Geography-Aware Sequential Location Recommendation

Rao Xuan

2021-9-30

Research direction

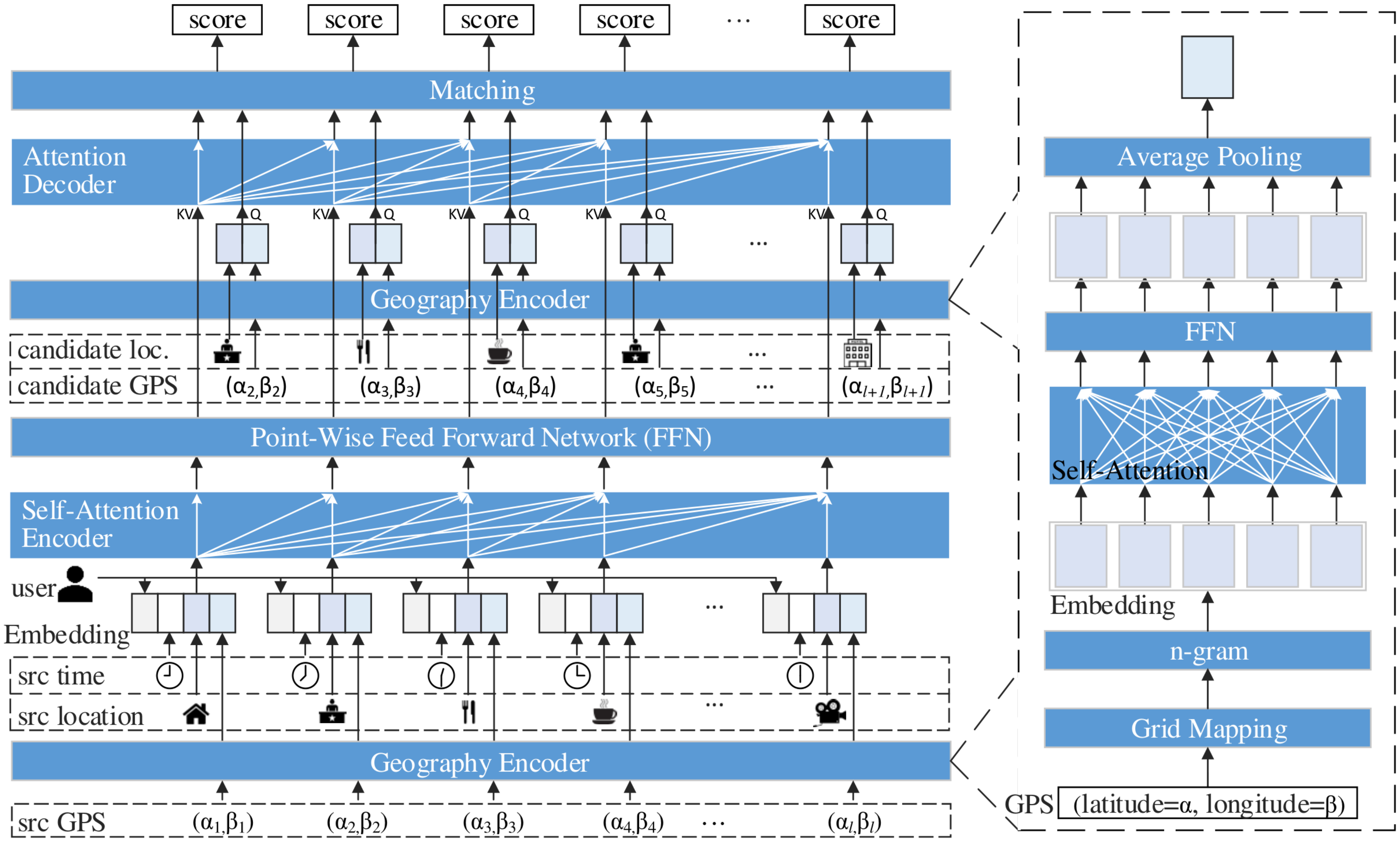
● Drawbacks

- Geographical information is still not effectively utilized.
- Use either the **BPR loss** or the **binary cross-entropy loss** for optimization by selecting **random samples** from unvisited locations.

