Spatial Summarization of Image Collections

Diego A. Ballesteros Villamizar

ETH Zürich

January 27th, 2016

Outline

Correct normalization

1. In previous presentation, latitude and longitude were normalized to full range, i.e. $[-90, 90]$ and $[-180, 180]$ respectively.
2. All features normalized to $[0, 1]$ range **over the data**.
3. However, no improvement of score.
4. Moreover, the phenomena previously seen on the bmW weights is still present, i.e. they are the same across dimensions d.

Why uniform weights across dimensions?

1. *Note 1:* Initialization of the $\boldsymbol{B}$ weights is obtained from a uniform distribution over $[0, 0.001]$.
2. *Note 2:* The gradient update for $\boldsymbol{B}$ is given by equations 1 and 2:

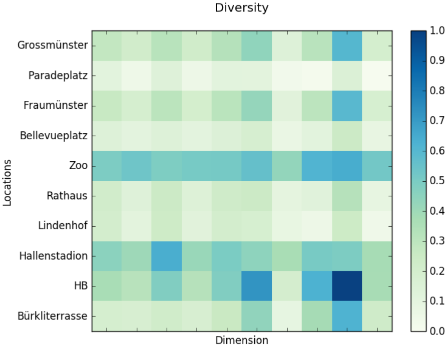
$$\left(\nabla_{\boldsymbol{B}} \log \frac{1}{\hat{Z}} \tilde{P}\left(S \mid \boldsymbol{a}, \boldsymbol{B}\right)\right)_{m,l} = x_{i^*,m} - \sum_{i \in S} x_{i,m} \tag{1}$$

$$i^* = \operatorname*{argmax}_{i \in S} \boldsymbol{x}_i \boldsymbol{b}_l \tag{2}$$

   Which shows that the only difference in the weights across dimensions is given by the value of i^*.
3. Because the initialization is small in comparison with the feature values and the variability is low, the value of i^* at every timestep is equal in all dimensions d as it only depends on the feature vector $\boldsymbol{x}_i$.

Fixing weight B distribution

1. To avoid initial updates choosing always the same i^*, the initialization for B should be larger. The new range is $[0, 1]$.
2. The learned weight matrix W is not uniform with the updated initialization condition.



Fixing weight B distribution

1. However, the scoring remains low. Worse than the score with the identity as feature matrix, i.e. the non-featurized model.

Model	Accuracy	MRR
Modular with features	17.38 ± 1.81	39.85 ± 1.52
FLID (d = 10)	$\mathbf{29.31 \pm 2.74}$	$\mathbf{52.19 \pm 1.78}$
FFLID (d = 1)	13.60 ± 1.60	37.54 ± 1.39
FFLID (d = 5)	13.48 ± 1.61	37.51 ± 1.45
FFLID (d = 10)	12.34 ± 1.51	36.94 ± 1.47

Outline

1. Featurized FLID
2. Extending the Model

Coherence term

$$P(S \mid \boldsymbol{a}, \boldsymbol{B}, \boldsymbol{C}) = \frac{1}{Z} \exp\left(\sum_{i \in S} u_i + Div(S, \boldsymbol{B}) + Coh(S, \boldsymbol{C})\right) \tag{3}$$

$$Div(S, \boldsymbol{B}) = \sum_{l=1}^{L} \left(\max_{i \in S} \boldsymbol{x}_i \boldsymbol{b}_l - \sum_{i \in S} \boldsymbol{x}_i \boldsymbol{b}_l\right) \tag{4}$$

$$Coh(S, \boldsymbol{C}) = \sum_{k=1}^{K} \left(\sum_{i \in S} \boldsymbol{x}_i \boldsymbol{c}_k - \max_{i \in S} \boldsymbol{x}_i \boldsymbol{c}_k\right) \tag{5}$$

$$\boldsymbol{u} = \boldsymbol{X}\boldsymbol{a} \quad \boldsymbol{X} \in \mathbb{R}^{|V| \times M} \quad \boldsymbol{u} \in \mathbb{R}^{|V|} \quad \boldsymbol{a} \in \mathbb{R}^{M} \tag{6}$$

$$\boldsymbol{W}_{\boldsymbol{B}} = \boldsymbol{X}\boldsymbol{B} \quad \boldsymbol{B} \in \mathbb{R}^{|M| \times L} \tag{7}$$

$$\boldsymbol{W}_{\boldsymbol{C}} = \boldsymbol{X}\boldsymbol{C} \quad \boldsymbol{C} \in \mathbb{R}^{|M| \times K} \tag{8}$$

$$\tag{9}$$

Intuition

1. Supermodular term encourages adding similar items, i.e. values with high $w_{c_{i,k}}$ close to the $\max_{i \in S}$.
2. This is a more natural notion in the setting of touristic places, e.g. someone who visits Fraumünster will likely visit Grossmünster as well.
3. The learning is analog to the diversity-only case, however more parameters are present so more noise samples should be used. The gradient update is the negative of the equation for the diversity case.

