ACL2018読み会@名大

Semi-supervised User Geolocation via Graph Convolutional Networks

Afshin Rahimi **Trevor Cohn** **Timothy Baldwin**
School of Computing and Information Systems
The University of Melbourne
arahimi@student.unimelb.edu.au
{t.cohn,tbaldwin}@unimelb.edu.au

発表者：廣中詩織（豊橋技術科学大学）@elnikkis

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2009-2019, 2018.

Abstract

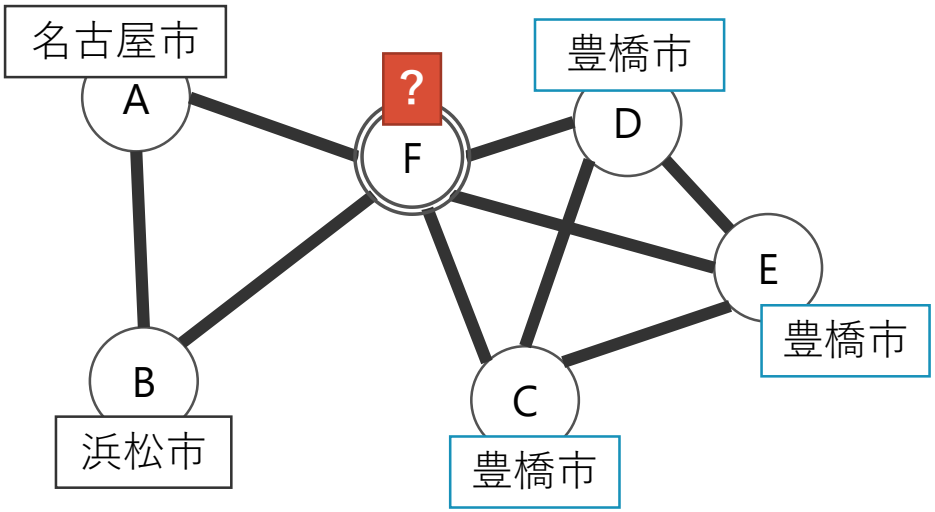
- ソーシャルメディアユーザの場所を推定するタスク
- Graph Convolutional Networksをもとにした
  マルチビュー位置推定モデルGCNを提案
  - テキスト情報とネットワーク情報との両方を使う
    semi-supervisedな手法
- 3つのデータセットを使ってstate-of-the-art（と提案した2つのベースライン）と比較
  - 十分なラベルデータがあるときSOTAに勝つまたは同等の性能
  - ラベルデータが少ないときベースラインより良い性能
- highway network gatesはGCNの有用な近傍の量を制御するために不可欠であることを発見

ユーザの場所推定

ソーシャルグラフを使う ネットワークベースの手法

ノード：ユーザ
(例：A, B, C, …)

エッジ：ユーザ間の関係
(例：@メンション)



グラフ上で近いユーザと
地理的距離が近いと仮定

ユーザの投稿を使う テキストベースの手法

ユーザA：
- **うなぎ**食べた
- 今日は晴れてる
- **静岡**に帰省してる

ユーザB：
- **山賊焼き**おいしい（写真）
- 今日は**カープ**の応援にきた

地名や地域の特徴的な単語を利用

両方使うハイブリッドの手法もある

Model

一部ユーザのラベルY_Sが与えられるときY_Uを予測する

- $X \in \mathbb{R}^{|U|\times|V|}$: text view, 各ユーザのbag of words
- $A \in \mathbb{1}^{|U|\times|U|}$: network view, ユーザのインタラクション

$|U|$: ユーザ数
$|V|$: 単語数

- y_iはラベルのone-hotベクトル（場所がラベル）
- ラベル付きユーザU_SとラベルなしユーザU_Hがいる
  （$U = U_S \cup U_H$）

Model - GCN

ニューラルネットワークモデル$f(X, A)$の各レイヤ

$$\hat{A} = \tilde{D}^{-\frac{1}{2}}(A + \lambda I)\tilde{D}^{-\frac{1}{2}}$$

Aが$\{0, 1\}$なのでなめらかにしているかんじ（前処理）

$$\mathrm{H}^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{(l)} + b\right)$$

X : text view

A : network view

- $H^0 = X$
- $\tilde{D}$はdegree matrixで、λは重み
- $W^{(l)}$とbは学習するパラメータ
  - $W^{(l)}$は$d_{\mathrm{in}} \times d_{\mathrm{out}}$の行列
  - bは$d_{\mathrm{out}} \times 1$の行列
- σはnonlinear activation function

