Data Science, 2019 Spring

Verifying whether the regions with high fine dust share the similar features.[1)]

2017320229, 심규현, 2017320202 박지수, 2017320217 박경남

요 약

최근 미세먼지 농도 증가에 따라, 미세먼지의 위험성에 대한 각종 연구들과 호흡기 및 시정장애 환자의 급증 현상이 각종 언론과 학계에서 보고되고 있다. 이 보고서에서는 고농도 미세먼지 현상을 설명하기 위한 다양한 가정 중 서울의 지역별 지리, 건물분포에 대한 특징에 따라 미세먼지 농도가 변한다는 가설을 검증한다. 첫 번째로 가설을 검증하기 위한 문제를 정의한다. 두 번째로 서로 다른 규모의 데이터를 같은 규모로 보관하는 방법을 소개하고, 딥러닝을 통해 지역 구간별 특징을 추출하는 방법을 소개한다. 결과에서는 미세먼지 데이터와 지역별 특징 벡터를 분석하고 비교하여 PoIs의 분포와 미세먼지 수치 간의 상관관계가 있음을 보인다.

1. 서 론

최근 미세먼지 농드 증가에 따라, 미세먼지의 위험성에 대한 각종 연구들과 호흡기 및 시정장애 환자의 급증 현상이 각종 언론과 학계에서 보고되었다. 이는 미세먼지 문제가 더이상 개인의 부주의 문제가 아닌 범국가적 차원의 문제임을 시사한다. 미세먼지 농도에 영향을 미치는 주된 요인은 기후여건과 지형적 특성이다[1]. 지형적 특성은 기후여건에 비해 정적인 속성이므로, 지형의 특성과 미세먼지 농도의 상관관계를 고려함으로써 지역 기반의 장기적인 미세먼지 정책 수행이 가능하다. 따라서 미세먼지 농도 데이터와 지역의 유사도 데이터를 비교분석 함으로써, 정책 수행의 효과를 예측할 수 있다.

본 보고서에서는 가설을 검증하기 위한 문제를 정의한다. 해당 가설은 주관적인 판단으로 이뤄질 수 있는 문제이므로, 파라미터를 사용자가 정의할 수 있도록 한다. 또한 서울을 중심으로 가설을 검증하는데, 필요한 미세먼지 데이터가 지리적으로 희소하게 존재했다. 이를 보간하기(interpolate) 위한 방법을 소개한다. 그리고 Efficient Similar Region Search with Deep Metric Learning[2]에서 특정 지역의 건물 분포와 종류에 대한 정보를 포함하는 벡터를 생성하는 방법이 제안되었다. 이를 사용하여 지역별 특징 벡터를 생성한다. 이를 이용해 지역별 유사도에 대한 순위를 만들고 보간한 미세먼지 데이터를 비교하여 가설을 검증한다.

2. 문제 정의

"높은 농도의 미세먼지를 가진 지역은 비슷한 특징을 가질 것이다."라는 가설을 증명하기 위해 아래의 문제를 정의한다.

정의 1. 미세먼지가 높은 지역 R_h는 아래와 같다.
시간별 미세먼지 평균 농도를 기준으로 지역들을 내림차순 했을 때 상위 x% 이내의 지역들을 말한다.(이때 x는 사용자에 의해 정의된다.)

지역 간의 유사도를 측정하기 위해 그 지역의 PoIs(건물 객체)를 벡터로 표현해야 한다. 그리고 아래의 정의를 이용해 서로 다른 두 지역의 유사도를 측정한다.

정의 2. 두 지역 R_1, R_2가 주어졌을 때, 두 지역의 특성 벡터를 v_1, v_2라고 하자. 이 때 두 지역의 유사도 $\mathrm{sim}(R_1,R_2)$는 다음과 아래와 같다.

