Utilization of Neural Embeddings to Predict Population Behavior During the COVID-19 Pandemic

Gihan Jayatilaka, Louiqa Raschid, Vanessa Frias-Martinez
University of Maryland
2024-July-10

This work is part of the PandEval project. All correspondence should go to vfrias AT umd DOT edu

Embeddings

- Machine learning models predicts and **output ($\hat{y}$)** for a given **input (x)**
  - $\hat{y} = f(x)$
- x can be images, text, numerical data, documents or ANYTHING.
- **Embeddings** (z) are an intermediate representations of inputs.
  - $z = g(x)$
  - $\hat{y} = h(x)$
- Generally, embeddings are “useful” to the prediction task at hand.

What is good about embeddings?

1. Embeddings are relatively lower dimensional than inputs. They are a compact representation.
2. Embeddings are vectors of floats. They are easier to handle compared to different forms of input data (text, image, tabular)

Ideally,

1. Embedding function **g()** would be handling noise, outliers, etc:.
2. **g()** would get rid of sparsity and redundancy of the data.
3. **g()** would highlight the relevant information for the task at hand and discard irrelevant information.

Learning **g( )**

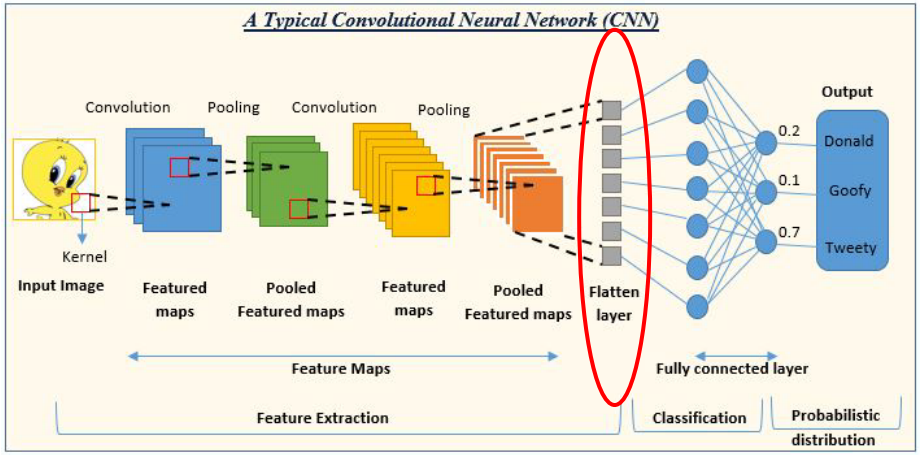
- Handcrafted
  - Explicit equations based on human intuition on what is important.
- Supervised learning
  - z is explicitly chosen to be meaningful for the task at hand.
- Unsupervised learning
  - z is chosen to enhance the patterns in data distribution.
  - Usually, the “patterns” are extracted based on priors about what is important.