◆ Solutions

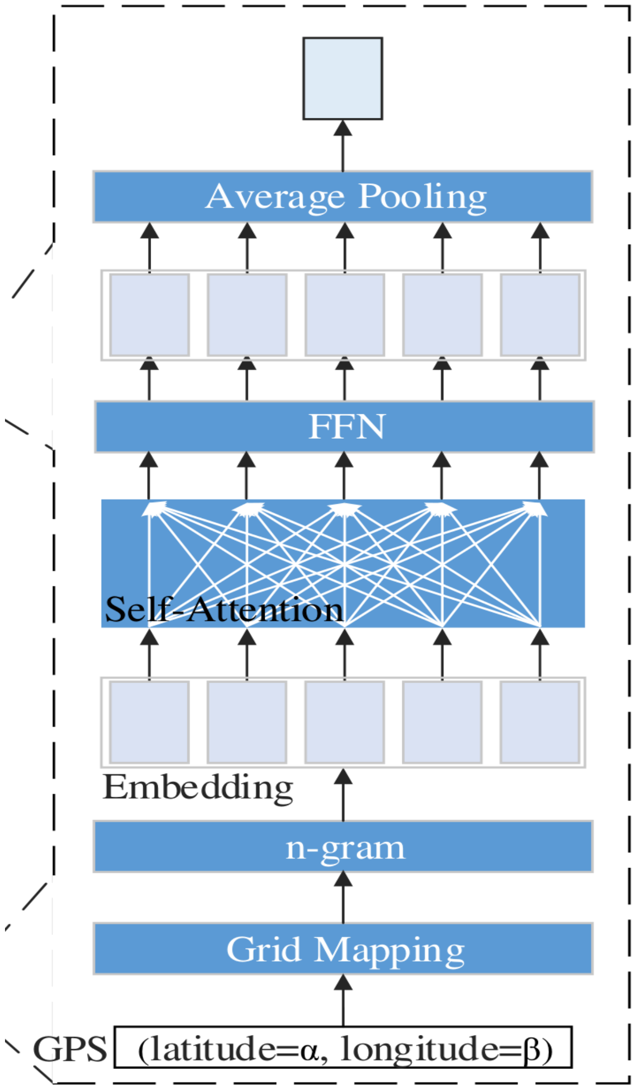
- Embed the exact GPS of the location with a novel **geography encoder**.
- Propose a **new loss function** based on **importance sampling** for optimization.
- Put forward **geography-aware negative samplers** to promote the informativeness of negative samples.