---

1) 데이터 분석에 활용한 Colaboratory notebook 코드가 너무 많아 ./code에 첨부하였습니다.

$$\mathrm{sim}(\mathrm{R}_1,\mathrm{R}_2) = \frac{1}{\|\mathrm{v}_1 - \mathrm{v}_2\|} \quad (2)$$

두 벡터의 거리는 노름(norm)으로 계산할 수 있다. 그리고 두 벡터의 유사도를 나타내는 sim은 두 벡터 간의 거리가 멀수록 작아지도록 하였다. 이 실험의 가설은 미세먼지가 높은 지역(정의 1)이 높은 유사도(정의 2)를 가진다는 것이다. 이를 판단하기 위한 기준을 아래와 같이 설정한다.

정의 3. 미세먼지가 높은 지역의 집합을 RH, 모든 지역의 집합을 R이라고 할 때, 아래의 조건을 만족하면 가설은 참이다.

$$\forall R_1, R_2 \in RH \ \wedge \ \forall R_3, R_4 \in R, \sim(\mathrm{R}_1,\mathrm{R}_2) \geq \sim(\mathrm{R}_3,\mathrm{R}_4) \quad (3)$$

공간적 특징만이 특정 지역의 미세먼지 농도를 결정짓지 않는다. 따라서 미세먼지 농도가 높은 날 특정 시간에서 **정의 3**을 만족하면 가설이 참이라고 판단한다.

3. 데이터 수집

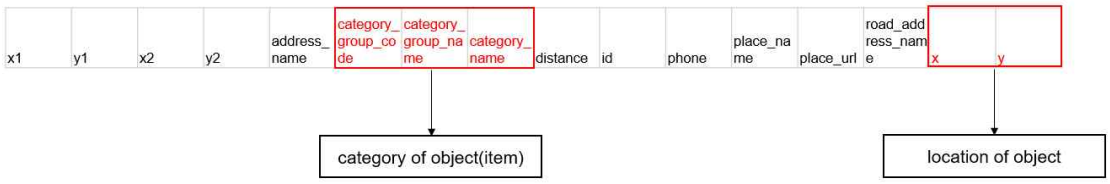
서울의 PoIs 데이터는 카카오 서비스 API[3]에서 수집하였다. REST API의 로컬 항목에서 카테고리로 장소 검색과 키워드로 장소 검색 두 가지 서비스를 이용하였다. 키워드로 장소 검색의 경우에는 편의점, 학교, 지하철역 등 기본으로 제공되는 18개의 카테고리 그룹 코드 전부에 대하여 장소데이터를 수집했다. 이때 응답에는 검색어에 검색된 문서 수와 관계없이 노출 가능한 문서 수가 최대 45로 정해져 있었다. 따라서 빠지는 자료가 없게 하도록 서울의 경계좌표로 만든 직사각형을 100×100으로 쪼갠 뒤, 총 10000개의 사각형 범위에 대해 장소 데이터를 요청하였다. 키워드로 장소 검색의 경우에는 기본 제공 카테고리에 속하지 않았으나 실험에 필요하다고 판단되는 분류인 '공장'의 위치 정보를 얻기 위하여 활용했으며, 카테고리로 장소 검색과 같은 방법으로 요청하였다.

미세먼지 데이터는 공공데이터포털[4]의 한국환경공단에서 제공하는 대기오염 정보 조회 서비스 REST API 중 측정소별 실시간 측정정보 조회 기능을 이용하였다. 서울의 25개 구마다 위치하는 도시대기 측정소의 최근 3개월간 측정 결과를 요청하였다. 그 결과, 공통으로 2019년 3월 1일 0시부터 2019년 5월 28일 23시까지 1시간 간격의 자료를 얻을 수 있었다.