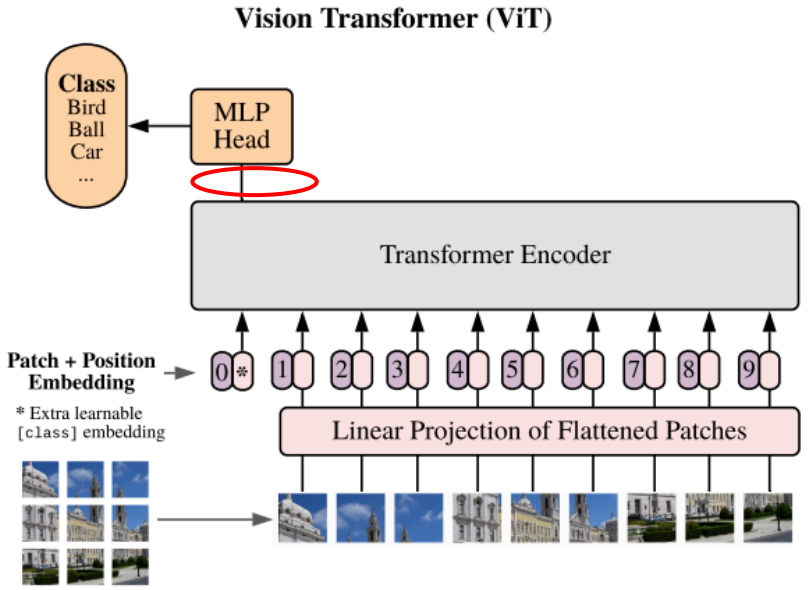
Handcrafted Embeddings

Variable	CDC-SVI[a]	CCVI[b]	PVI[c]
Percentage of individuals living below the federal poverty level	x	x	x
Percentage of civilians aged ≥16 y who are unemployed	x	x	x
Per-capita income	x	x	x
Percentage of individuals aged ≥25 y with no high school diploma	x	x	x
Percentage of individuals aged ≥65 y	x	x	x
Percentage of individuals aged ≤17 y	x	x	
Percentage of the civilian noninstitutionalized population with a disability	x	x	
Percentage of single-parent households with children aged <18 y	x	x	
Percentage of individuals who belong to racial and ethnic minority groups (everyone except non-Hispanic White people)	x	x	
Percentage of individuals aged ≥5 y who speak English "less than well"	x	x	
Percentage of housing in structures with ≥10 units	x	x	x
Percentage of mobile homes	x	x	x
Percentage of crowded households (households with more people than rooms)	x	x	x
Percentage of households with no vehicle available	x	x	x
Percentage of individuals living in group quarters	x	x	x
Percentage of the population that has no health insurance		x	x
Percentage of households without access to indoor plumbing		x	
Annual cancer incidence per 100 000 population		x	
Number of individuals living with HIV per 100 000 population		x	
Percentage of adults diagnosed with high cholesterol		x	
Percentage of adults diagnosed with a stroke		x	
Percentage of adults diagnosed with heart disease		x	
Percentage of adults diagnosed with chronic obstructive pulmonary disease, emphysema, or chronic bronchitis		x	
Percentage of adults reporting being obese		x	x
Percentage of adults ever diagnosed with diabetes		x	x
Percentage of adults who report smoking cigarettes		x	x
Intensive care unit beds per 100 000 population		x	
Hospital beds per 100 000 population		x	x
Epidemiologists per 100 000 population		x	
Agency for Healthcare Research and Quality Prevention Quality Indicator overall composite score		x	
State-level health spending per capita		x	
Aggregate cost of medical care		x	
Percentage of population with a primary care physician		x	
Public Health Emergency Preparedness cooperative agreement total funding per capita		x	
Health laboratories per 100 000 population		x	
Emergency services per 100 000 population		x	
Long-term care facility residents per 100 000 population		x	
Prison population per 100 000 population		x	
Percentage of the population employed in an industry in which frequent contact with other people in the workplace occurs		x	
Population density or estimated daytime population		x	x
Number of transmissible cases of COVID-19			x
COVID-19 disease spread			x
Average traffic volume per meter of major roadway in the county			x
Numerical score based on social distancing scoreboard			x
Population divided by the number of COVID-19 tests performed			x
Percentage of population that self-identifies as Black or African American			x
Percentage of population that self-identifies as American Indian or Alaska Native			x
Average daily density of fine particulate matter in μg per cubic m			x
Years of potential life lost before age 75 y per 100 000 population (age-adjusted)			x

Abbreviations: CCVI, COVID-19 Community Vulnerability Index; CDC-SVI, Centers for Disease Control and Prevention Social Vulnerability Index; PVI, Pandemic Vulnerability Index.