Model - GCN

つまり、ユーザu_iのレイヤlの出力は次のように計算

$$\vec{h}_i^{l+1} = \sigma\left(\sum_{j \in \mathrm{nhood}(i)} \hat{A}_{ij}\vec{h}_j^{l} W^{l} + b^{l}\right)$$

- W^lとb^lは学習するパラメータ
- $\mathrm{nhood}(i)$はu_iの隣接ノード集合

- レイヤ数は何ホップ先までの情報を使うか
  - レイヤ数3にするのは3ホップ先までの情報を使うということになる

Model - Highway GCN

GCNのレイヤを増やしていくと
伝搬されてくる情報にノイズが増える

なので、ノードごとにどのくらいの距離までの情報を渡すか制御する

gating weights $T\left(\vec{h}^{l}\right)$を使う

$$T\left(\vec{h}^{l}\right) = \sigma\left(W_t^{l}\vec{h}^{l} + b_t^{l}\right)$$

$$\vec{h}^{l+1} = \vec{h}^{l+1} \circ T\left(\vec{h}^{l}\right) + \vec{h}^{l} \circ \left(1 - T\left(\vec{h}^{l}\right)\right)$$

次のレイヤの値か
元のレイヤの値かを選択

$\circ$: 要素ごとの積
σ : シグモイド関数

Model - (Highway) GCN

