深度学习实验

工具：

- 语言：python3.5
- 工具包：KERAS（基于 tenserflow 版，是一个深度学习包），CUDA（用来 gpu 加速，可大幅度缩短深度网络训练时间）。
- IDE:pycharm

数据集：

- 该数据集共包含记录 47269982 条，全部为住院数据。
- 住院数据包含人员 92652 人，住院人次 181541，数据时间跨度从 2006-12-15 到 2015-05-27。

实验目的：

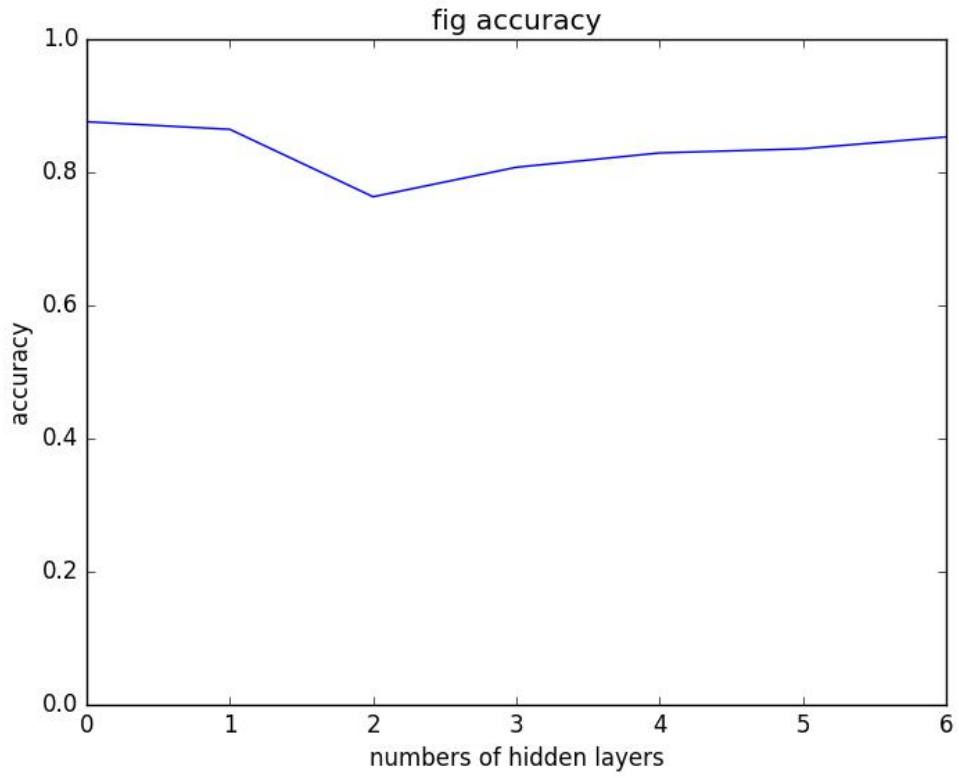
- 我想针对一种疾病，现在是以脑梗塞为例，预测一个病人是否会患上脑梗塞，而忽略掉再住院的情况，即只考虑新患脑梗塞的情况。
- 若该试验结果较好，我们就可以应用到疾病的预防中去，来减少脑梗塞的发病率。

流程与实验结果

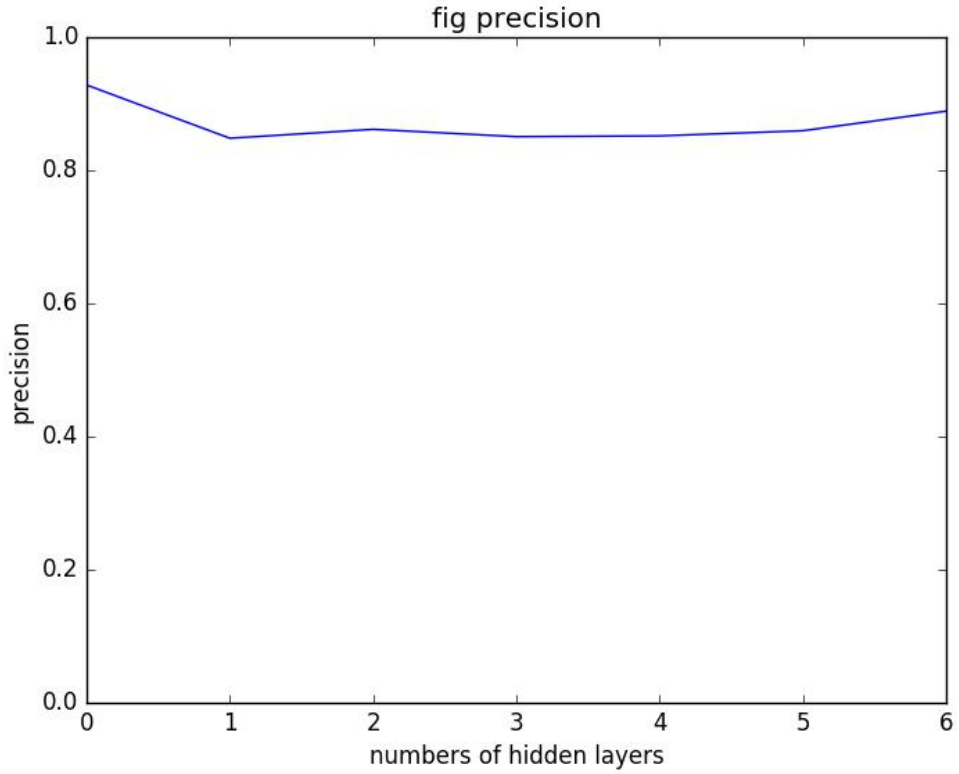
- 数据预处理部分
  a) 首先我找到 2014 年以后新患脑梗塞的人 3877 人。
  b) 为了避免非平衡问题带来的影响，我们从剩余数据集中找到了未新患脑梗塞的人员 3750 人。
  c) 将这两部分人合并，得到数据集人员 7627 人。
  d) 我们使用每个人的医疗项目费用做预测，医疗项目为 5764 个。
  e) 最终深度学习网络的输入数据为 7877*5764 项，其中非零项为 840095，占总数据项的 1.85%。部分数据如图所示