Framework



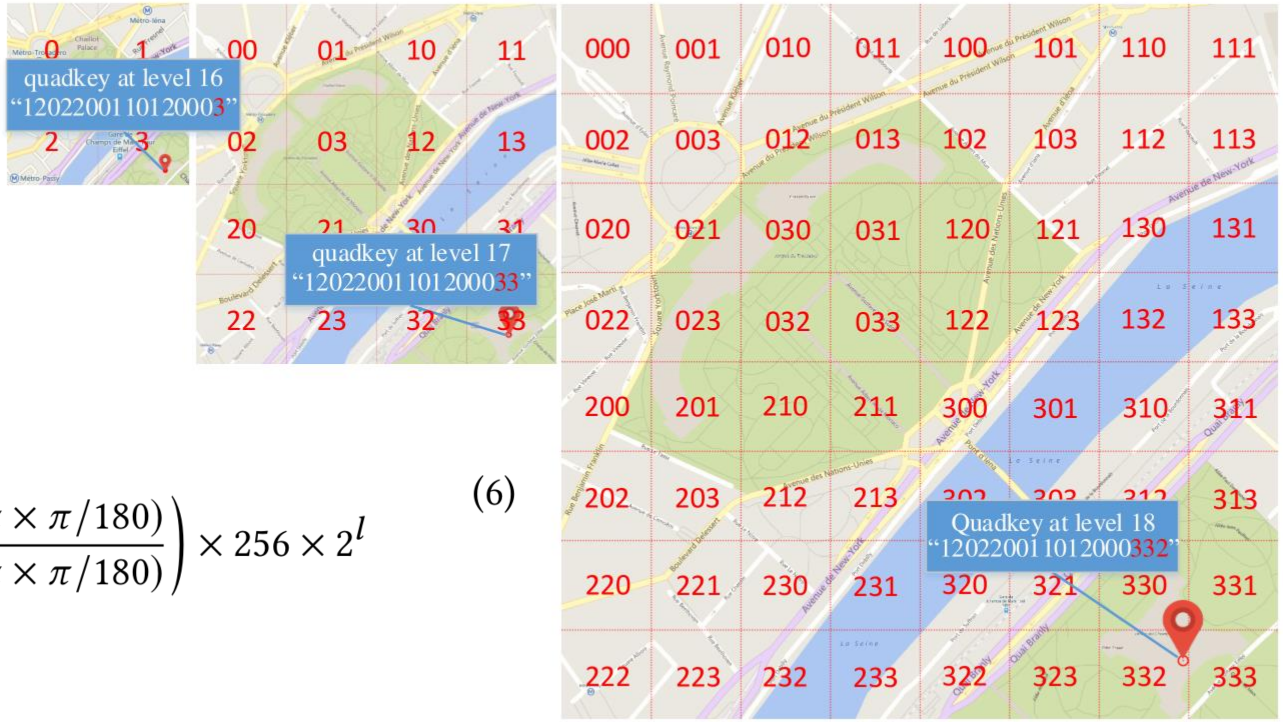
score
score
score
score
···
score
Matching
Attention Decoder
KV
Q
Geography Encoder
candidate loc.
candidate GPS
(α2,β2)
(α3,β3)
(α4,β4)
(α5,β5)
(αl+1,βl+1)
Point-Wise Feed Forward Network (FFN)
Self-Attention Encoder
user
Embedding
src time
src location
Geography Encoder
src GPS
(α1,β1)
(α2,β2)
(α3,β3)
(α4,β4)
(αl,βl)
Average Pooling
FFN
Self-Attention
Embedding
n-gram
Grid Mapping
GPS
(latitude=α, longitude=β)