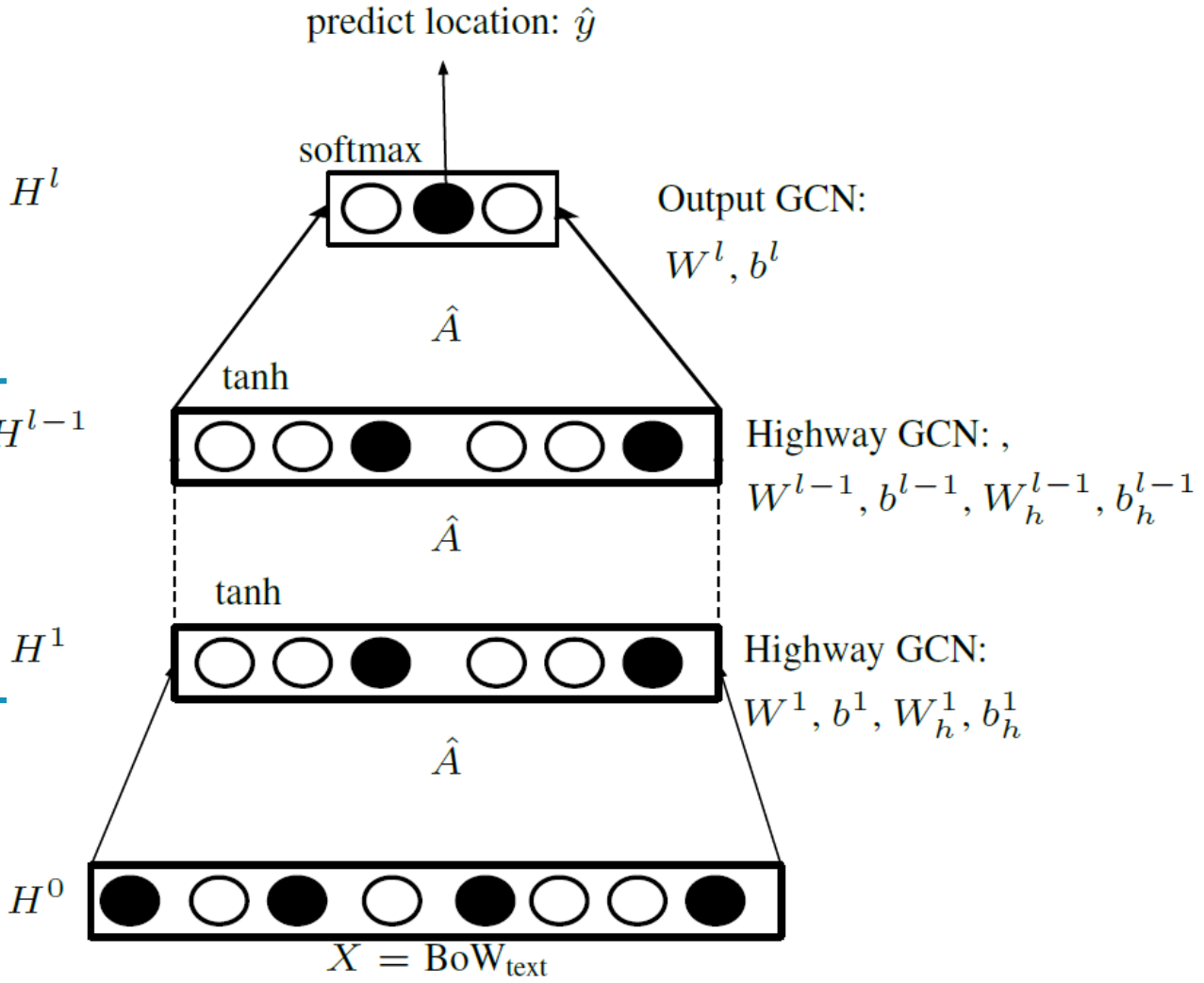
3.
softmaxをとって
ラベルを出力

2.
語彙を圧縮しつつ伝搬

隠れ層のレイヤの
サイズは変わらない

実験では
レイヤ数=3くらい

1.
各ユーザの語彙を入力



引用：Figure 1

Model - DCCA

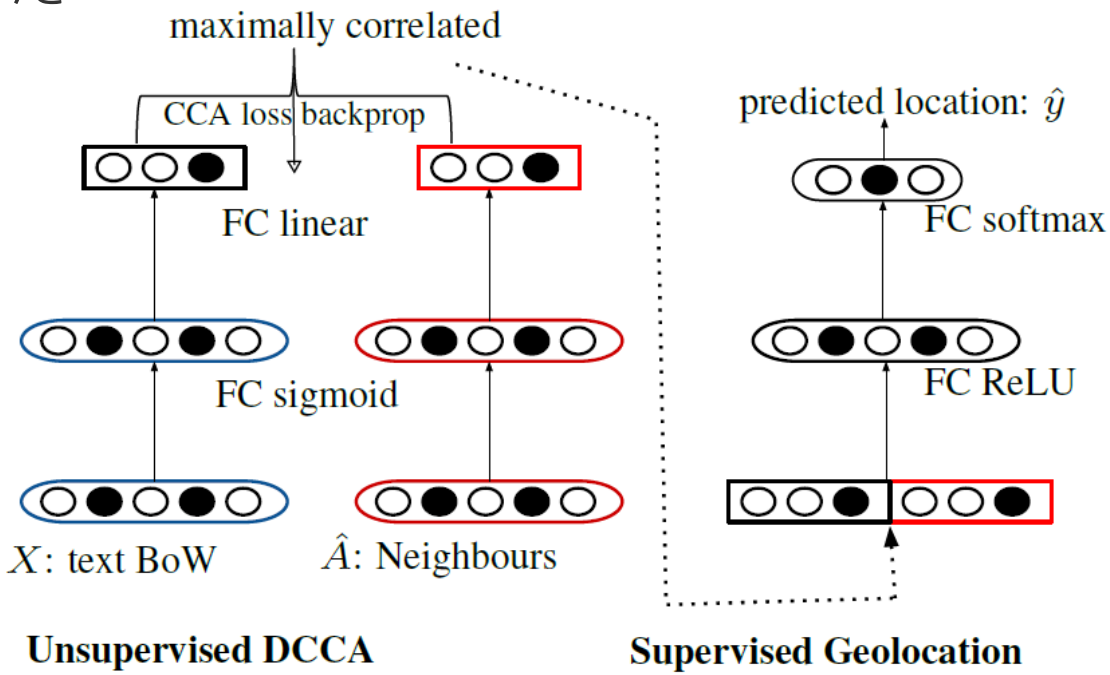
DCCA: Deep Canonical Correlation Analysis

1. $f_1(X)$と$f_2(\hat{A})$の相関を最大化するf_1とf_2を
   多層パーセプトロンで学習　　$\rho = \mathrm{corr}\left(f_1(X), f_2(\hat{A})\right)$

2. $f_1(X)$と$f_2(\hat{A})$を結合した
   ものを使って推定

Deepといっても
隠れ層は1つ

$f_1(X)$と$f_2(\hat{A})$はそれぞれ、
相関しないノイズを減少させた
Xと$\hat{A}$の圧縮された表現になる



引用：Figure 2

Baselines

- GCN-LP
  - GCNのH^0へテキスト情報の代わりにネットワーク情報を入れたもの
  - ネットワーク情報のみを使うGCN
- MLP-TXT+NET
  - text view Xとnetwork view $\hat{A}$とを結合したものを多層パーセプトロン（隠れ層1）に入力したもの
  - テキスト情報とネットワーク情報を単純に使ったニューラルネット

