Text-based Geolocation Prediction of social media users

Ismini Lourentzou, Alex Morales and ChengXiang Zhai

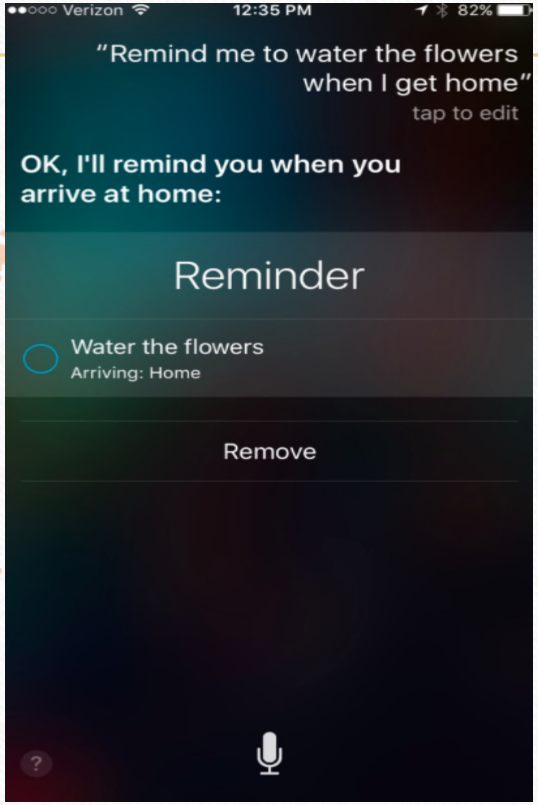
CS @ University of Illinois at Urbana – Champaign

IEEE Big Data 2017

http://sifaka.cs.uiuc.edu/ir/

Social Media era

https://ghanatalksbusiness.com/wp-content/uploads/2017/05/social-media-marketing.png



http://3.bp.blogspot.com/-QZNK17a[illegible]Nmig-OzBpAXsP[illegible]1600/Screen%2BShot%2B2016-06-17%2Bat%2B12.37.42%2BPM.png

http://pm1.narvii.com/6442/50898c483c04e58712820feb61632bcacab4c635_hq.jpg

https://metrouk2.files.wordpress.com/2015/11/safetycheck1.jpg

http://smallbusinessu.org/wp-content/uploads/2016/09/check-in.jpg

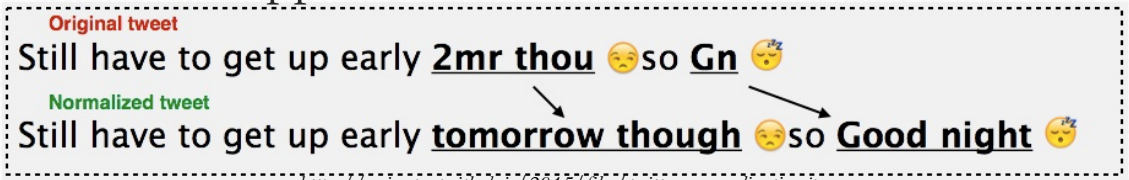
https://pbs.twimg.com/media/CUHq2d-UAAAibSQ.jpg

- 😕 Missing
- 😕 Highly unstructured
- 😕 Non-geographical
- 😕 A tiny proportion of users geotagging their posts

- ✓ Unreliable
- ✓ Imprecise
- ✓ Need for predictive models

Geolocation prediction: which data?

- **User generated-content, such as posts**
  - Simple. easily adjusted to new datasets for real-time applications
  - Text normalization



https://noisy-text.github.io/2015/files/twitter_normalization.jpg

- **Metadata**
  - Time zone, number of followers, likes etc.
  - Availability depends highly on the provider and can vary among social media platforms

- **Network information**
  - User-friend, user-mentions, etc.
  - Time consuming for large social networks