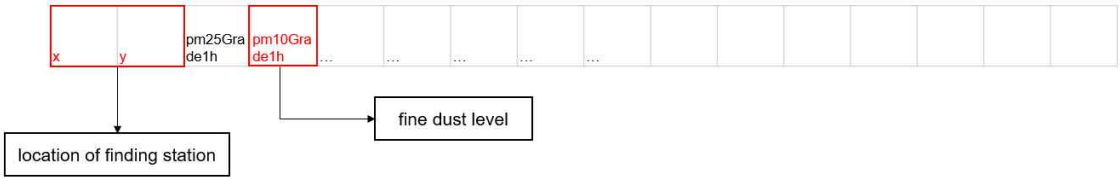
또한, 앞서 이용한 카카오 서비스 API에서 좌표-행정구역정보 변환 서비스를 이용하였다. 실험에서는 서울을 위도, 경도 각각 0.01 단위로 쪼갠 셀을 하나의 단위로 사용하였다. 이때 이 셀이 서울에 있는지를 판단하는 기준을 셀의 중심좌표의 행정구역으로 정하였다. 따라서 각 셀의 중심좌표를 행정구역정보로 변환한 데이터가 필요했고, 이에 해당 서비스에서 데이터를 요청하였다.

4. 데이터 가공 및 분석

[그림 1]은 카카오 서비스 API를 통해 수집한 데이터를 보여준다. 하나의 PoIs가 여러 속성을 가지고 있지만, 데이터 분석을 위해 필요한 것은 PoIs의 좌표와



[그림 1] 카카오 API를 통해 수집한 PoIs 데이터



[그림 2] 공공데이터포털을 통해 수집한 미세먼지 데이터

시간	강남구	강동구	강북구	강서구	...
2019-05-28 23:00	12	15	15	15	...
2019-05-28 22:00	12	12	16	13	...
2019-05-28 21:00	10	7	11	10	...
...	...	...	...	...	...

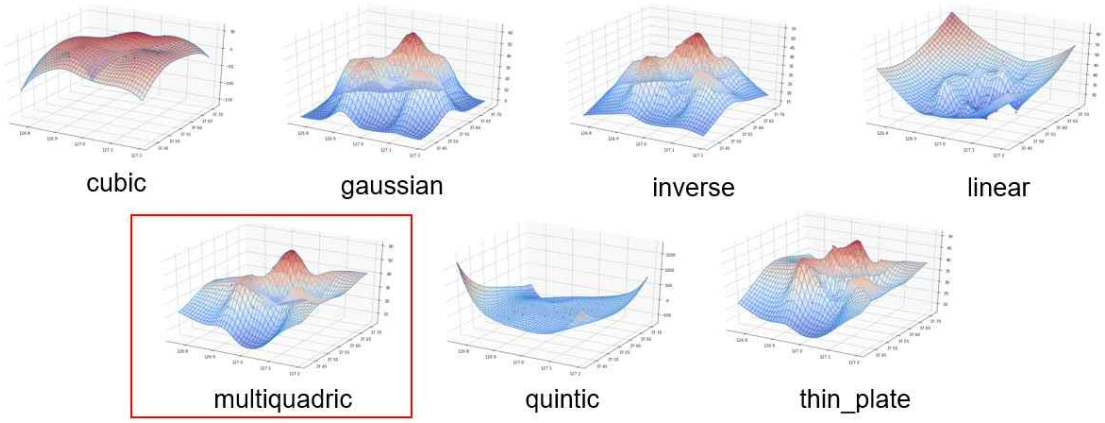
[표 1] 미세먼지 데이터를 가공하여 통합한 결과

종류이다. 이 정보를 추출하여 이용하였다.

