Editing of Pig DNA May Lead to More Organs for People

Youngduck Choi, Amy Jung, Vaughn Tajirian, Katie Westerlund

CILVR Lab, New York University

December 6, 2015

Motivation

- In recent years the field of machine learning has had a major break-through with a family of models called deep neural networks.
- The applicability of deep learning, however, is not yet comprehensive.
- Can we get a state-of-the-art result for a well defined supervised task in medical domain using deep learning?

Problem: Early Detection of Diabetes from Claims

1. An STR locus on chromosome 6 has many alleles, each with a different number of CA repeats. In a particular population, the alleles have these frequencies:

allele	allele frequency
20	0.10
21	0.001
22	0.12
23	0.05
24	0.30
25	0.10
26	0.009
27	0.09
29	0.23

These frequencies describe the local gene pool for the STR locus. Assume that the population is at Hardy-Weinberg equilibrium.
a. What's the expected frequency of 25/25 homozygotes?

b. What's genotype is expected to be the most common genotype in this population?

c. What's the expected heterozygosity at this locus? (You'll want a calculator for this one.)

Figure 1 : Various supervised tasks

Neighbors I

Top 50 cosine distance codes:
4394155 (religious affiliation) : 1.0
4319594 (sufficiency of income for needs) : 0.886792323995
4320469 (low motivation) : 0.875451444993
4419678 (auditory and visual hallucinations) : 0.861213899833
4276205 (inability to cope) : 0.860558423511
4864857 (poor cognition) : 0.860319009683
4548639 (morphine 20 mg) : 0.858606201488
4122208 (swearing) : 0.858280502692
4247373 (fidgeting) : 0.855690125451
4121999 (labile affect) : 0.855292082459
4311835 (incoherent) : 0.854193874673
4528103 (morphine 20 mg/ml oral solution) : 0.850018284495
4120301 (moderate anxiety) : 0.849993724872

Neighbors II

Top 50 cosine distance codes:
4000978 (alzheimer's disease) : 1.0
4305675 (aricept) : 0.927681307567
4749486 (alzheimer's disease pathway kegg) : 0.90958817619
4305676 (donepezil) : 0.906635582954
4656147 (namenda) : 0.903731091715
4298078 (dementia) : 0.884320489297
4012831 (memantine) : 0.883449173077
4005600 (presenile dementia) : 0.864689763427
4605035 (mild cognitive disorder) : 0.859568216842
4183905 (infarction, lacunar) : 0.847736673577
4464452 (demented) : 0.836936999719

Defining the Surrogate Measures for Embedding Space Properties

- NDR-RT relation rank statistics score : let (x, y) be a pair that encodes certain medical relationship z(i.e. may treat). We say that if the embedded structure has the entity y as in the neighborhood of x where the neighborhood of x is defined as the top-k entities that are sorted by the inner product score with respect to x, then it exhibits "relatedness" with respect to z.

Defining the Surrogate Measures for Embedding Space Properties

- UMLS semantic type rank statistics score: let (x,t) be a pair of medical entity x with type t. We say that if the embedded structure has an entity y of same type t in the neighborhood of x where the neighborhood of x is defined as the top-k entities that are sorted by the inner product score with respect to x, then it exhibits 'conceptual similarity."

UMLS Semantic Type Scoring System

4003436 (carcinoma, non-small-cell lung, C0007131, ['Neoplastic Process']) : 1.0
4069419 (small cell carcinoma of lung, C0149925, ['Neoplastic Process']) : 0.955599426233
4394316 (carcinoma of lung, C0684249, ['Neoplastic Process']) : 0.933888909808
4125384 (malignant neoplasm of lung, C0242379, ['Neoplastic Process']) : 0.928970432186
4070138 (adenocarcinoma of lung (disorder), C0152013, ['Neoplastic Process']) : 0.924754262378
4555365 (tarceva, C1135136, ['Organic Chemical', 'Pharmacologic Substance']) : 0.917757636841
4069342 (lung mass, C0149726, ['Finding']) : 0.914073299934

Empirical Result I

Table 1 : Neighbor and Analogy Results on the May Treat Relationship. Total 722 drugs were queried.

	NEIGHBORS	ANALOGY MEAN	ANALOGY MAX
MEDCOPORA	9.8338%	21.7299%	41.4127%
10BIL,30D,S500,NS20	27.7008%	20.2710%	43.9058%
10BIL,30D,S300,NS20	28.5319%	20.4457%	42.6593%
10BIL,30D,S400,NS50	**30.7479%**	25.6753%	45.7064%
1BIL,1D,S300,NS20	30.6094%	25.4884%	**47.6454%**
1BIL,7D,S300,NS20	30.3324%	24.9829%	46.2604%
SVD,S100,NS10	31.4404%	18.7605%	31.7175%
SVD,S300,NS10	35.4571%	22.8062%	41.1357%
SVD,S500,NS10	39.1967%	24.6856%	42.2438%
SVD,S1000,NS10	42.9363%	26.8280%	43.6288%

Empirical result II

Table 2 : The mean DCG result for the top 40 neighbors of CUIs associated with certain type.

	MedCopora	GraphSGD (Time)
Pharmacologic Substance	**6.74 ± 3.21**	2.95 ± 2.15
Disease or Syndrome	**5.41 ± 2.48**	4.28 ± 1.60
Neoplastic Process	**6.74 ± 3.47**	4.54 ± 0.11
Clinical Drug	**1.01 ± 0.12**	0.12 ± 0.18
Finding	**2.85 ± 1.90**	2.15 ± 1.35
Injury or Poisoning	2.67 ± 2.40	**2.92 ± 2.80**

Conclusion

- Under the introduced formal definitions, we empirically show that the embedded structure (while having the algorithmic nature analogous; slight tweak) exhibits relatedness when learned on distributional pattern in time, whereas the embedded structure exhibits conceptual similarity when learned learned on distributional pattern in corpora even in the lower dimensional space.

Future work

- More work needs to be done incorporating these representations for the specific given medical task. The best model using these representations (not hand crafted) does not over perform the internal baseline of gradient boosted decision tree models using large set of hand-engineered features.