Data

データセット名	ユーザ数	ラベルの種類数	ユーザの場所
GEOTEXT [Eisenstein 10]	9k	129	最初の位置情報付きツイートの座標
TWITTER-US [Roller 12]	449k	256	最初の位置情報付きツイートの座標
TWITTER-WORLD [Han 12]	1.3m	930	最も近いcityの中心座標

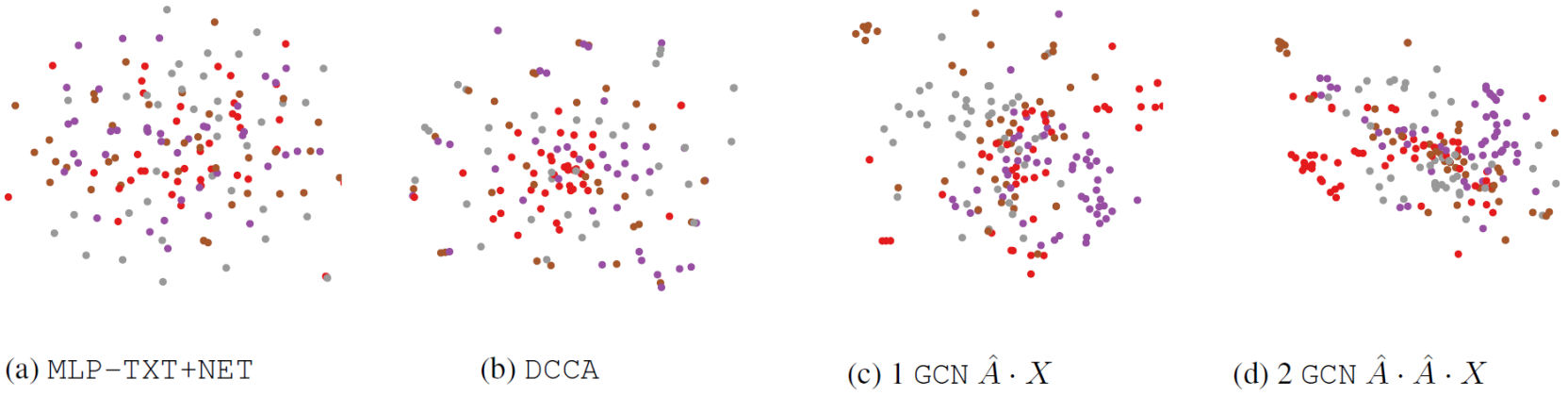
- ラベル：座標をk-d treeを使って離散化
- network view: メンションをもとに作る
- text view: ユーザの投稿を使ったtf-idf
- train, dev, testは最初から分割されている

Representation

t-SNEで可視化

DCCAはただ2つのviewを
合体させたものより良い

GCNはもっとイケてる



(a) MLP−TXT+NET (b) DCCA (c) 1 GCN $\hat{A} \cdot X$ (d) 2 GCN $\hat{A} \cdot \hat{A} \cdot X$

Figure 3: Comparing t-SNE visualisations of 50 training samples from each of 4 randomly chosen regions of GEOTEXT using various data representations: (a) concatenation of $\hat{A}$ (Equation 1); (b) concatenation of DCCA transformation of text-based and network-based views X and $\hat{A}$; (c) applying graph convolution $\hat{A} \cdot X$; and (d) applying graph convolution twice $\hat{A} \cdot \hat{A} \cdot X$