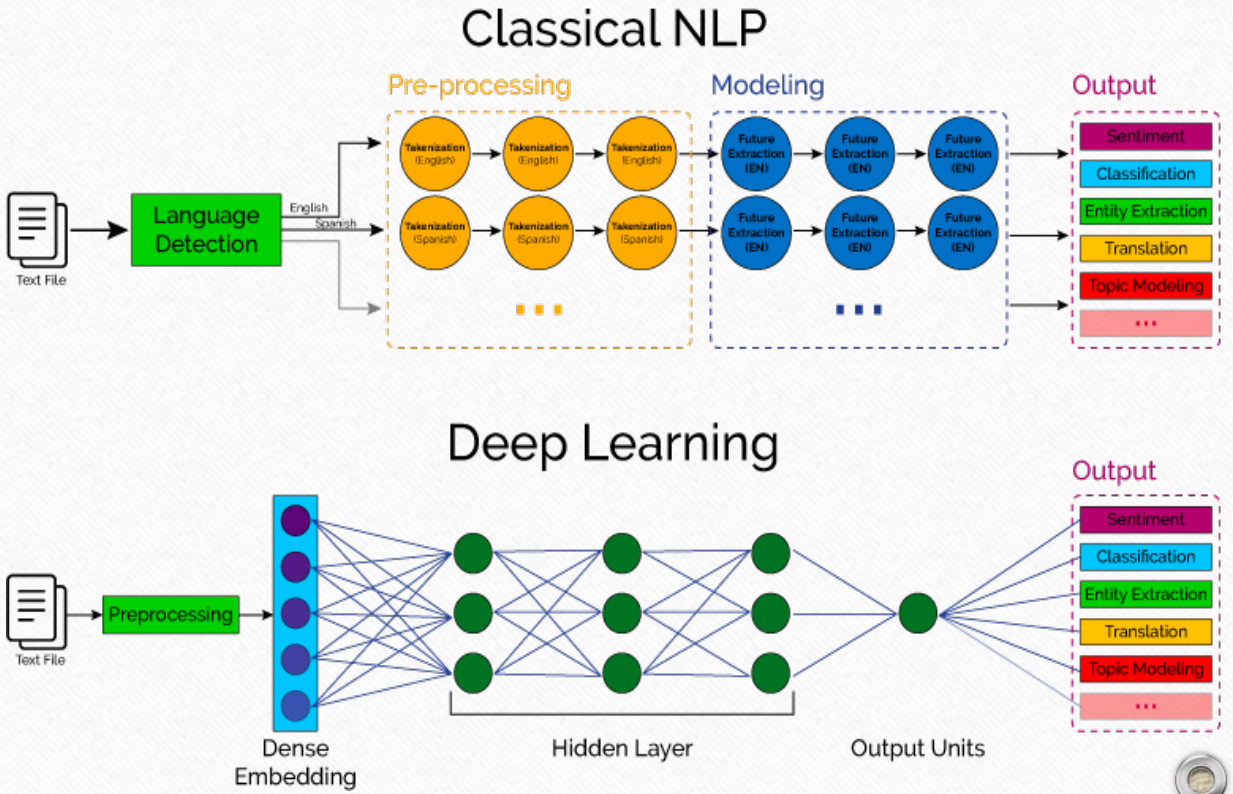
http://online.wsj.com/media/crt_facebook_F_20101220120320.jpg

And why Deep Learning?

Geolocation with traditional ML requires:

- Effective feature construction
  - Simply combining surface features would not do
  - Need for increased engineering efforts
- Complex multi-resolution algorithms
  - Previous work relied on grid-based methods & ensemble models
  - Combine different levels of geographical resolution

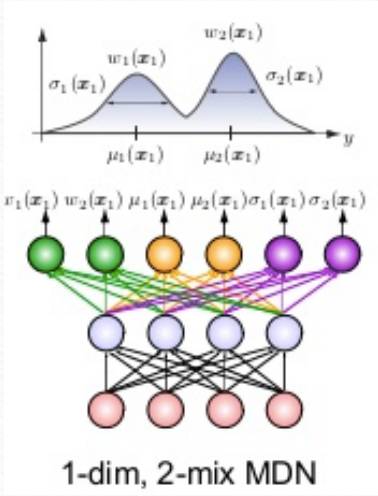
Simple input + Neural Network ?



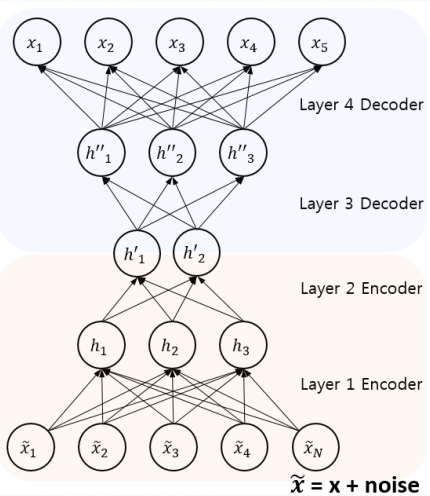
https://pbs.twimg.com/media/DJKdejZW0AELJsR.png

Neural Geolocation (related work)