Geography encoder



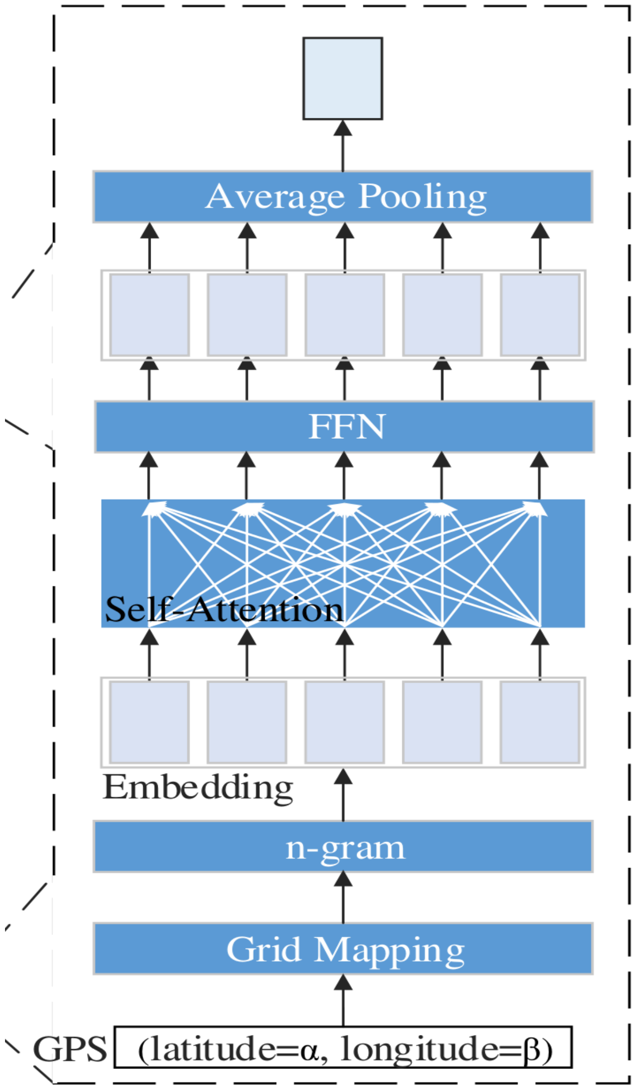
Map Gridding

- Project the entire world into a flat plane, which is cut into 4 grids of 256 x 256 pixels.
- Each grid is divided into four sub-grids of the same size when the number of levels increases by one.



$$x = \frac{\beta + 180}{360} \times 256 \times 2^{l}$$
$$y = \left(\frac{1}{2} - \frac{1}{4\pi} \log \frac{1 + \sin(\alpha \times \pi/180)}{1 - \sin(\alpha \times \pi/180)}\right) \times 256 \times 2^{l} \qquad (6)$$