引用：Figure 3

Labeled Data Size

ラベル付きデータが10%以下
だとGCNとDCCAが良い

ラベル付きデータが95〜98%以上
あるとMLP-TXT+NETの性能が最高

良い

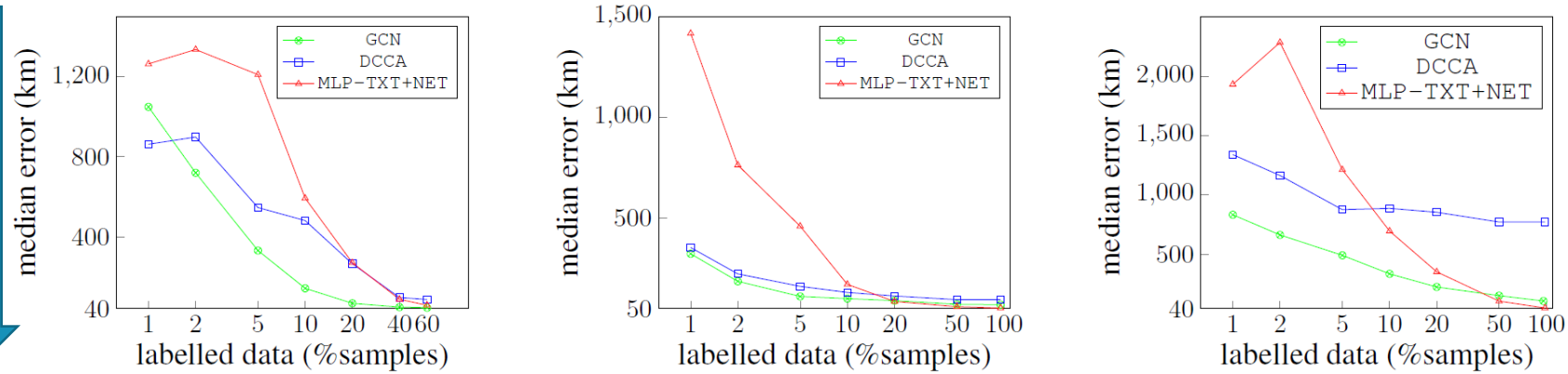


Figure 4: The effect of the amount of labelled data available as a fraction of all samples for GeoText, Twitter-US, and Twitter-World on the development performance of GCN, DCCA, and MLP-TXT+NET models in terms of Median error. The dataset sizes are 9k, 440k, and 1.4m for the three datasets, respectively.

引用：Figure 4

Highway Gates

highway gateはGCNの
性能を出すために大事

最短経路長の分布と適合して、
レイヤ数4のところに
性能のピークがある

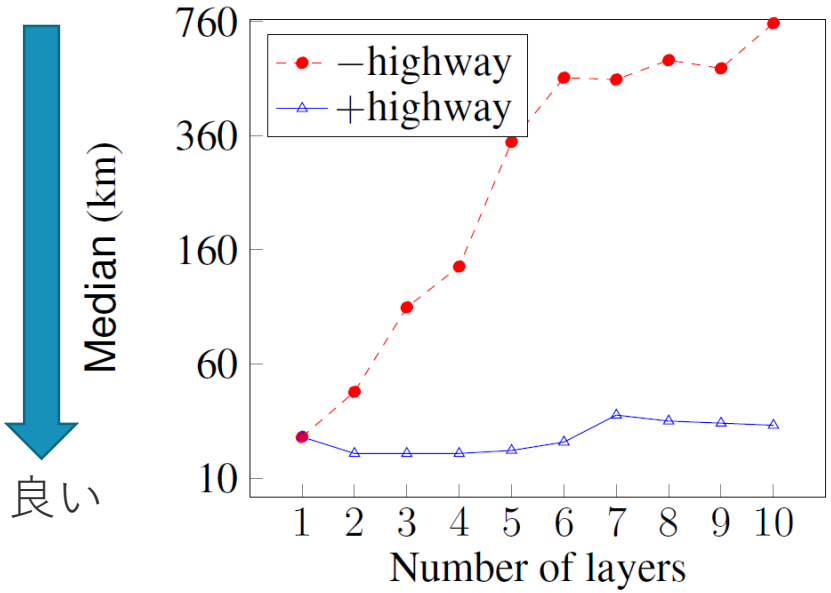


Figure 5: The effect of adding more GCN layers (neighbourhood expansion) to GCN in terms of median error over the development set of GEOTEXT with and without the highway gates, and averaged over 5 runs.