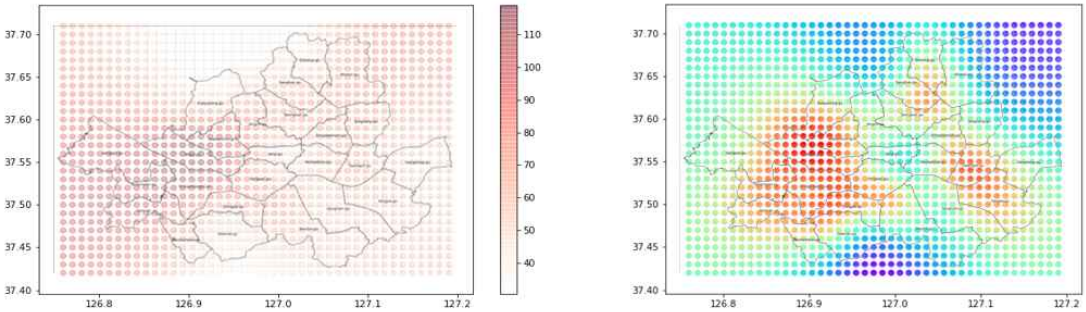
[그림 2]는 공공데이터포털을 통해 수집한 미세먼지 데이터를 보여준다. 한 행이 같는 특성은 측정 시간, 오존 수치, $PM_{2.5}$ 수치, PM_{10} 수치 등을 비롯하여 총 33개이다. 그중에서 실험의 목적과 맞는 미세먼지, 즉 PM_{10} 수치와 비교를 위한 측정 시간, 측정소의 위치를 추출하여 이용하였다. 남은 자료는 측정 시간과 측정소를 축으로 하여 미세먼지 수치를 나타낼 수 있게 하나의 파일로 통합했다. 그 결과는 [표 1]과 같다.

가. 미세먼지 수치 보간

학습을 위해서는 셀마다 미세먼지 수치를 알아야 하는데, 측정소는 구마다 하나 밖에 없다. 측정소 사이와 밖에 있는 미세먼지 데이터를 보간하기 위해 컴퓨터그래픽스 분야에서 연구되어 온 보간(interpolation)을 이용하였다. scipy.interpolate.Rbf에서는 7가지 보간 및 추정 함수를 지원한다. 각 셀이 몇 개의 방향을 참고하여 보간되었는 지를 나타내는 비선형성과 해당 기술의 신뢰도를 나타내는 인용수를 참고하여 보간방법을 정하였다. [그림 3]은 7개의 함수를 적용



[그림 3] 셀마다 보간된 미세먼지 값을 나타내는 그래프



[그림 4] 보간된 미세먼지 데이터를 시각화한 결과

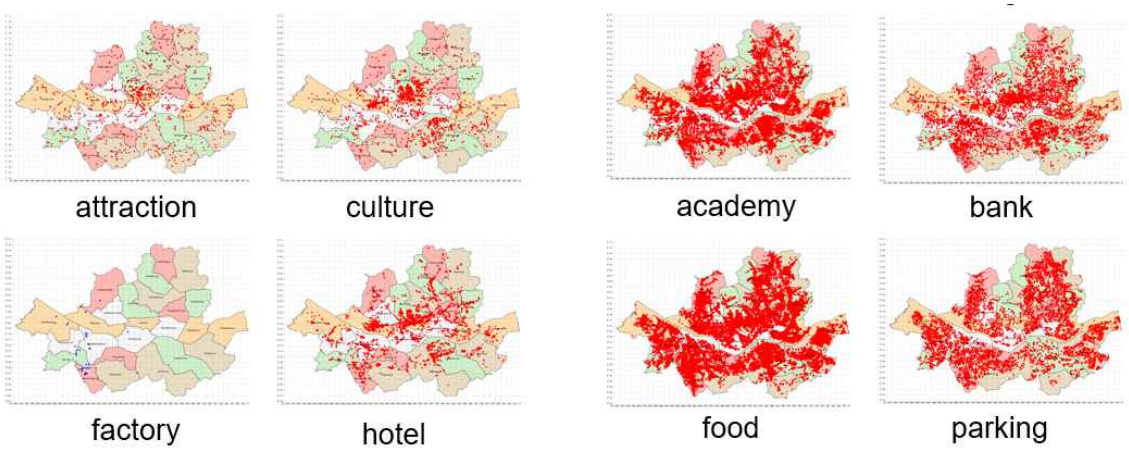
하여 보간한 결과를 나타낸다. multiquadric[5]이 가장 많은 방면의 데이터를 참고하여 보간을 시행하기 때문에 곡면이 가장 부드러워 후에 학습이 필요할 때 다루기 쉽고, 인용수 또한 319로 충분히 신뢰있는 방법이라고 판단하였다. [그림 4]는 결과를 2차원 그래프에 시각화한 것을 보여준다.