Geography encoder



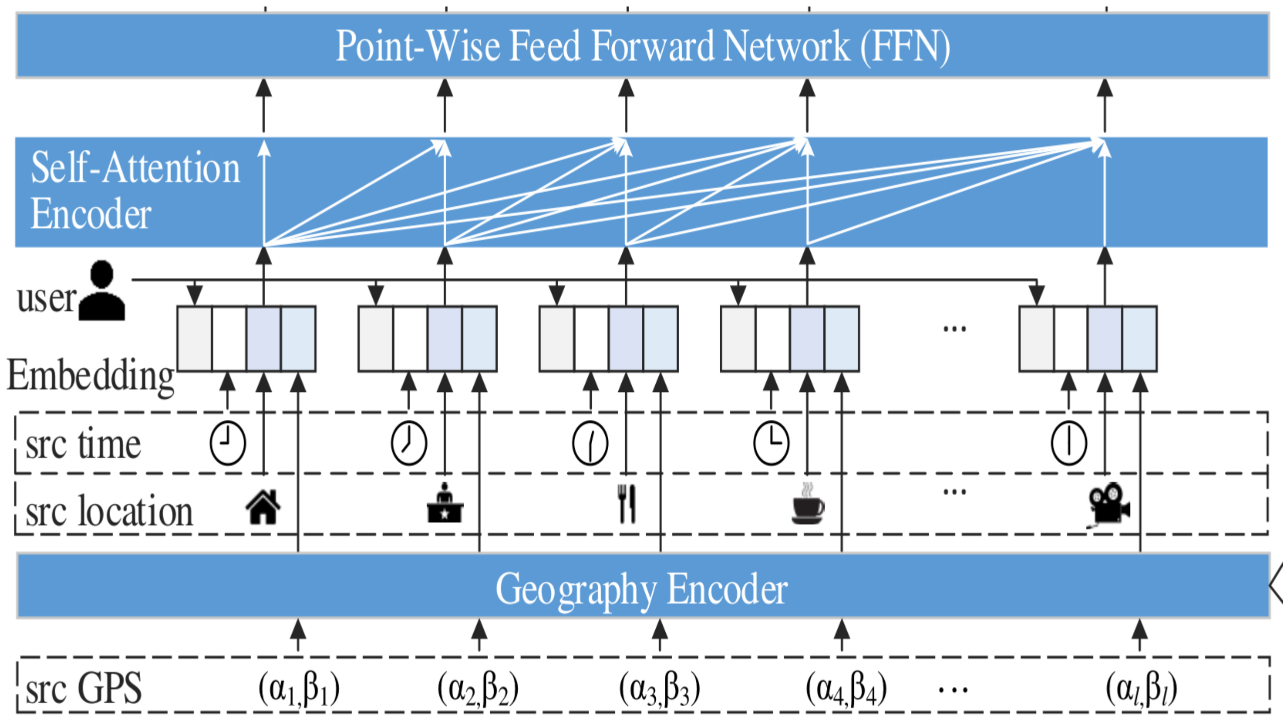
● Encoding Quadkeys

- Because of grid division like quadtree, each grid can be identified with a **quadtree key** (quadkey for short), which can be interpreted as a base-4 number and whose length equals the level of detail.

- Transform each quadkey into the sequence of **n-grams** first.

- After embedding the sequence of n-grams, we apply a stacked self-attention network for capturing sequential dependence, and then aggregate the sequence of n-gram representations via average pooling.

Self-Attention encoder



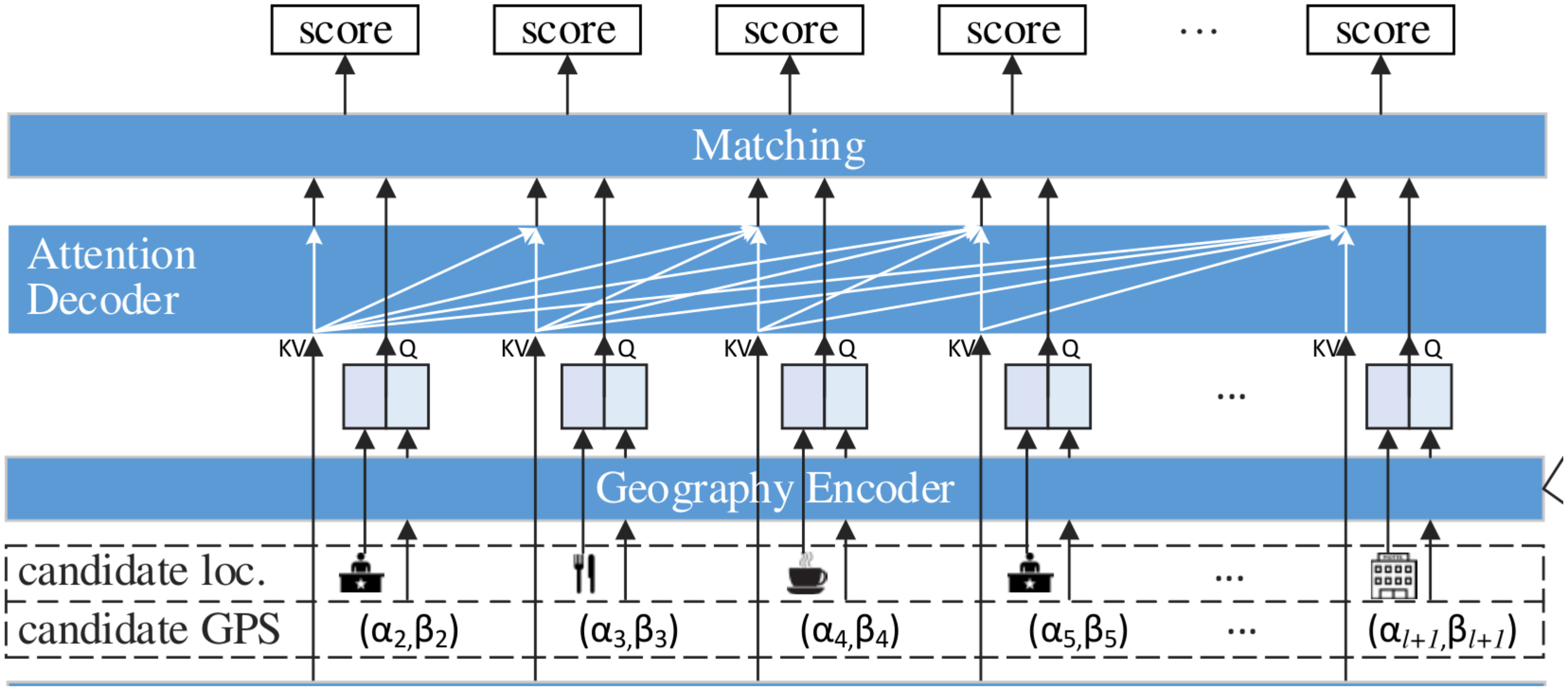
- Embed **user**, **hour** of week and **location** of each behavior, and encode the **exact GPS position** with geography encoder. These vectors are concatenated, forming the representation matrix of the input sequence.
- The self-attention encoder stacks **multiple** self-attention blocks, each of which consists of a **self-attention layer** and a **point-wise feed-forward** network (FFN).
- When stacking multiple self-attention blocks, residual connection and layer normalization are applied in FFN and the self-attention layer.