Supervised learning for Embeddings



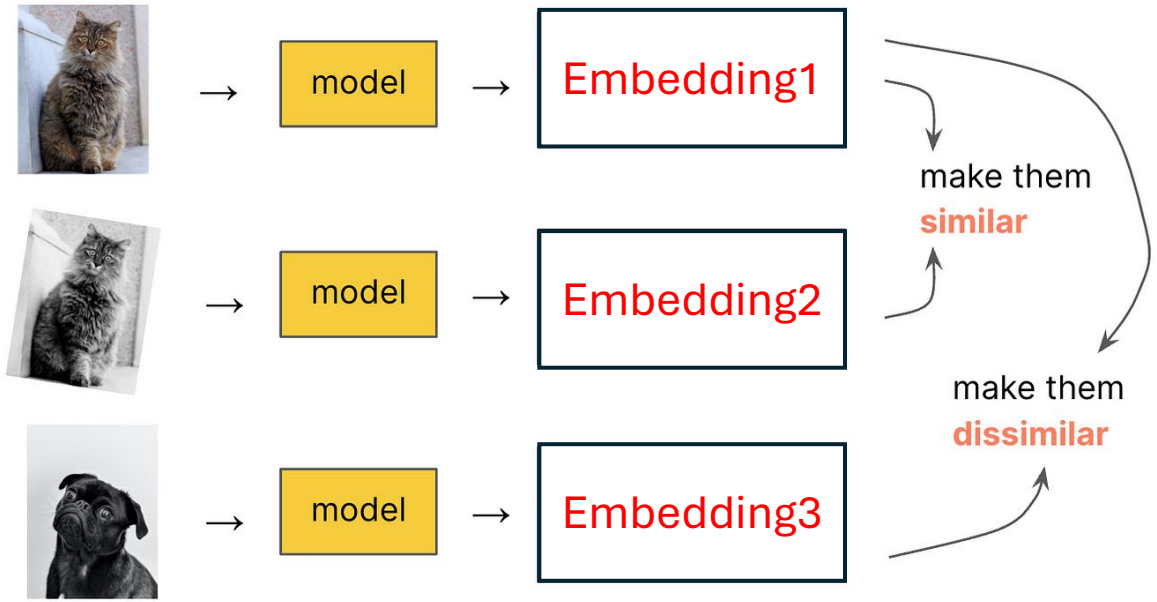


- The embeddings are marked by red ovals.

Unsupervised Embeddings

- Unsupervised learning broadly refers to learning in the absence of labels.
- Self Supervised Learning is a new paradigm in computer vision, which creates “pseudo labels” from within unlabeled data.
- For this presentation, we will lump both techniques together.

Self Supervised Learning



- The embeddings are learnt such that "similar" images are closer and "dissimilar" images are further in embedding space.
- "Similar images" are generated by assuming invariant properties of images – (color pallet, rotation, lighting).

Self Supervised Learning (Pretext tasks)

Supervised embeddings	Unsupervised embeddings
Low dimensional, compact vectors of floats.	
Noise, outliers, sparsity and redundancy are handled.	
Retains information that is relevant to task at hand.	Retains information that based on prior understanding of what is meaningful in general.
Discards information that are not useful for the particular task.	Doesn't
	Generalizes better.
Annotated data is required. Difficult to obtain in larger scales.	**Can use data in the wild.** **Easy to obtain.**
	Prone to noise and dataset biases.
	Requires another annotated dataset/task to measure the quality.

Example work (Hui et. al. 2020)

- We will look into the following paper as a case study to explain all the techniques used in the domain.

  Bo Hui, Da Yan, Wei-Shinn Ku, and Wenlu Wang. 2020. **Predicting Economic Growth by Region Embedding: A Multigraph Convolutional Network Approach.** *In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland. ACM, New York, NY, USA,* []

- Task :
  - Predict the economic growth for a given zip code based on information about the demographic, social, economic and housing information of the zip code and it's connectivity to other zip codes.
  - Predict the change in the number of NAICS 2 digit industry establishment counts for a given zip code based on ACS data of the zip code and it sharing school district, county and direct flights to other zip codes.
- Note:
  - This work does not handle temporal aspects explicitly!

(Hui et. al. 2020)

ACS data

- $\mathbf{x}_{i,c_1}$ for Category c_1 with 82 demographic features;
- $\mathbf{x}_{i,c_2}$ for Category c_2 with 150 social features;
- $\mathbf{x}_{i,c_3}$ for Category c_3 with 114 economic features;
- $\mathbf{x}_{i,c_4}$ for Category c_4 with 142 housing features,

