Categorical Feature Compression via Submodular Optimization

Mohammad Hossein Bateni, Lin Chen, Hossein Esfandiari, Thomas Fu, Vahab Mirrokni, and Afshin Rostamizadeh

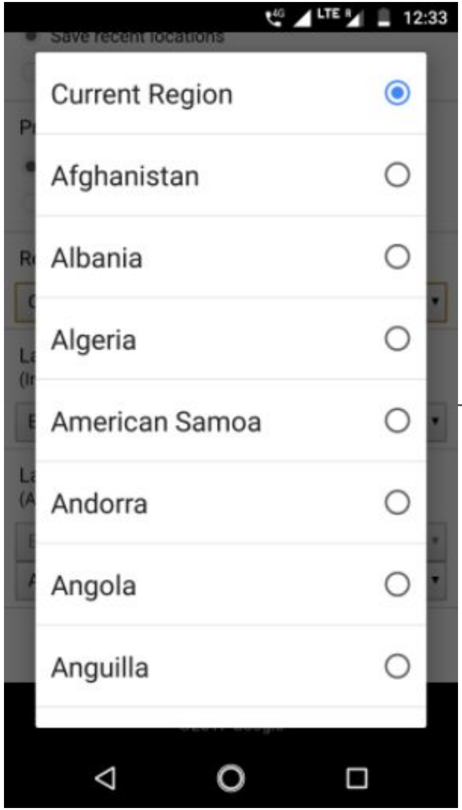
Pacific Ballroom #142

Google

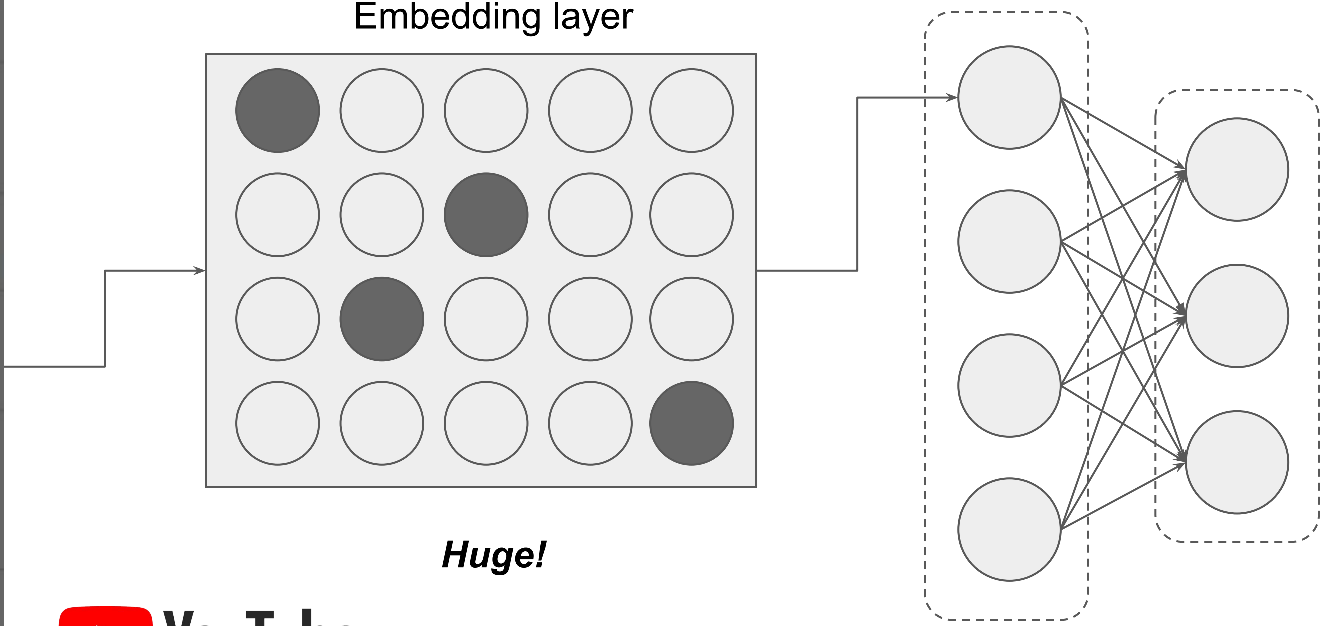
Yale

Why Vocabulary Compression?

Why Vocabulary Compression?