- Rahimi et al., 2017
  - (Gaussian) Mixture Density Networks
  - Output: probability distribution over all location points
  - Uncertainty estimates over all the coordinate space
- Liu and Inkpen, 2015
  - 3-hidden layer Stack Denoising Autoencoder



https://www.slideshare.net/danilosoba1/statistical-parametric-speech-synthesis-heiga-zen



https://wikidocs.net/images/page/3413/sDA.png

❑ Liu and Inkpen, "Estimating user location in social media with stacked denoising auto-encoders." 1st Workshop on Vector Space Modeling for NLP, NAACL, 2015
❑ Rahimi et al. "Continuous representation of location for geolocation and lexical dialectology using mixture density networks," EMNLP, 2017

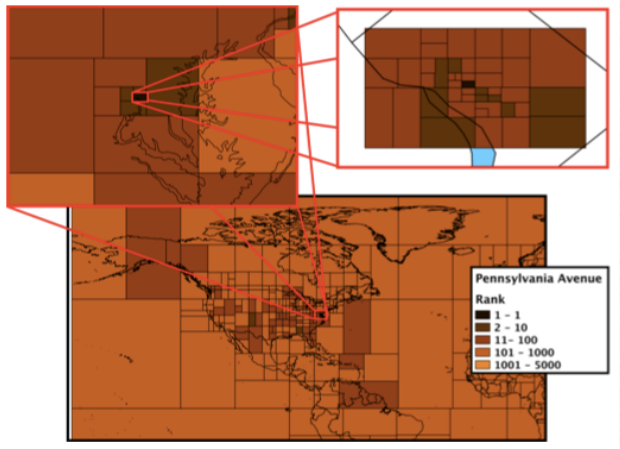
Related work before DL (1)

- Eisenstein et al., 2010
  - Geo-topic model
  - Limited to small datasets due to computational complexity
- Cha et al., 2015
  - Sparse coding + Voting-based grid nearest neighbors model
  - Incorporating word order information, i.e. word sequences

❑ Eisenstein et al. "A latent variable model for geographic lexical variation." EMNLP 2010
❑ Cha et al. "Twitter geolocation and regional classification via sparse coding," ICWSM 2015

Related work before DL (2)

- Wing and Baldridge, 2011
  - Divide Earth into uniform grids and construct pseudo-document for each grid
  - Uniform grids do not take into account the skewness of the pseudo-document distribution
- Roller et al., 2012
  - Constructing grids using adaptive k-d trees
- Wing and Baldridge, 2014
  - Logistic regression on hierarchy of grids



Picture from Wing and Baldridge, 2014

- Wing and Baldridge, "Simple supervised document geolocation with geodesic grids," ACL-HLT 2011
- Wing and J. Baldridge, "Hierarchical discriminative classification for text-based geolocation." EMNLP 2014
- Roller et al. "Supervised text-based geolocation using language models on an adaptive grid," EMNLP 2012

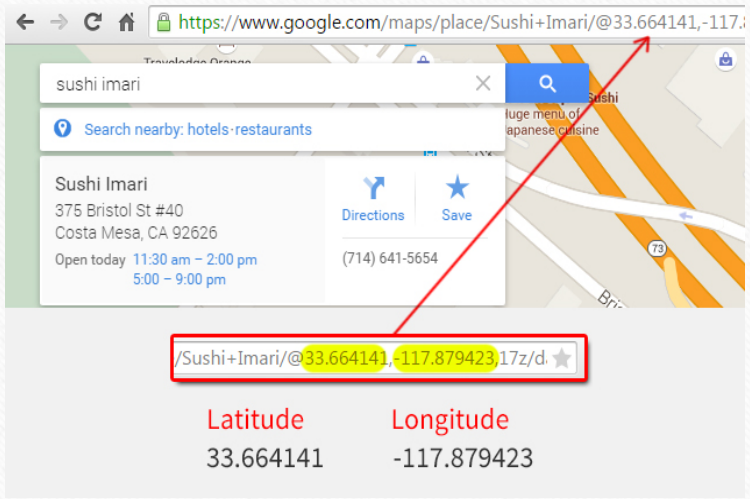
Datasets

- Each training example is the collection of all tweets by a single user **(word counts)**
- The earliest geotagged tweet determines the user's location
- Already split in training, development and test sets
- Removed tweets non-English and not near a city
- Filtered non-alphabetic, overly short and overly infrequent words

Dataset Name	Users	Sample Size	Region
GeoText	9.5K	380K tweets	Contiguous US
TwUS	450K	38M tweets	North America
TwWORLD	1.4M	12M tweets	English World Wide