Connectivity

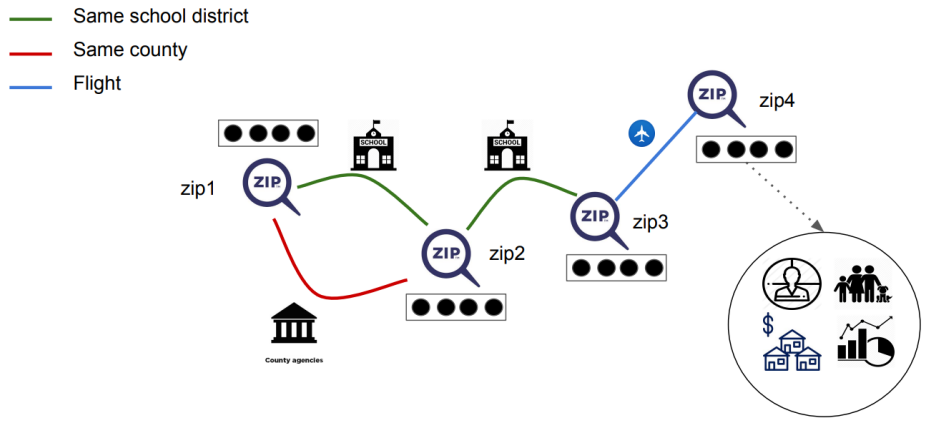


Figure 4: An illustration of the Multigraph

(Hui et. al. 2020) – overall architecture

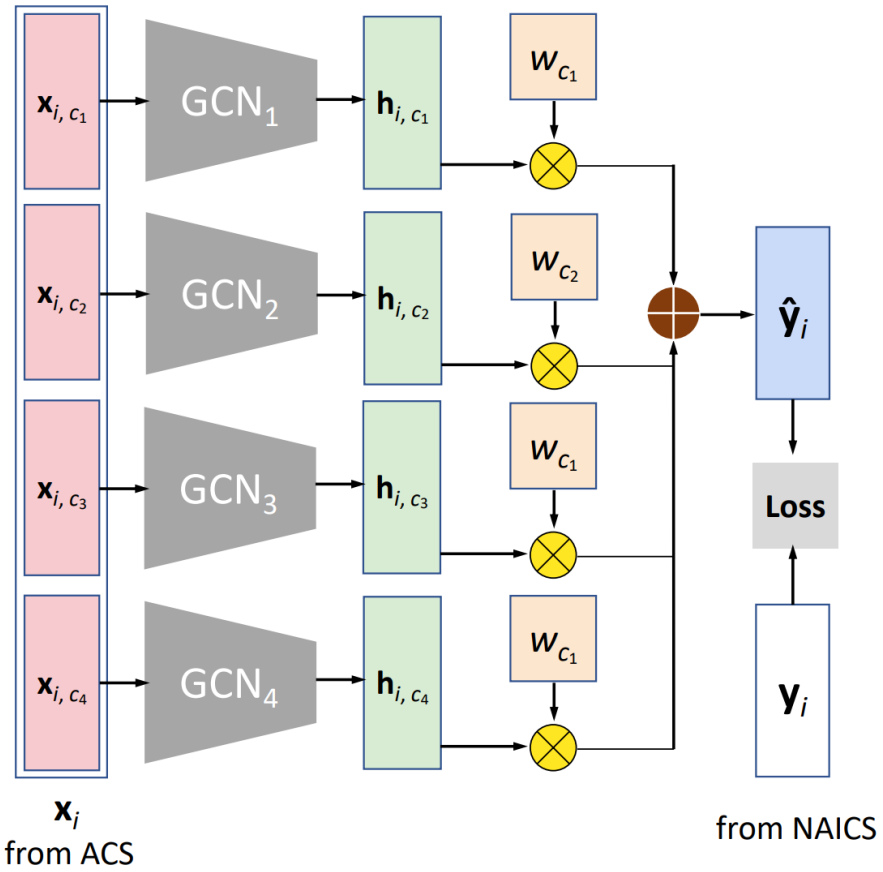
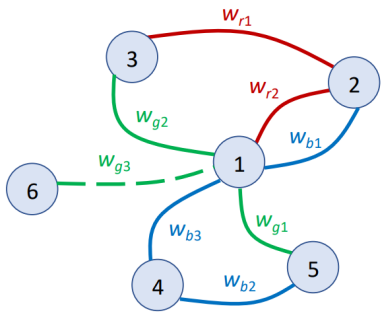