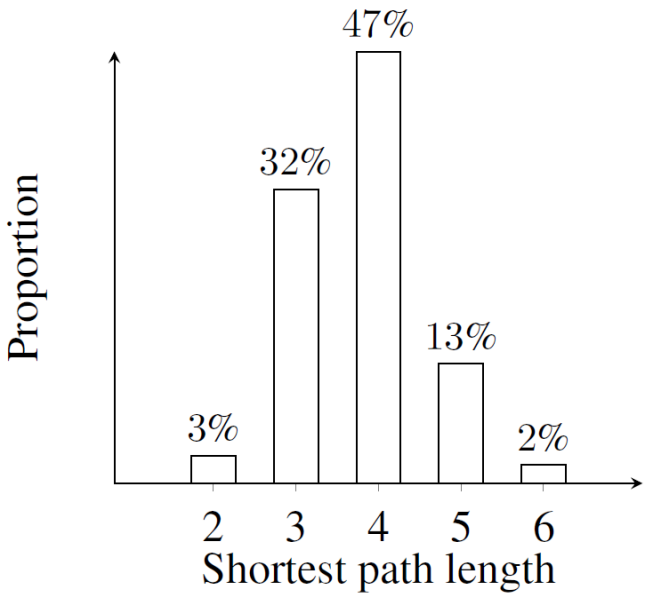


Figure 6: The distribution of shortest path lengths between all the nodes of the largest connected component of GEOTEXT's graph that constitute more than 1% of total.

引用：Figure 5, 6

Performance

	GeoText			Twitter-US			Twitter-World		
	Acc@161↑	Mean↓	Median↓	Acc@161↑	Mean↓	Median↓	Acc@161↑	Mean↓	Median↓
Text (inductive)									
Rahimi et al. (2017b)	38	844	389	54	554	120	34	1456	415
Wing and Baldridge (2014)	—	—	—	48	686	191	31	1669	509
Cha et al. (2015)	—	581	425	—	—	—	—	—	—
Network (transductive)									
Rahimi et al. (2015a)	58	586	60	54	705	116	45	2525	279
GCN−LP	58	576	56	53	653	126	45	2357	279
Text+Network (transductive)									
Do et al. (2017)	**62**	**532**	**32**	**66**	433	**45**	53	1044	118
Miura et al. (2017)	—	—	—	61	481	65	—	—	—
Rahimi et al. (2017b)	59	578	61	61	515	77	53	1280	104
MLP−TXT+NET	58	554	58	**66**	**420**	56	**58**	**1030**	**53**
DCCA	56	627	79	58	516	90	21	2095	913
GCN	60	546	45	62	485	71	54	1130	108
Text+Network (transductive)									
MLP−TXT+NET 1%	**8**	1521	1295	14	1436	1411	8	3865	2041
DCCA 1%	7	1425	979	38	869	348	14	3014	1367
GCN 1%	6	**1103**	**609**	**41**	**788**	**311**	**21**	**2071**	**853**

Table 1: Geolocation results over the three Twitter datasets for the proposed models: joint text+network MLP−TXT+NET, DCCA, and GCN and network-based GCN−LP. The models are compared with text-only and network-only methods. The performance of the three joint models is also reported for minimal supervision scenario where only 1% of the total samples are labelled. "—" signifies that no results were reported for the given metric or dataset. Note that Do et al. (2017) use timezone, and Miura et al. (2017) use the description and location fields in addition to text and network.

引用：Table 1

Performance

	GeoText			Twitter-US			Twitter-World		
	Acc@161↑	Mean↓	Median↓	Acc@161↑	Mean↓	Median↓	Acc@161↑	Mean↓	Median↓
Text (inductive)									
Rahimi et al. (2017b)	38	844	389	54	554	120	34	1456	415
Wing and Baldridge (2014)	—	—	—	48	686	191	31	1669	509
Cha et al. (2015)	—	581	425	—	—	—	—	—	—
Network (transductive)									
Rahimi et al. (2015a)	58	586	60	54	705	116	45	2525	279
GCN−LP	58	576	56	53	653	126	45	2357	279
Text+Network (transductive)									
Do et al. (2017)	**62**	**532**	**32**	**66**	433	**45**	53	1044	118
Miura et al. (2017)	—	—	—	61	481	65	—	—	—
Rahimi et al. (2017b)	59	578	61	61	515	77	53	1280	104
MLP−TXT+NET	58	554	58	**66**	**420**	56	**58**	**1030**	**53**
DCCA	56	627	79	58	516	90	21	2095	913
GCN	60	546	45	62	485	71	54	1130	108
Text+Network (transductive)									
MLP−TXT+NET 1%								3865	2041
DCCA 1%								3014	1367
GCN 1%								**2071**	**853**

MLP-TXT+NETとGCNはテキストまたはネットワーク情報のどちらかのみを使う方法に勝っている

Table 1: Geolocation results over the three Twitter datasets for the proposed models: joint text+network MLP−TXT+NET, DCCA, and GCN and network-based GCN−LP. The models are compared with text-only and network-only methods. The performance of the three joint models is also reported for minimal supervision scenario where only 1% of the total samples are labelled. "—" signifies that no results were reported for the given metric or dataset. Note that Do et al. (2017) use timezone, and Miura et al. (2017) use the description and location fields in addition to text and network.