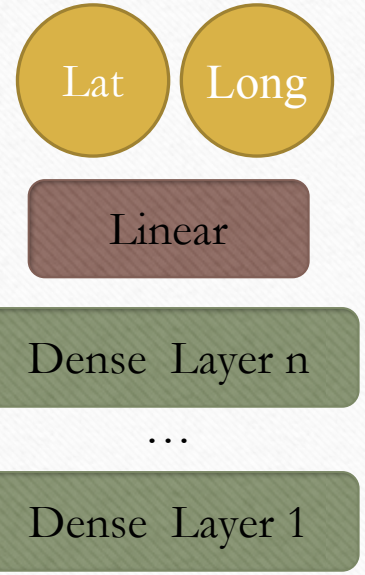
- Eisenstein et al. "A latent variable model for geographic lexical variation." EMNLP 2010
- Roller et al. "Supervised text-based geolocation using language models on an adaptive grid." EMNLP 2012
- Han et. al. "Text-based twitter user geolocation prediction." Journal of Artificial Intelligence Research 2014

Regression



https://whitespark.ca/uploads/grab-geo-coordinates.jpg

GeoText **TWUS** **TWWORLD**



(# users, 2)
Coordinates
latitude, longtitude

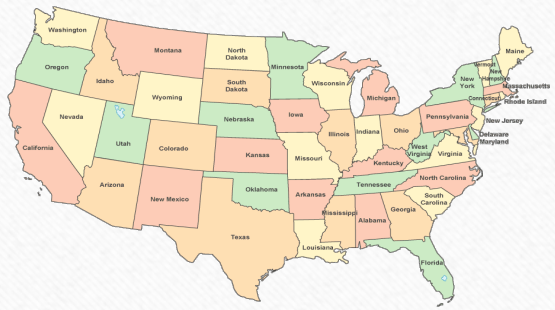
TF-IDF tweet representation $w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$

#Seahawks #Colts Sunday Night Football showdown

Loss: Haversine

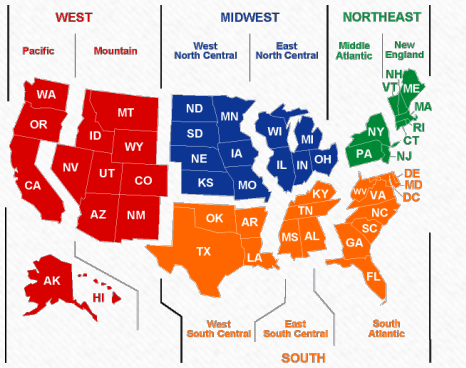
Mean and Median error distance (km) + Acc@161km radius

Classification



http://www.united-states-map.com/usa-conic-1256-916.gif

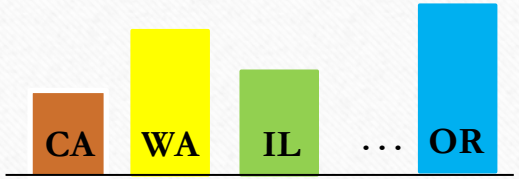
49 classes



https://www.pinterest.com/pin/376965431287282767/

4 classes

GeoText



(# users, # classes)
49 states/4 regions

Softmax $softmax(x_j) = \frac{e^{x_j}}{\sum_k e^{x_k}}$

Dense Layer n

…

Dense Layer 1

TF-IDF tweet representation $w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$

#Seahawks #Colts Sunday Night Football showdown

Loss: Cross-entropy

Accuracy: proportion of correctly classified users

Design choices

Regularization	Activation functions	Architecture size
Batch Normalization	Non-linear (simgoid)	Number of layers
Dropout	Linear non-parametric (ReLU)	Neurons per layer
	Parameterized linearity (PReLU)	

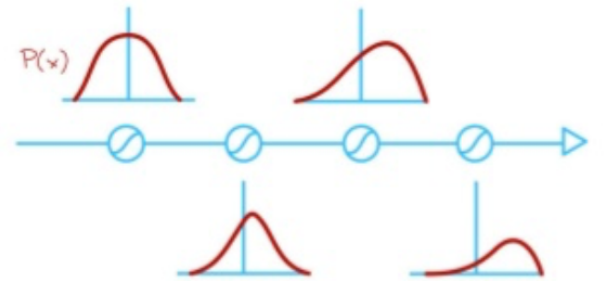
Regularization

Batch Normalization

Internal Covariate shift: as learning progresses, the distribution of layer input changes due to parameter updates, which slows down learning