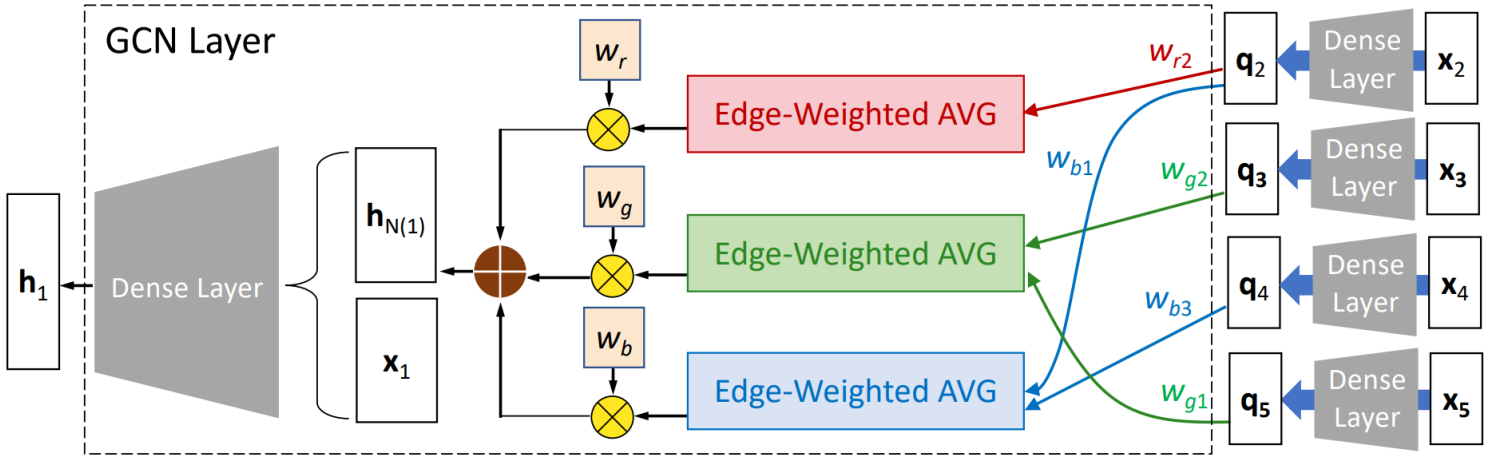
Figure 5: Model Overview

- $h_{i,c_j} \in \mathbb{R}^{20}$
- $\hat{y}_i \leftarrow w_{c_1} \cdot \mathbf{h}_{i,c_1} + w_{c_2} \cdot \mathbf{h}_{i,c_2} + w_{c_3} \cdot \mathbf{h}_{i,c_3} + w_{c_4} \cdot \mathbf{h}_{i,c_4}$,
- GCN_i and w_{c_j} are trainable
- Loss is MSE
- Note that individual GCN_i directly predicts NAICS change.

(Hui et. al. 2020) – Connectivity



(a) A Toy Multigraph



(b) Our GCN Layer Design

Figure 6: GCN Design

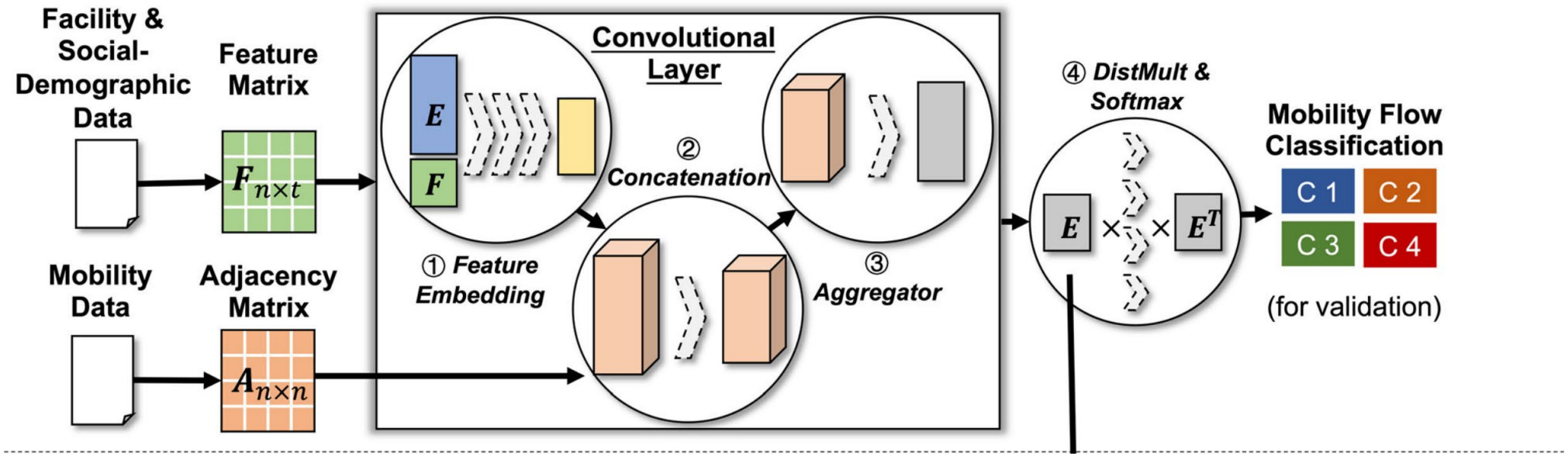
- Note:
  - x_i in this slide is for a particular category of ACS features.
  - w_g, w_g, w_b are trainable

Alternative ideas for connectivity

- Supervised
  - Graph Convolution with learnable weights (Hui et. al.)
  - Graph Convolution with naïve weights.
  - 2D convolution based ideas.

- Unsupervised
  - Random walks
  - node2vec

Graph Convolution with naïve weights.



- **n** is the number of rectangular grids (not counties or census tracts)
- **A** is a binary matrix.
- Aggregator is summing for one's in **A**
- The white arrow head is an MLP.

Fan, C., Yang, Y. and Mostafavi, A., 2024. **Neural embeddings of urban big data reveal spatial structures in cities**. Humanities and Social Sciences Communications, 11(1), pp.1-15.

Table 1 Selection of features of urban areas for a county.

