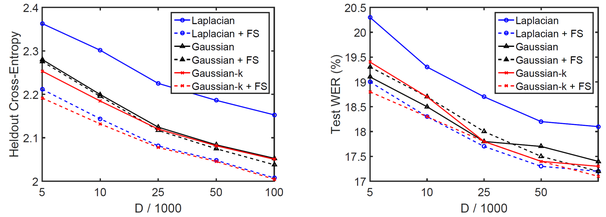


(a) Test results on Cantonese



(b) Test results on Bengali

Broadcast News Results



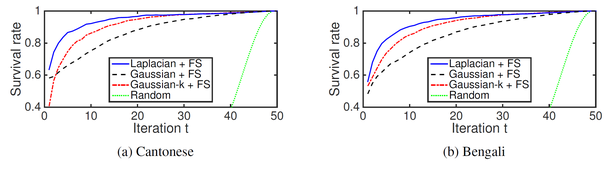
Comparison to DNNs

Method	Cantonese		Bengali	
	Cross-Entropy	CER	Cross-Entropy	WER
Best DNN	1.81	**67.3%**	1.28	**71.3%**
Best Kernel	1.81	68.6%	1.30	72.7%
Method	BN50 (CE LR Decay)		BN50 ("ERP" [2] LR Decay)	
	Cross-Entropy	WER	Cross-Entropy	WER
Best DNN	2.04	**16.5%**	2.35	**16.5%**
Best Kernel	2.00	17.1%	2.06	16.6%

[2] ICASSP 2016: USC, Columbia, IBM. "A Comparison Between Deep Neural Networks and Kernel Acoustic Models for Speech Recognition." **Come See Poster Tomorrow!!**

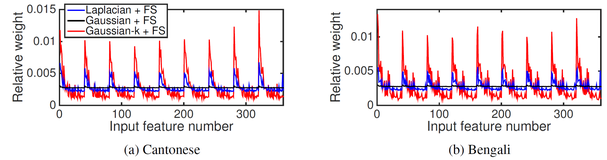
Effects of Feature Selection

Do features selected in iteration t survive future rounds?



(a) Cantonese (b) Bengali

Are some input features more important than others?



(a) Cantonese (b) Bengali

Future Work

- Why are DNNs typically better than kernels models in terms of WER, when they have comparable cross-entropy?
- Can we reduce the number of parameters needed by the kernel models, and maintain strong performance? Could we then train models with even more random features?
- How does feature selection compare with backpropagation?
- Can we leverage sparse random projections for faster training?
- Theoretical guarantees for algorithm?