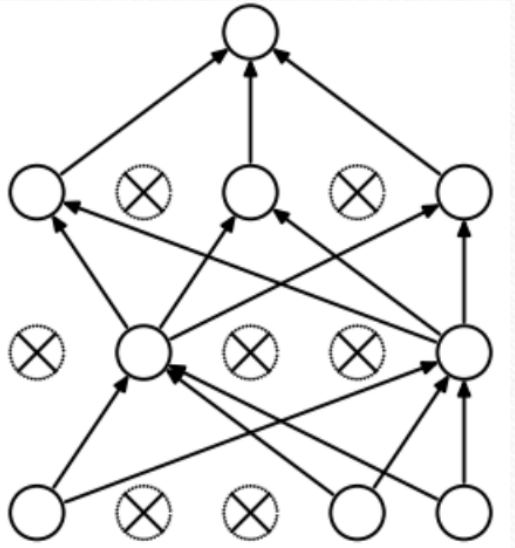
Batch Normalization: whitened inputs (i.e. zero mean, unit variances)

- Networks converge faster
- Allows much higher learning rates
- Reduces the sensitivity to the weight initialization
- Makes more activation functions viable
- Provides regularization



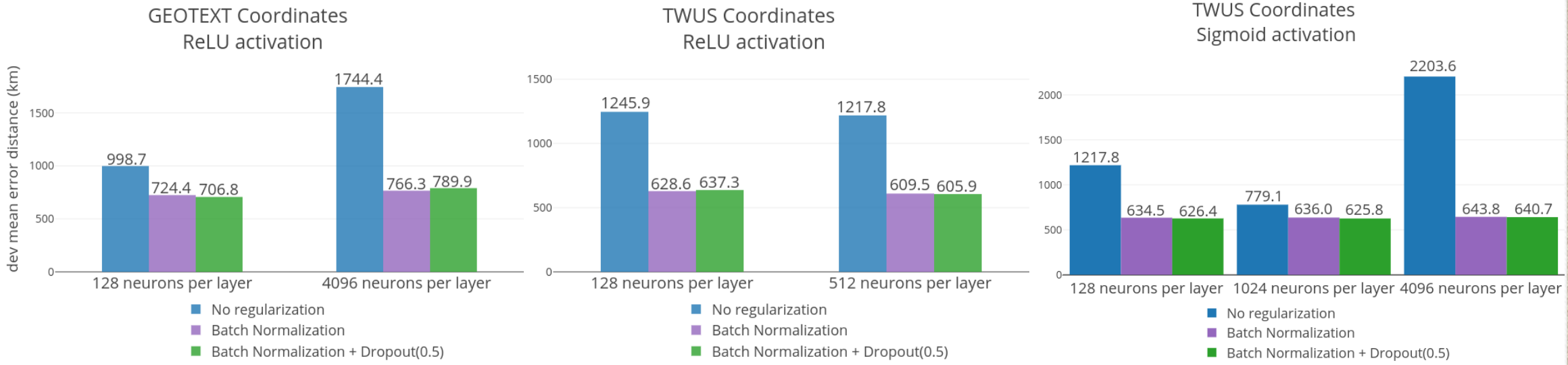
https://image.slidesharecdn.com/dlcv2017d2l1optimization-170622143746/95/optimization-for-deep-networks-d2l1-2017-upc-deep-learning-for-computer-vision-8-638.jpg?cb=1498142501

Dropout



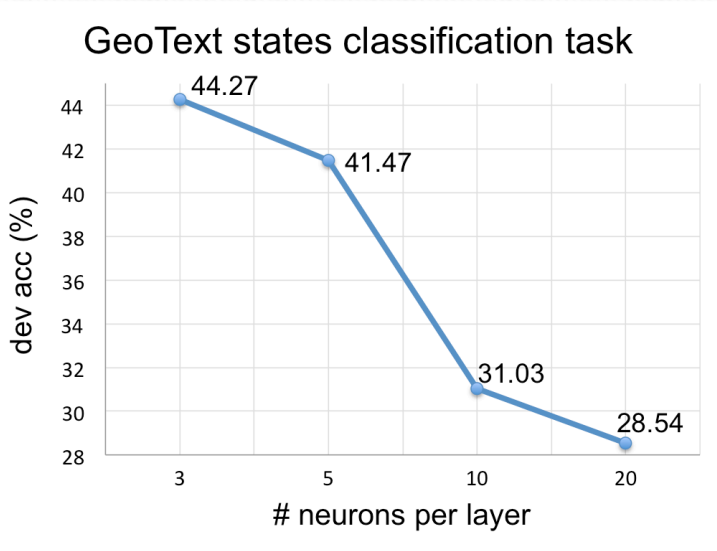
*https://cdn-images-1.medium.com/max/1044/1*iWQzxhVlwadk6VAJjsgXgg.png*