Category	Feature details			Source
Socio-demographics	Percentage of minority	Per capita income		U.S. Census
	Percentage of people older than 65	Percentage of crowded structures		
Facility services	Utilities	Construction	Manufacturing	Points of Interest (POI)
	Wholesale trade	Retail trade	Finance and insurance	
	Professional, scientific, and technical services	Transportation and warehousing	Real estate rental and leasing	
	Information	Administrative services	Educational services	
	Health care and social assistance	Arts, entertainment, and recreation	Accommodation and food services	
	Other services (except public administration)	Public administration	-	
Population	Residential population size			Mobile phone data
Network characteristics	In-degree			
	Weighted in-degree			

2D CNN based Ideas

FIGURE 5: Set of neighbor coordinates δ.

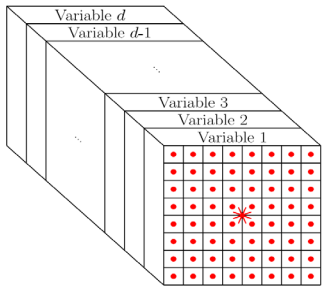


FIGURE 6: Data square cuboid γ_{δ}.

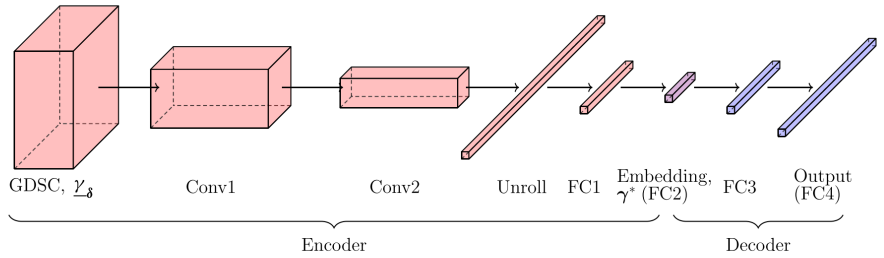


FIGURE 7: Convolution-based geographic embedding model.

- The adjacency is only based on the geographical locality.
- Unable to model heterogenous affinity parameters.

C Blier-Wong, H Cossette, L Lamontagne, E Marceau. **Geographic ratemaking with spatial embeddings [AAAI 2020]**

Node2vec

$$\max_f \sum_{u \in V} \log Pr(N_S(u)|f(u)).$$

$$Pr(N_S(u)|f(u)) = \prod_{n_i \in N_S(u)} Pr(n_i|f(u)).$$

$$Pr(n_i|f(u)) = \frac{\exp(f(n_i) \cdot f(u))}{\sum_{v \in V} \exp(f(v) \cdot f(u))}.$$

Identifying Personas

- Xiao Qian, Utkarsh Gangwal, Shangjia Dong, Rachel Davidson. **“A Deep Generative Framework for Joint Households and Individuals Population Synthesis”**
- Objective:
  - **Input:** ACS microdata (household or individual level) and ACS population data.
  - **Output:** A population of synthetic personas.
  - **Constraint:** The synthetic population should be realistics. i.e. the marginal distribution of the synthetic population should match state and national level distribution for different measures.