	100012	11702	11705	200001	200002	200003	200004	200005	200006	200007	200008	200009	200010	200011	200012	200013
zy1	0	0	0	0	90	0	0	0	47	0	0	0	80	0	18	0
zy10007	0	0	0	0	14	0	189	14	0	0	0	0	18	14	0	0
zy100110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy100183	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy100239	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy100240	0	0	0	0	0	0	108	0	0	0	0	0	0	0	0	0
zy10026	0	0	0	0	38	0	60	0	0	0	0	1	40	0	0	0
zy100271	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy100524	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0
zy100569	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy100834	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy101294	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy101350	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy101646	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy102007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
zy102359	0	0	0	0	30	0	25.5	0	0	0	0	0	17	20	0	0
zy102390	0	0	0	0	69	0	0	270.3	0	0	0	0	121.4	19	7	0
zy102448	0	0	0	0	28.6	0	0	321.9	0	0	0	0	59.4	0	27	0
zy10269	0	0	0	0	112	0	0	64	36	50	0	7	68	82	1	1
zy102699	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zy102759	0	0	0	0	0	0	45	45.6	0	0	0	0	0	0	8	0
zy10284	0	0	0	0	0	0	0	46	0	0	0	0	46	0	0	0
zy103084	0	0	0	0	0	0	0	17.1	0	15.2	0	3.8	5.8	0	1	0
zy103206	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

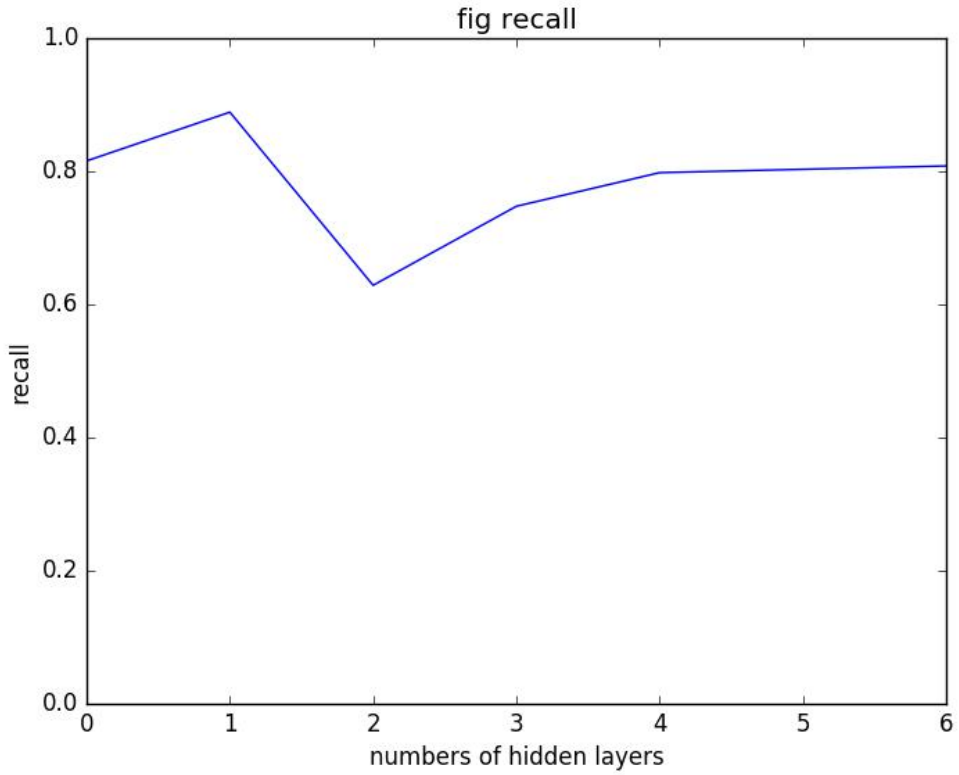
- 实验方法
  a) 我们使用层叠自动编码器对深度网络进行 pre-training,再使用 mlp（多层感知机）进行 fine_tuning，希望得到较低维度的 feature representation，可以作为高维数据的表征。
  b) 我们使用了 PCA,ICA 等降维方法与之进行了比较。
- 实验结果
  a) 深度网络层数与 accuracy 的关系，0 层代表未使用深度网络。



  b) 深度网络层数与 precison 的关系，0 层代表未使用深度网络。



c) 深度网络层数与 recall 的关系，0 层代表未使用深度网络。



d) 层叠自动编码器与降维方法的比较（6 层隐藏层）

方法	Accuracy	Presicion	Recall
层叠自动编	0.8531645569620253	0.8888888888888888	0.8080808080808081

码器+深度网络			
PCA	0.7746835443037975	0.8562091503267973	0.6616161616161617
ICA	0.7772151898734178	0.8293413173652695	0.6994949494949495
Xgboost	0.8759493670886076	0.9281609195402298	0.8156565656565656

- 实验结论
  a) 使用深度学习对数据进行降维，可以保证对于预测问题有相近的精确度，准确度和召回率。
  b) 由于数据维度大大降低，减少了分类问题的处理时间
  c) 深度学习的方法与其他降维方法相比跟能保留数据的 feature representation。