Regularization: results

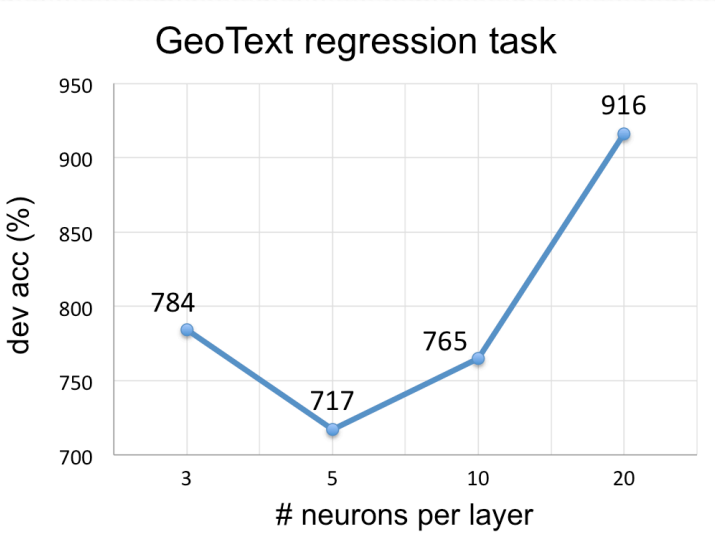


- Comparable results for classification tasks
- Dropout's effect on the performance is marginal
- Batch Normalization: robust performance improvements