引用：Table 1

Performance

	GeoText			Twitter-US			Twitter-World		
	Acc@161↑	Mean↓	Median↓	Acc@161↑	Mean↓	Median↓	Acc@161↑	Mean↓	Median↓
Text (inductive)									
Rahimi et al. (2017b)	38	844	389	54	554	120	34	1456	415
Wing and Baldridge (2014)	—	—	—	48	686	191	31	1669	509
Cha et al. (2015)	—	581	425	—	—	—	—	—	—
Network (transductive)									
Rahimi et al. (2015a)	58	586	60	54	705	116	45	2525	279
GCN-LP	58	576	56	53	653	126	45	2357	279
Text+Network (transductive)									
Do et al. (2017)	**62**	**532**	**32**	**66**	433	**45**	53	1044	118
Miura et al. (2017)	—	—	—	61	481	65	—	—	—
Rahimi et al. (2017b)	59	578	61	61	515	77	53	1280	104
MLP-TXT+NET	58	554	58	**66**	**420**	56	**58**	**1030**	**53**
DCCA	56	627	79	58	516	90	21	2095	913
GCN	60	546	45	62	485	71	54	1130	108
Text+Network (transductive)									
MLP-TXT+NET 1%	**8**						8	3865	2041
DCCA 1%	7						14	3014	1367
GCN 1%	6						**21**	**2071**	**853**

MLP-TXT+NETとGCNは
Rahimi (2017b)も上回っている

DoとMiuraはタイムゾーン情報やプロフィール情報も
使っているのでフェアじゃない（だから負けている）

Table 1: Ge… …ext+network MLP-TXT+… …ith text-only and networ… …for minimal supervision scenario where only 1% of the total samples are labeled. — signifies that no results were reported for the given metric or dataset. Note that Do et al. (2017) use timezone, and Miura et al. (2017) use the description and location fields in addition to text and network.

引用：Table 1

Performance

	GeoText			Twitter-US			Twitter-World		
	Acc@161↑	Mean↓	Median↓	Acc@161↑	Mean↓	Median↓	Acc@161↑	Mean↓	Median↓
Text (inductive)									
Rahimi et al. (2017b)	38	844	389	54	554	120	34	1456	415
Wing and Baldridge (2014)	—	—	—	48	686	191	31	1669	509
Cha et al. (2015)	—	581	425	—	—	—	—	—	—
Network (transductive)									
Rahimi et al. (2015a)	58	586	60	54	705	116	45	2525	279
GCN-LP	58	576	56	53	653	126	45	2357	279
Text+Network (transductive)									
Do et al. (2017)	**62**	**532**	**32**	**66**	433	**45**	53	1044	118
Miura et al. (2017)	—	—	—	61	481	65	—	—	—
Rahimi et al. (2017b)	59	578	61	61	515	77	53	1280	104
MLP-TXT+NET	58	554	58	**66**	**420**	56	**58**	**1030**	**53**
DCCA	56	627	79	58	516	90	21	2095	913
GCN	60	546	45	62	485	71	54	1130	108
Text+Network (transductive)									
MLP-TXT+NET 1%	**8**	1521	1295	14	1436	1411	8	3865	2041
DCCA 1%	7	1425	979	38	869	348	14	3014	1367
GCN 1%	6	**1103**	**609**	**41**	**788**	**311**	**21**	**2071**	**853**

現実的な設定（1%のサンプルにのみラベルがついている）
ではGCNとDCCAがMLP-TXT+NETを上回る
→学習するパラメータの数が影響

Table 1: C +network
MLP-TXT text-only
and netw minimal
supervisi sults were
reported for the given metric or dataset. Note that Do et al. (2017) use timezone, and Miura et al. (2017) use the description and location fields in addition to text and network.