나. PoIs 데이터 가공

[그림 5]는 PoIs 데이터의 좌표를 이용하여 서울 지도에 사영한(project) 결과를 보여준다. 시각화를 통해 분포가 매우 균등(uniform)하고 거의 모든 셀에 포함된 카테고리와 분포가 균등하지 않고(skewed) 일부 셀에 포함된 카테고리가 있었다. 전자의 경우 지역의 특징 벡터를 생성하고 지역을 구별하는 데 영향을 않는다 판단했다. 그래서 [그림 5]의 attraction, culture, factory, hotel과 같이 특정 셀에 분포하고 밀도가 적은 데이터만 특징 벡터 생성에 사용하기로 하였다.

다. 특징 이미지 생성

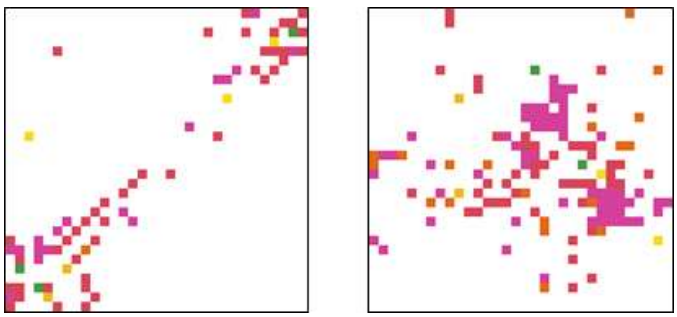
Efficient Similar Region Search with Deep Metric Learning[2]에서는 지역의 특징을 나타내는 벡터를 생성하는 방법을 제안했다. PoIs의 분포를 담은 이미지를 생성 후 CNN 계층과 Triplet[6]을 이용하여 유사도 함수를 학습한다. 그래서 2.나. PoIs 데이터 가공한 결과 나온 데이터를 가지고 입력 이미지를 생성했다. 이를 특징 이미지라고 한다. [그림 6]은 특징 이미지 생성에 사용된 PoIs-RGA 대응 테이블을 보여준다. 하나의 PoIs는 이미지의 한 픽셀에 대응되어 카테고리에 대응되는 색깔을 나타낸다. 이를 이용하여 32x32 이미지를 생성하였고 결과는 [그림 6]과 같다.



[그림 5] PoIs 데이터의 좌표를 이용하여 서울 지도에 사영한 결과

```
map={
    "academy":(96,56,45), #dark skin
    "attraction":(206,142,123), #light skin
    "bank":(85,112,161), #blue sky
    "brokerage":(77,102,46), #foliage
    "cafe":(129,118,165), #blue flower
    "convenience":(114,199,176), #bluish green
    "culture":(219,104,24), #orange
    "factory":(22,22,22), #black
    "food":(56,89,174), #purplish blue
    "hospital":(211,67,87), #moderate red
    "hotel":(207,62,151), #magenta
    "kinder":(160,193,57), #yellow green
    "market":(230,182,29), #orange yellow
    "oil_station":(26,32,145), #blue
    "parking":(72,148,65), #green
    "pharmacy":(197,27,37), #red
    "public_inst":(241,212,36) #yellow
}
```

[그림 6] 특징 이미지 생성에 사용된 PoIs-RGB 매핑 테이블



[그림 7] 생성된 특징 이미지

5. 데이터 학습

가. 특징 벡터 생성

데이터수집 및 가공 뒤, 의미 있는 정보를 예측 및 도출하기 위해 모델링 작업이 필요하다. 제대로 된 결과를 도출하기 위해서는 적절한 모델의 선택이 필요하다. 이 실험에서의 목표는 유사도가 높은 지역 간의 논리적인 거리를 줄이도록 학습하여 비슷한 PoIs 분포를 갖는 셀 간의 거리를 줄이는 것이다. 이 실험에서는 유사한 데이터의 거리를 0으로 수렴시키는 metric을 갖는 triplet network[6]를 활용하였다. 이 실험에서 활용된 메트릭과 Loss는 [그림 8]과 같다.