Embedding layer

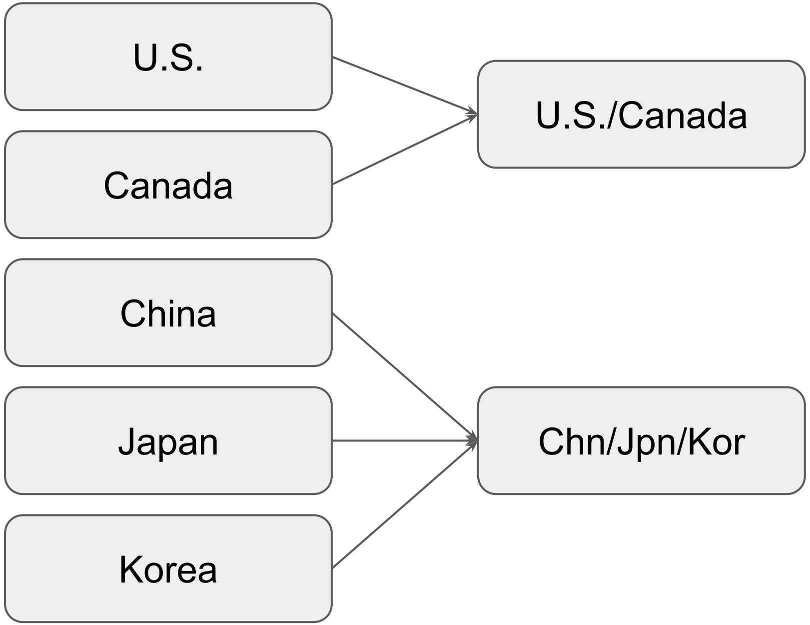


Huge!

Video ID: ~7 billion values
99.9% of neural net

How to Compress Vocabulary?

How to Compress Vocabulary



Group similar feature values into one.

Good compression preserves ***most information of labels***.

Supervised

Problem Formulation

Problem Formulation

User ID	Featur e	Compressed feature	Favorite fruit (label)
#1843	China	China/Japan/Korea	
#429	Japan	China/Japan/Korea	
...			
#9077	Brazil	Brazil/Argentina	

Random variable X ∈ {Afghanistan, Albania, …, Zimbabwe}

Compressed feature f(X) ∈ {China/Japan/Korea, Brazil/Argentina, U.S./Canada}

Random variable C ∈ {pear, apple, …, mango}

Max I(f(X); C)

s.t. f(X) can take at most *m* values

Our Results

Our Results

Max I(f(X); C)

s.t. f(X) can take at most *m* values

There is a ***quasi-linear*** (O(n log n)) algorithm that achieves ***63%*** f(OPT) if label is ***binary***.

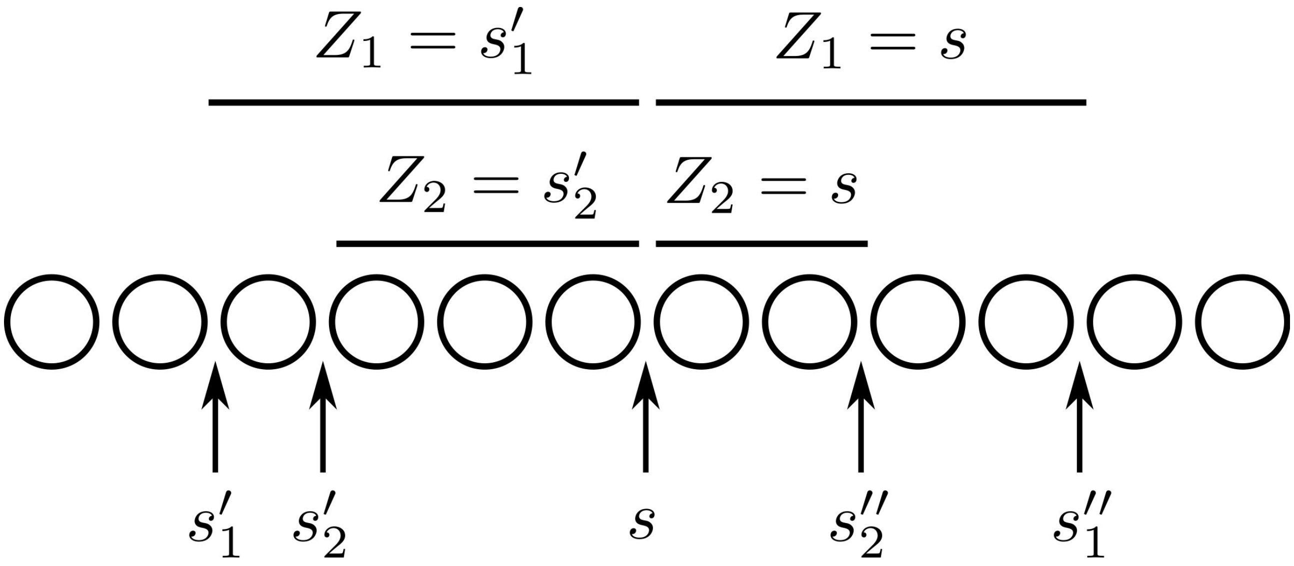
- Design a new submodular function after re-parametrization

There is a ***log(n)***-round distributed algorithm that achieves ***63%*** f(OPT) with ***O(n/k)*** space per machine.