Qian et. al.,

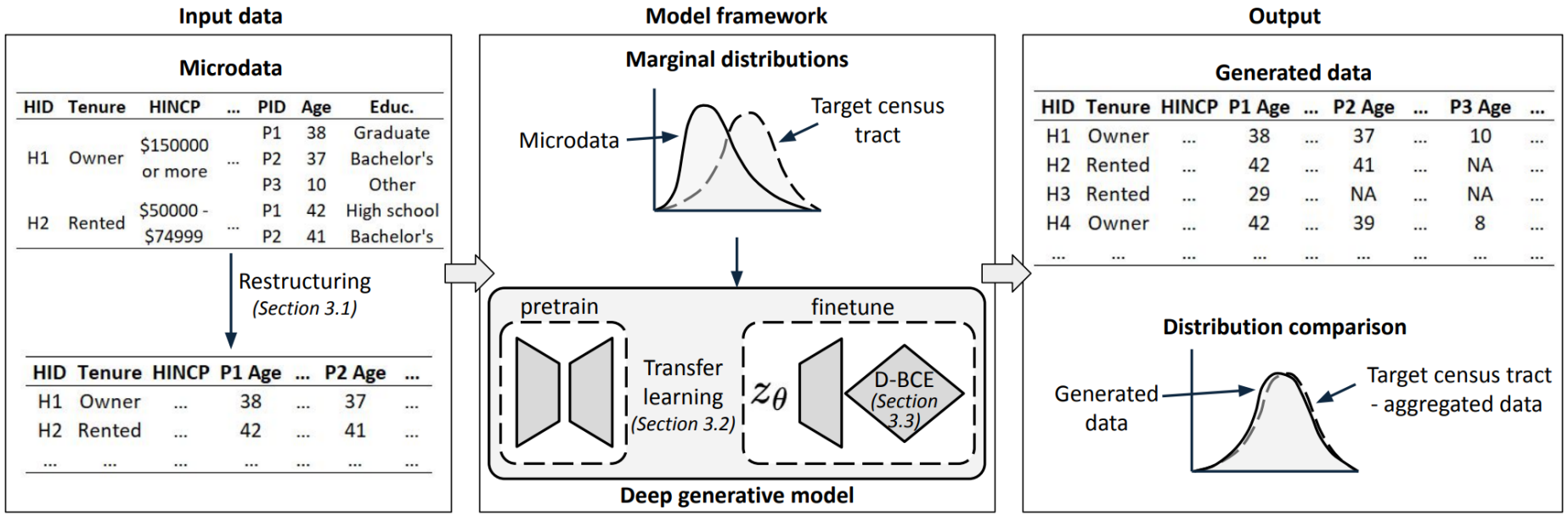
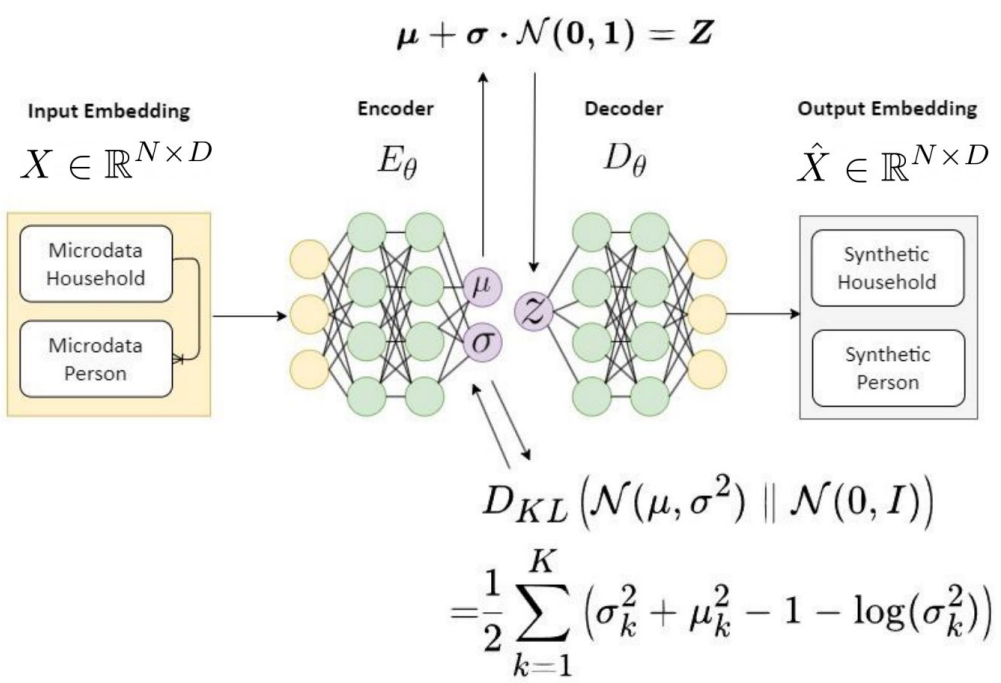
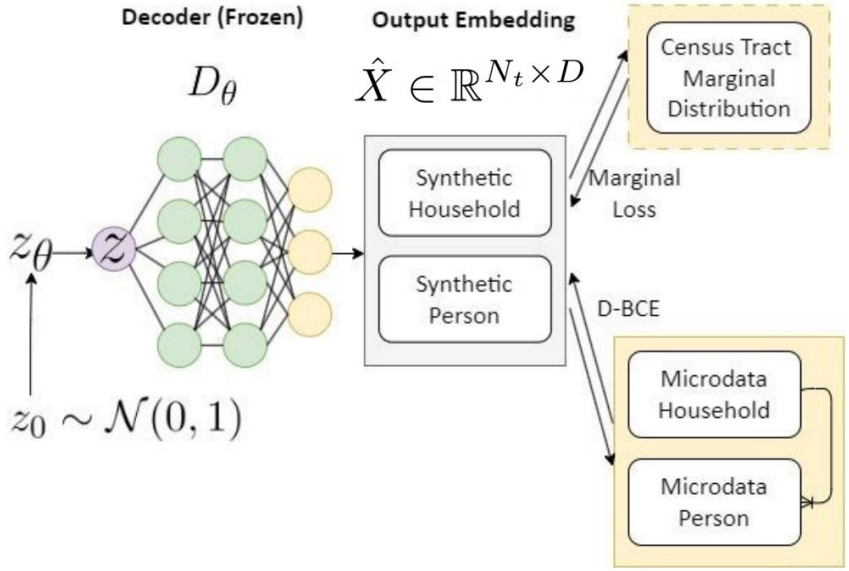