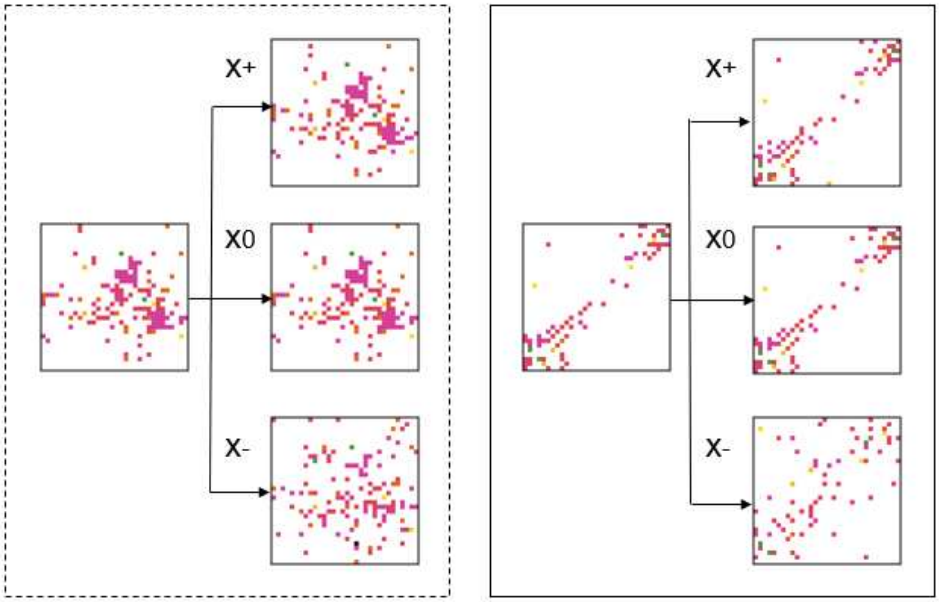
triplet network은 하나의 데이터를 학습시킬 때 원본 데이터 x_0와, x_0에 적은 noise를 더한 x_+, 많은 noise를 추가하여 다른 클래스로 분류돼야 할 x_-를 입력으로 받는다. 이 실험에서는 x_+에 10% noise를 추가하였고, x_-에는 50%의

$$d_+ = \frac{e^{\|Net(x)-Net(x^+)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2} + e^{\|Net(x)-Net(x^-)\|_2}}$$

$$d_- = \frac{e^{\|Net(x)-Net(x^-)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2} + e^{\|Net(x)-Net(x^-)\|_2}}.$$

$$Loss(d_+, d_-) \rightarrow 0 \text{ iff } \frac{\|Net(x)-Net(x^+)\|}{\|Net(x)-Net(x^-)\|} \rightarrow 0$$

[그림 8] triplet network에서 사용된 손실함수



[그림 9] 학습할 입력 데이터

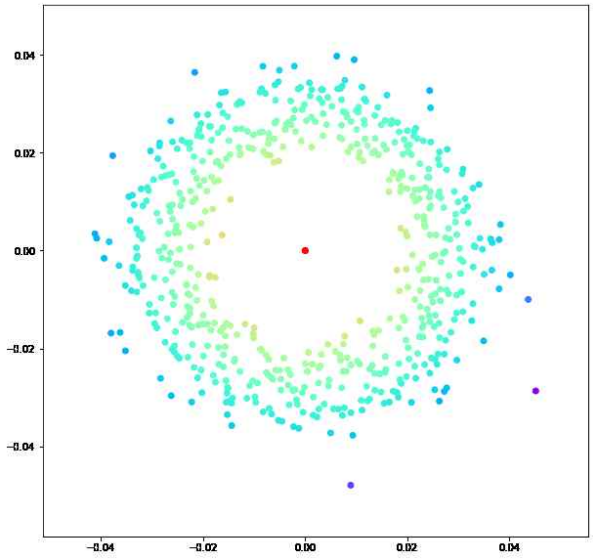
noise를 추가하였다. noise는 Efficient Similar Region Search with Deep Metric Learning[2]에서 정의되었다. [그림 9]은 각 데이터에 대해 x_+와 x_-를 보여준다.