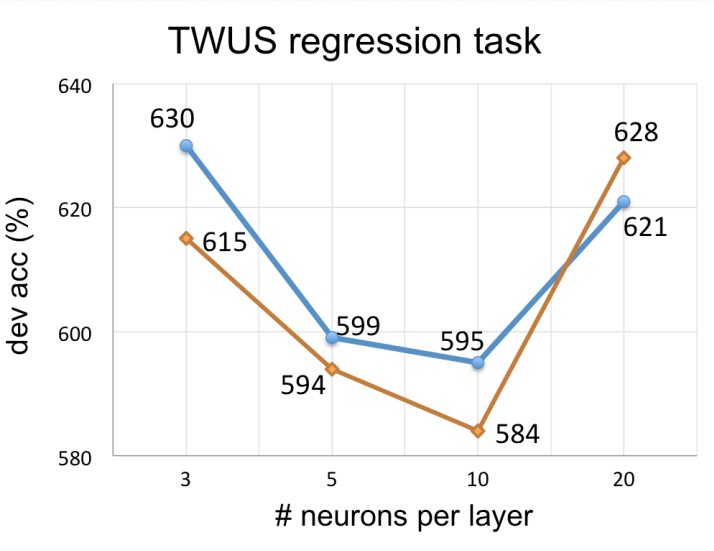
Varying number of layers



Less is better



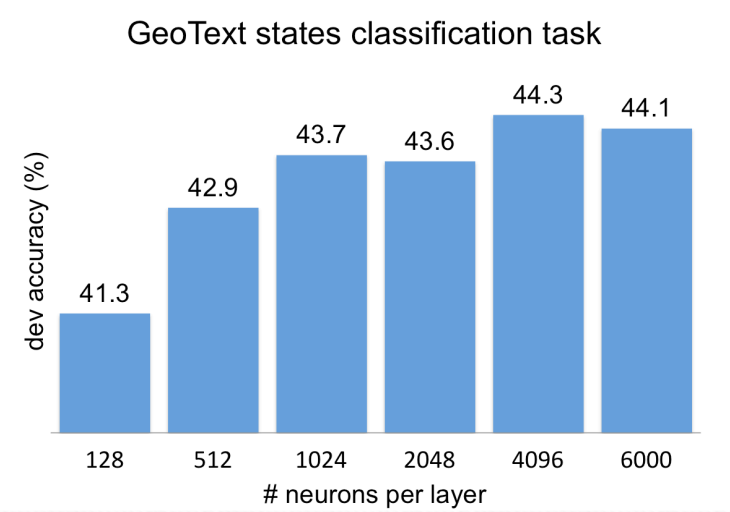
? is better



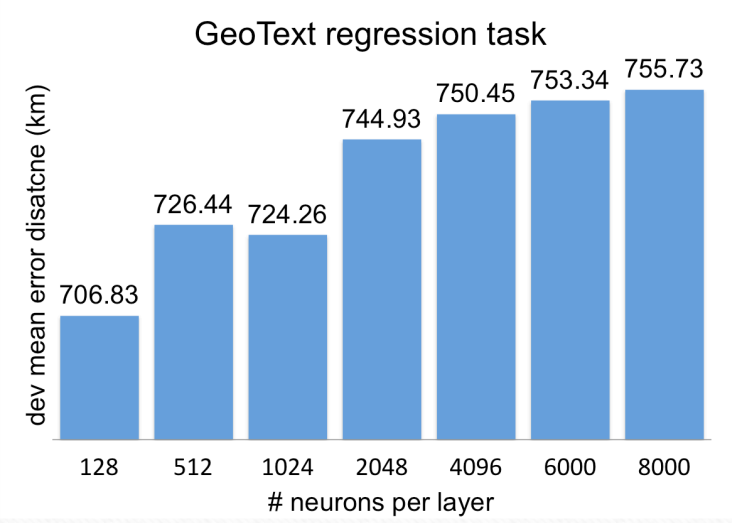
? is better

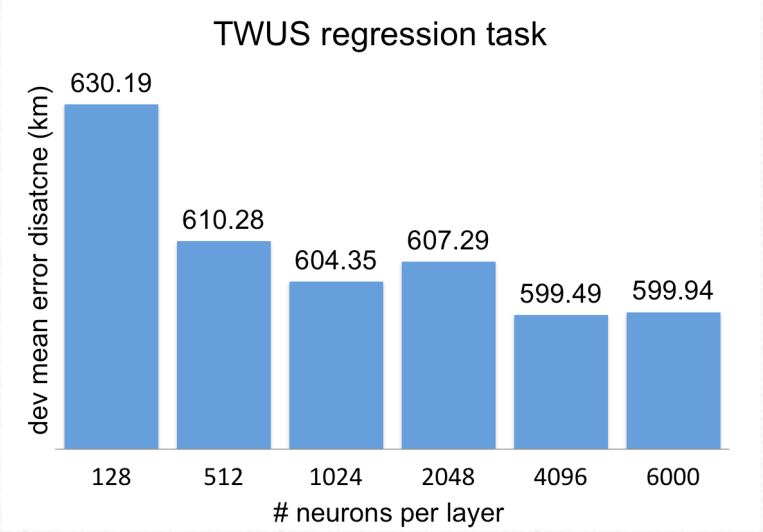
"no solution fits all"

Varying number of neurons per layer



More is better





More is better

Shallow and wide architectures

Best performing hyper-parameter settings

	dropout	hidden	activation	layers
GEOTEXT states	0.5	4096	PReLU	3
GEOTEXT regions	0.5	512	ReLU	3
GEOTEXT regression	0.5	128	ReLU	3
TWUS	0	4096	PReLU	5
TWWORLD	0	4096	Sigmoid	3

Evaluation: GeoText

GeoText (Accuracy %)	States	Regions
Proposed Method	44.3	67.3
Liu and Inkpen, 2015 (SDA)	34.8	61.1
Eisenstein et al., 2010 (Ge		
Cha et al., 2015 (SC+wor		

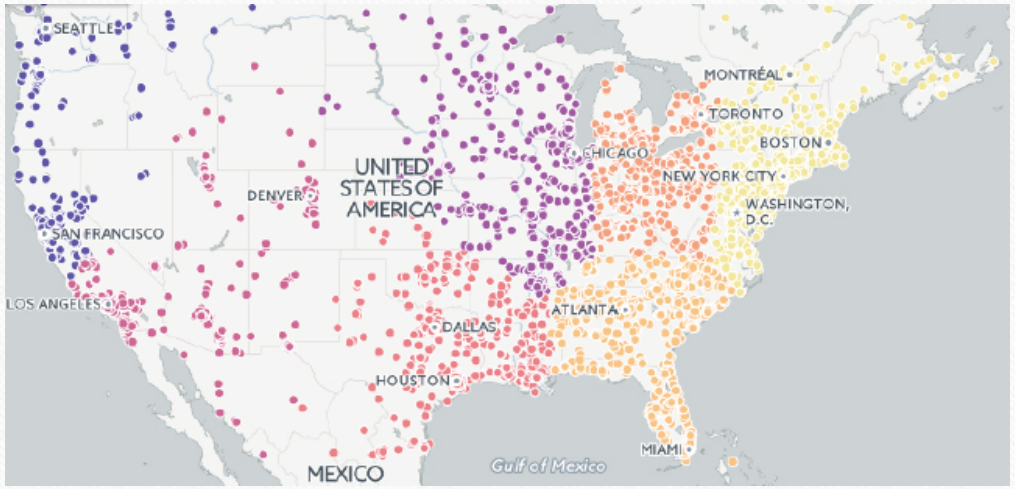
GeoText (Errors in km)	Mean	Median	Acc@161
Proposed Method	747	448	29
Rahimi et al., 2017 (MDN-Shared)	865	412	39
Liu and Inkpen, 2015 (SDA)	856	-	-
Cha et al., 2015 (SC+word counts)	926	497	-
Cha et al., 2015 (SC+**word sequences**)	581	425	-
Roller et al., 2012 (UnifKd)	890	473	34
Eisenstein et al., 2010 (Geo topic model)	900	494	-