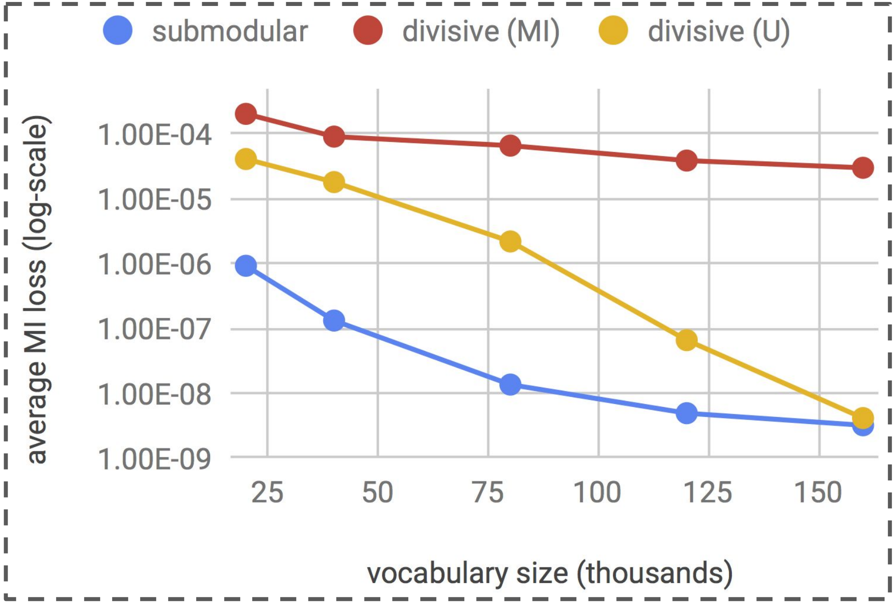
- *k* is # of machines

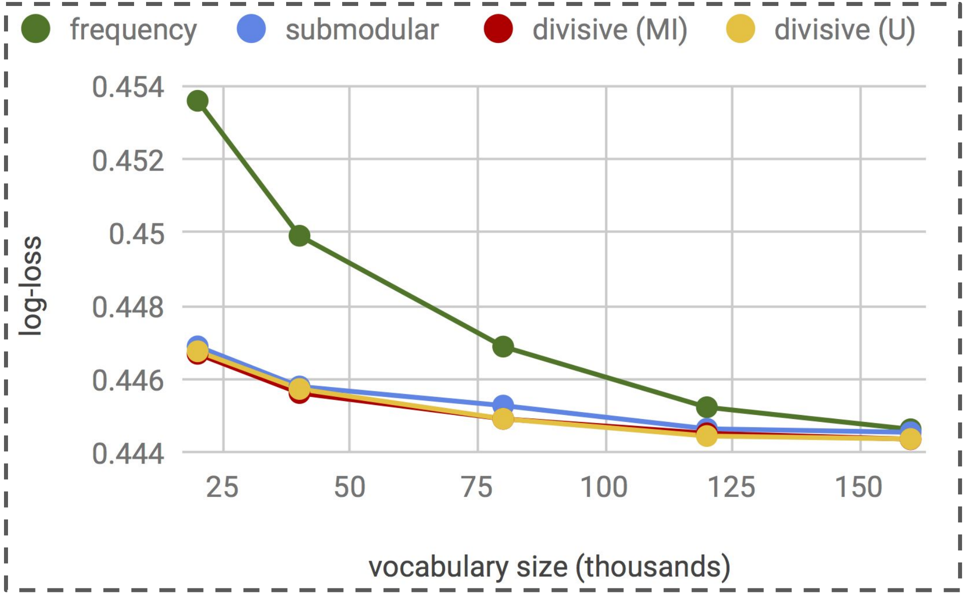
Reparametrization for Submodularity

- Sort feature values ***x*** according to ***P(X=x|C=0)***.
- A problem of placing separators
- ***I(f(X); C)*** is a function of the set of separators.



Experiment Results





Pacific Ballroom #142

See you this evening