triplet network는 각각의 입력에 대한 embedded network가 각각 하나씩 존재한다. 해당 코드의 경우, embedded network는 CNN으로 구현되어 있다. 해당 CNN은 conv1은 (3, 10, kernel_size=5)이고, conv2는 (10, 20, kernel_size=5), fc1(fully connected node)는 (10, 50), fc2는 (50, 10)으로 구성되어있는 네트워크로, 출력은 50×1 벡터로 나타난다. 학습 시에는 32 × 32의 x_0, x_+와 x_- 를 입력으로 넣는다. 학습 후 결과를 도출할 때는 특정 셀 c의 특징 이미지를 x_0, x_+에 넣고, x_-에 c와 비교할 다른 셀 c'의 특징 이미지를 넣는다. 그 결과 [그림 10]의 벡터를 얻을 수 있다. [그림 11]은 c의 결과 벡터를 중심으로 c'의 결과 벡터를 이차원 평면에 시각화한 것이다. 각 점은 모든 c'을 나타내며 중심과의 거리는 $d = \|c' - c\|$를 나타낸다. d가 클수록 셀 간의 특징이 다르다는 것을 의미한다.

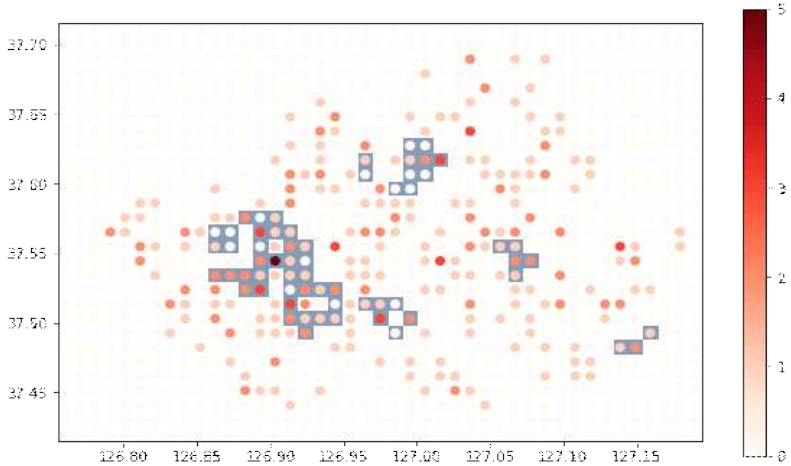
6. 결과

```
tensor([1.0367e-02, 6.6945e-03, 6.1685e-03, 7.0397e-03, 8.3771e-03, 8.3591e-03,
        6.4552e-03, 6.2223e-03, 1.3288e-02, 1.0715e-02, 5.0273e-03, 2.8213e-03,
        9.4373e-03, 1.2803e-02, 6.8384e-03, 2.7707e-03, 7.1757e-03, 3.7348e-03,
        3.4851e-03, 3.4254e-03, 3.1623e-06, 3.1623e-06, 5.6952e-03, 7.4830e-03,
        1.2924e-02, 4.5831e-03, 3.7333e-03, 1.0552e-02, 4.2145e-03, 5.4165e-03,
        4.2557e-03, 3.2425e-03, 1.1747e-02, 1.2575e-02, 8.9830e-03, 3.7621e-03,
        5.5539e-04, 8.4529e-03, 9.1091e-03, 7.5360e-03, 1.2032e-02, 5.9275e-03,
        8.7390e-03, 7.8946e-03, 1.2398e-02, 2.5809e-02, 5.7441e-03, 3.8305e-03,
        3.1623e-06, 3.1623e-06], grad_fn=<NormBackward1>)
```