Figure 2: The end-to-end deep generative pipeline for synthetic household-individual inventory development

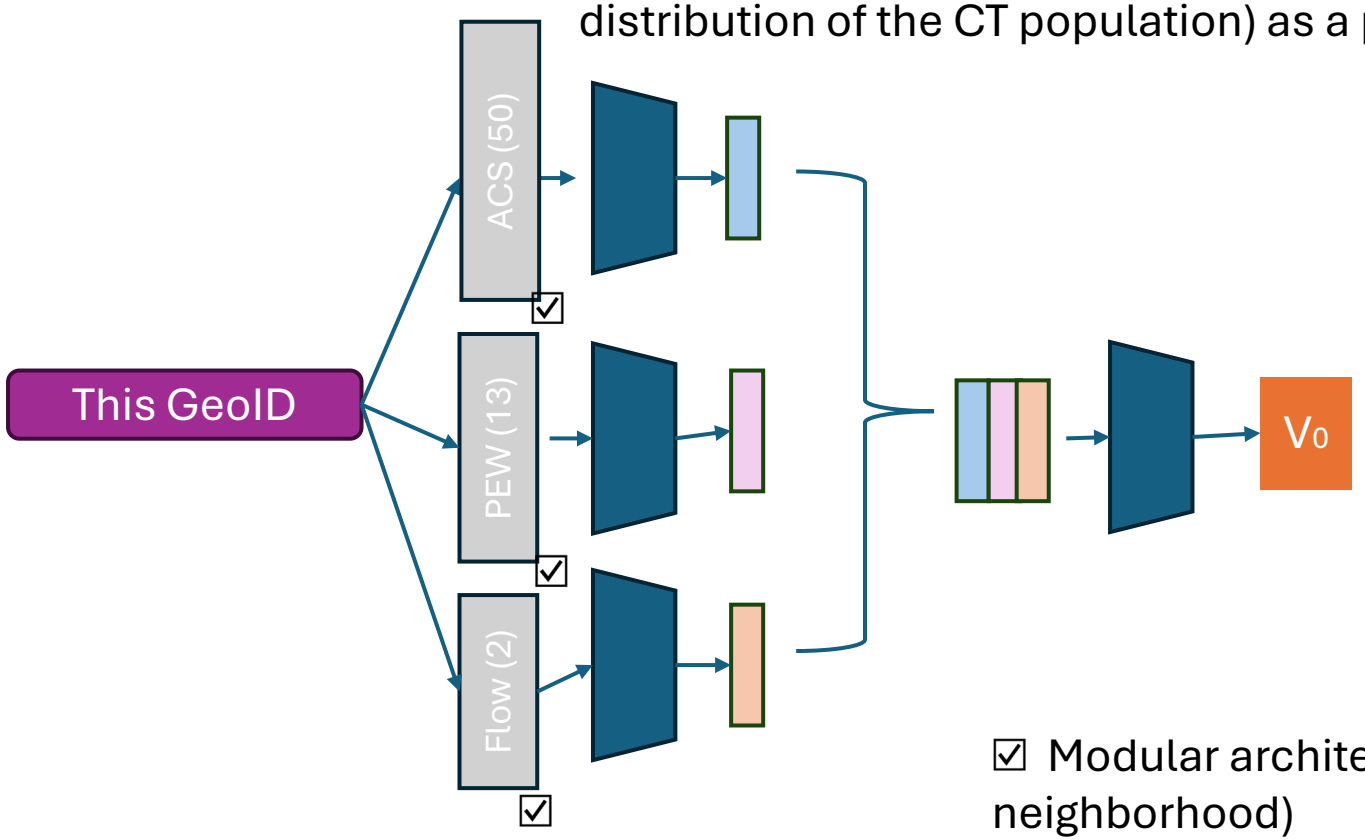
Qian et. al. -- Training





Ongoing Implementation(s)

We are using demographic data (gender, age, race distribution of the CT population) as a placeholder.



☑ Modular architecture (which can apply on a neighborhood)

Picking Neighbors

☑ Based on Geographical Distance

☑ Based on Flow

Output Variable

☑ Ratio drop of Total Flow

To do:

☐ Debug the Dimensions

☐ End to end training