$$\boldsymbol{S} = \mathrm{SA}(\boldsymbol{E}) = \mathrm{Attention}(\boldsymbol{E}\boldsymbol{W}_Q, \boldsymbol{E}\boldsymbol{W}_K, \boldsymbol{E}\boldsymbol{W}_V) \tag{1}$$

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V} \tag{2}$$

$$\boldsymbol{F}_i = \mathrm{FFN}(\boldsymbol{S}_i) = \max(0, \boldsymbol{S}_i\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{3}$$

Target-aware Attention decoder



- Most existing self-attention based recommenders **directly** feed these outputs into the matching module, which may be suboptimal.
- Matching module: compute **preference score** for each candidate location with any matching function.

$$A = \text{decoder}(F^{(l)}|T) = \text{Attention}(T, F^{(l)}W, F^{(l)}) \qquad (4)$$

$$y_{i,j} = f(A_i, T_j), \qquad (5)$$

Loss Function with Importance Sampling

- Unvisited locations with **large preference scores** can contribute more to gradient, so they are **more informative** and should be sampled with **high probability**.

$$-\sum_{S^u \in S} \sum_{i=1}^{n} \left(\log \sigma(y_{i,o_i}) + \sum_{k \notin L^u} P(k|i) \log(1 - \sigma(y_{i,k})) \right), \qquad (8)$$

$$P(k|i) = \frac{\exp(r_{i,k}/T)}{\sum_{k' \notin L^u} \exp(r_{i,k'}/T)} \qquad (9)$$

- This still suffers from **low efficiency** of computing **normalization** in the probability.

- Propose to approximate the expectation with importance sampling.

- Suppose the proposal distribution is Q(k|i), denote by ˜Q(k|i) the unnormalized probability of Q(k|i).

$$-\sum_{S^u \in S} \sum_{i=1}^{n} \left(\log \sigma(y_{i,o_i}) + \sum_{k=1}^{K} w_k \log\left(1 - \sigma(y_{i,k})\right) \right), \qquad (10)$$

$$w_k = \frac{\exp\left(r_{i,k}/T - \ln \tilde{Q}(k|i)\right)}{\sum_{k'=1}^{K} \exp\left(r_{i,k'}/T - \ln \tilde{Q}(k'|i)\right)}$$

Geography-aware Negative Sampler

KNN-uniform based Negative sampling

- **geographical information** can also be effective to distinguish negative from potentially positive in unvisited locations.
- For example, when he/she visits the **target location** O, the unvisited locations **around O** maybe more likely to be **negative**.

◆ Retrieve K nearest locations to the target location.

◆ Randomly draw negative samples from these K candidates.