引用：Table 1

Error Analysis

エラー距離が大きいのは田舎

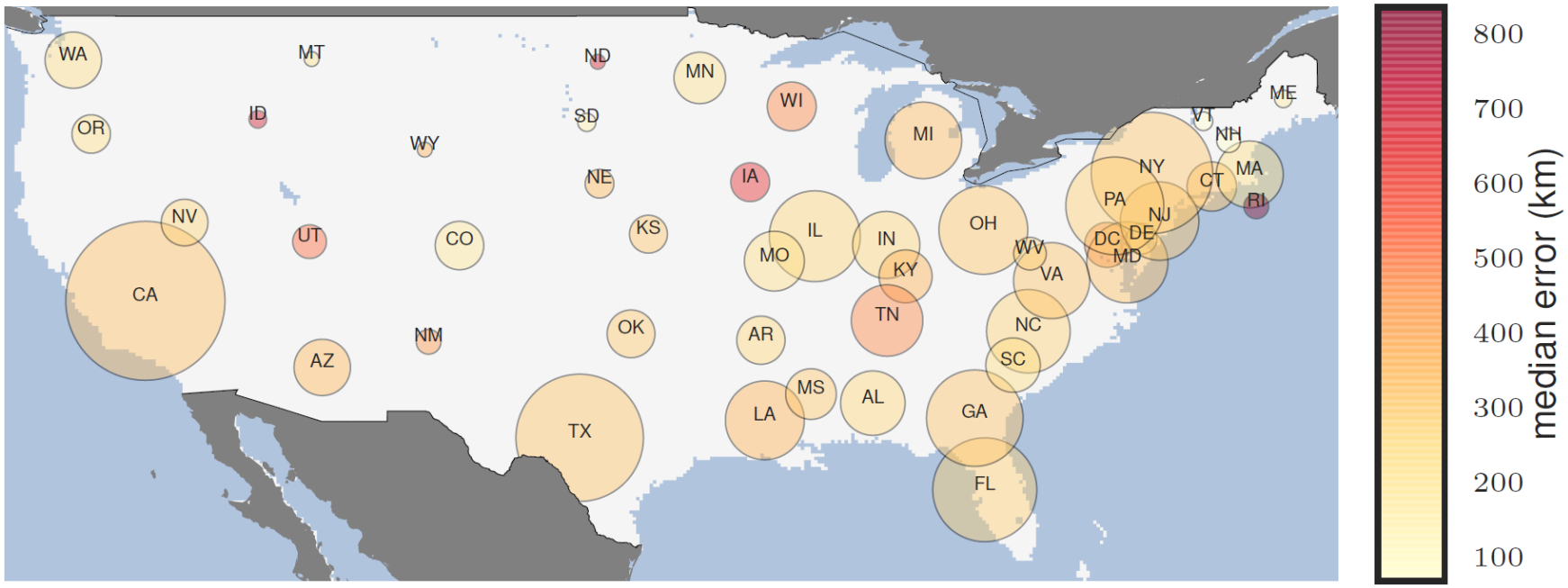


Figure 7: The geographical distribution of Median error of GCN using 1% of labelled data in each state over the development set of TWITTER-US. The colour indicates error and the size indicates the number of development users within the state.

Local Terms

Seattleの#goseahawksは
ラベル付きデータには入っていない単語

Seattle, WA	Austin, TX	Jacksonville, FL	Columbus, OH	Charlotte, NC	Phoenix, AZ	New Orleans, LA	Baltimore, MD
#goseahawks	stubb	unf	laffayette	#asheville	clutterbuck	mcneese	bhop
smock	gsd	ribault	#weareohio	#depinga	waffels	keela	#dsu
traffuck	#meatsweats	wahoowa	#arcgis	batesburg	bahumbug	pentecostals	chestertown
ferran	lanterna	wjct	#slammin	stewey	iedereen	lutcher	aduh
promissory	pupper	fscj	#ouhc	#bojangles	rockharbor	grogan	umbc
chowdown	effaced	floridian	#cow	#occupyraleigh	redtail	suela	lmt
ckrib	#austin	#jacksonville	mommyhood	gville	gewoon	cajuns	assistly
#uwhuskies	lmfbo	#mer	beering	sweezy	jms	bmu	slurpies

Table 2: Top terms for selected regions detected by GCN using only 1% of TWITTER-US for supervision. We present the terms that were present only in unlabelled data. The terms include city names, hashtags, food names and internet abbreviations.

Conclusion

- テキストとネットワークの情報を使ってユーザの場所を推定するモデルGCN, DCCA, MLP-TXT+NETを提案した
- 先行研究より性能が優れていることを示した
- GCNとDCCAがラベルデータの少ない環境でうまく機能することを示した

ソースコードとデータが公開されている
https://github.com/afshinrahimi/geographconv