Evaluation: TWUS & TWWORLD

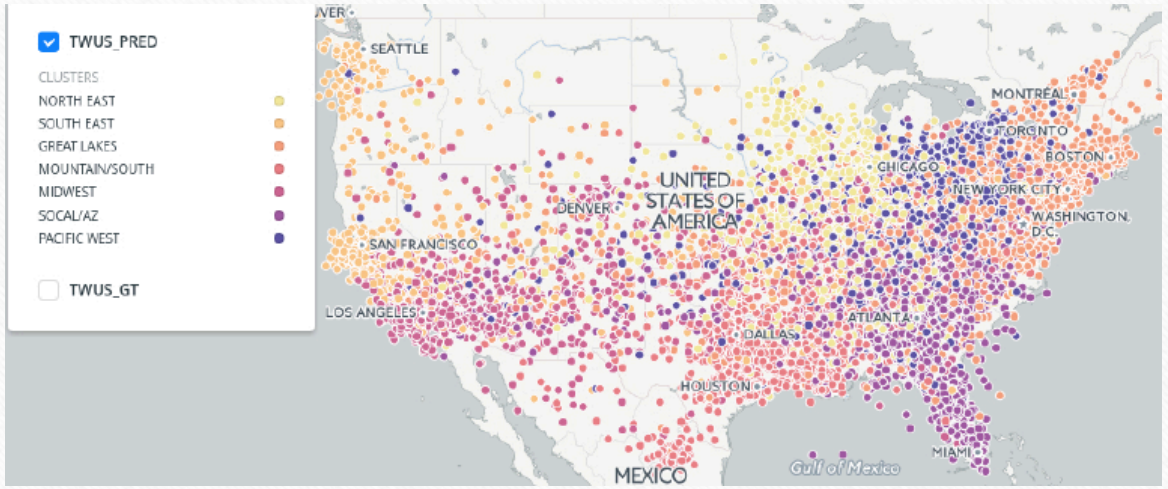
TWUS (Errors in km)	Mean	Median	Acc@161
Proposed Method	570	223	43
Rahimi et al., 2017 (MDN-Shared)	655	216	42
Liu and Inkpen, 2015 (SDA)	733	377	24
Wing and Baldridge, 2014 (HierLR Uniform)	704	171	49
Wing and Baldridge			
Roller et al., 20			

TWWORLD (Errors in km)	Mean	Median	Acc@161
Proposed Method	1338	495	21
Rahimi et al., 2017 (MDN-Shared)	-	-	-
Liu and Inkpen, 2015 (SDA)	-	-	-
Wing and Baldridge, 2014 (HierLR Uniform)	1715	490	33
Wing and Baldridge, 2014 (HierLR kd)	1670	509	31

Error analysis - Map available online



Ground truth clusters



Our model predictions

http://bit.do/geoDL

Conclusions & Future Work

- ✓ Explore several design choices and compare on three different task settings
- ✓ Show how hyper-parameter changes impacts our models and to what effect
  - Particularly useful in the case of transfer learning
  - When the classification task is refined, what options are available for keeping the performance at the same level
- ✓ Batch Normalization leads to better regularization and has the highest performance increase

---

- Investigate effect of information beyond text, such as metadata, user or network information
- Jointly learn "user embeddings" alongside with word embeddings
- Word order information (CNNs, LSTMs etc.)

https://giphy.com/gifs/thanks-thank-you-thnx-3o6ozuHcxTtVWJJn32/download