[그림 10] triplet network에서의 결과 벡터 형태



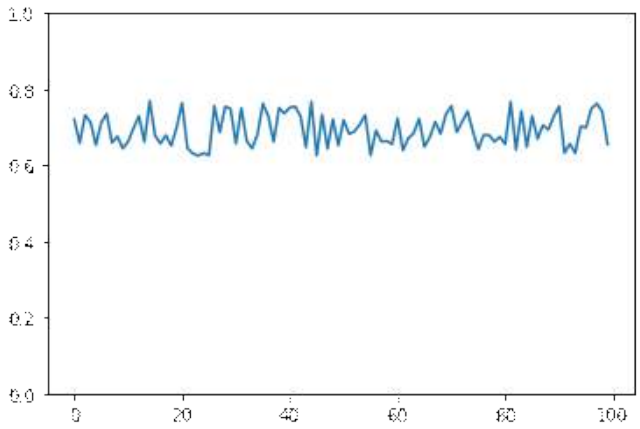
[그림 11] 임의의 셀을 중심으로 한 거리 시각화



[그림 12] 미세먼지가 높은 지역에 대한 상위 5개 빈도 heatmap

결과를 분석하기 위해서는 미세먼지를 수치를 보여주는 그리드와 셀 간의 거리를 나타내는 그리드를 비교해야 한다. 먼저 시각적으로 특정 셀을 분석하고자 한다. [그림 12]은 서울에서의 미세먼지 농도 상위 5% 구역(총 62개 구역)을 회색으로 강조한 것을 나타낸다. 이 62개의 셀에 대해 d가 작은 셀들을 전부 추출하였다. 그리고 [그림 12]는 셀이 나타난 빈도를 최대 5로 정규화하여 나타낸 heat map을 보여준다. 정의3을 만족하기 위해서는 회색으로 색칠된 셀에서 빈도가 그렇지 않은 셀의 빈도보다 커야 한다. 직관적으로 보면 회색으로 색칠된 곳에서 높은 빈도를 가진 셀들이 많지만, 빈도가 매우 낮은 셀들도 조금씩 존재한다. 다른 시간에 대해 결과를 도출해도 이러한 현상이 발생하였다.

문제는 데이터를 제작할 때 구(강남구, 성북구)를 표현하기 위해 큰 지역을 작은 셀로 나누어 표현한 것으로부터 발생하였다. 이러한 방법으로 인해 같은 지역(구)를 나타내는 셀임에도 불구하고, c의 특징 벡터 v_C 간의 편차가 크게 생겼다. 따라서 바로 옆의 위치임에도 불구하고 완전히 다른 특징 분포를 갖는다고 판단되는



[그림 13] 시간-비율 그래프

현상이 생겼다.

[그림 13]은 x축을 미세먼지가 높은 날의 시간으로, y축을 미세먼지가 높은 셀과 특징이 비슷한 셀이 겹치는 비율을 나타낸 그래프이다. 모든 시간대에서 비율이 0.6에서 0.8 사이로 가지고 있다. 이를 통해 정의 3을 엄격하게 만족하지는 않지만, 건물로 인해 나타난 특징과 미세먼지의 농도는 상관관계를 갖고 있다고 판단하였다. 그리고 모든 시간에 대해 높은 y값을 갖지 않는 이유는 너무 그리드를 잘게 쪼개 비슷한 같은 지역(구)임에도 불구하고, 편향이 생겼기 때문이라고 결론 내렸다.

7. 참고 문헌

[1] 이용기, 이기종, 이재성, & 신은상. (2012). 미세먼지 입경농도 분포의 지역별 특성. 한국대기환경학회지 (국문), 28(6), 666-674.

[2] Liu, Y., Zhao, K., & Cong, G. (2018, July). Efficient Similar Region Search with Deep Metric Learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1850-1859). ACM.

[3] Kakao service API, https://developers.kakao.com/features/kakao.

[4] public data portal, https://www.data.go.kr/dataset/15000581.

[5] Golberg, M. A., Chen, C. S., & Karur, S. R. (1996). Improved multiquadric approximation for partial differential equations. Engineering Analysis with boundary elements, 18(1), 9-17.

[6] Hoffer, E., & Ailon, N. (2015, October). Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition* (pp. 84-92). Springer, Cham.

[7] triplet-network,https://github.com/andreasveit/triplet-network-pytorch