Zürich Path Data

Table: Frequency of Item Sets

Set	Frequency
$[0, 2]$	101
$[9, 3]$	66
$[0, 5]$	63
$[2, 5]$	62
$[5, 6]$	52
$[9, 1]$	50
$[2, 6]$	49
$[9, 2]$	44
$[0, 2, 5]$	43
$[0, 3]$	41

Table: Locations

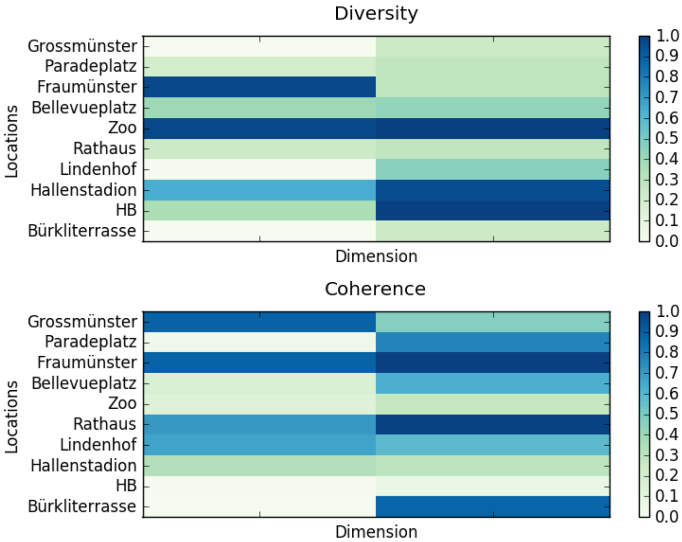
Index	Location
0	Grossmünster
1	Paradeplatz
2	Fraumünster
3	Bellevue
4	Zoo
5	Rathaus
6	Lindenhof
7	Hallenstadion
8	HB
9	Bürkliplatz

Model without Features

$$\boldsymbol{X} = \mathbb{I}$$

		K			
		0	2	5	10
	0	18.15 ± 3.08	24.85 ± 7.37	31.06 ± 6.89	30.72 ± 3.22
L	2	22.39 ± 2.66	$\mathbf{34.07 \pm 2.63}$		
	5	24.04 ± 3.22		34.35 ± 2.15	
	10	29.31 ± 2.74			35.48 ± 2.16

Model without Features

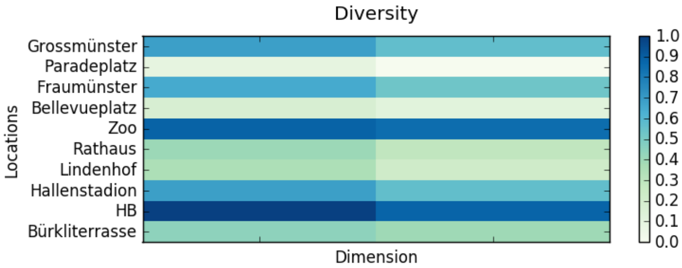


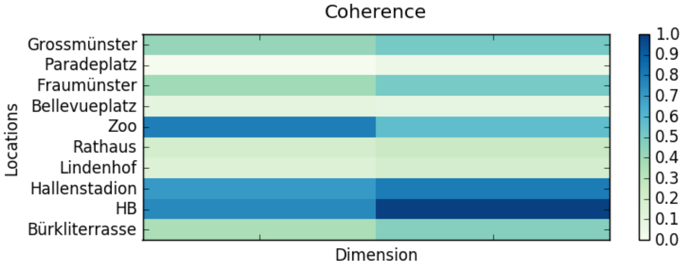
Model with Features

$$\boldsymbol{X} \in \mathbb{R}^{10\times 4}$$

		K			
		0	2	5	10
	0	17.38 ± 1.81	16.69 ± 1.67	16.77 ± 2.13	17.90 ± 1.65
L	2	13.39 ± 1.58	13.63 ± 1.64		
	5	13.48 ± 1.61		13.21 ± 1.44	
	10	12.34 ± 1.51			12.96 ± 1.14

Model with Features





Experiment Performance

1. As the number of parameters increase, what is the cost on running time performance?
2. As in Tschiatschek et al. (2016), let's define $\kappa = \max_{S \in \mathcal{D} \cup \mathcal{N}} |S|$.
3. The gradient update operations per iteration are:
    1. Updating the utility vector a: $O(\kappa M)$.
    2. Updating the weights B: $O(\kappa M L)$.
    3. Updating the weights C: $O(\kappa M K)$.
4. Then the overall performance is: $O(|\mathcal{D} \cup \mathcal{N}| \kappa M(L + K))$.
5. This is similar to the performance reported on Tschiatschek et al. (2016): $O(|\mathcal{D} \cup \mathcal{N}| \kappa D)$, as long as $L + K \approx D$ and $M \ll \kappa$.

References

Tschiatschek, S., Djolonga, J., and Krause, A. (2016). Learning probabilistic submodular diversity models via noise contrastive estimation. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*.