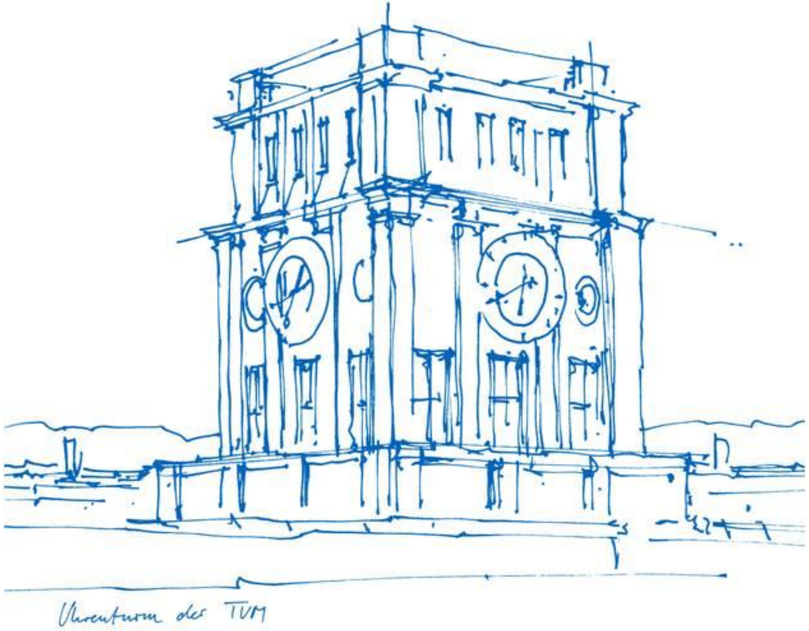
Multi-Lingual Theme Prediction of Customer Reviews Using Deep Pre-Trained Embeddings

Team 06

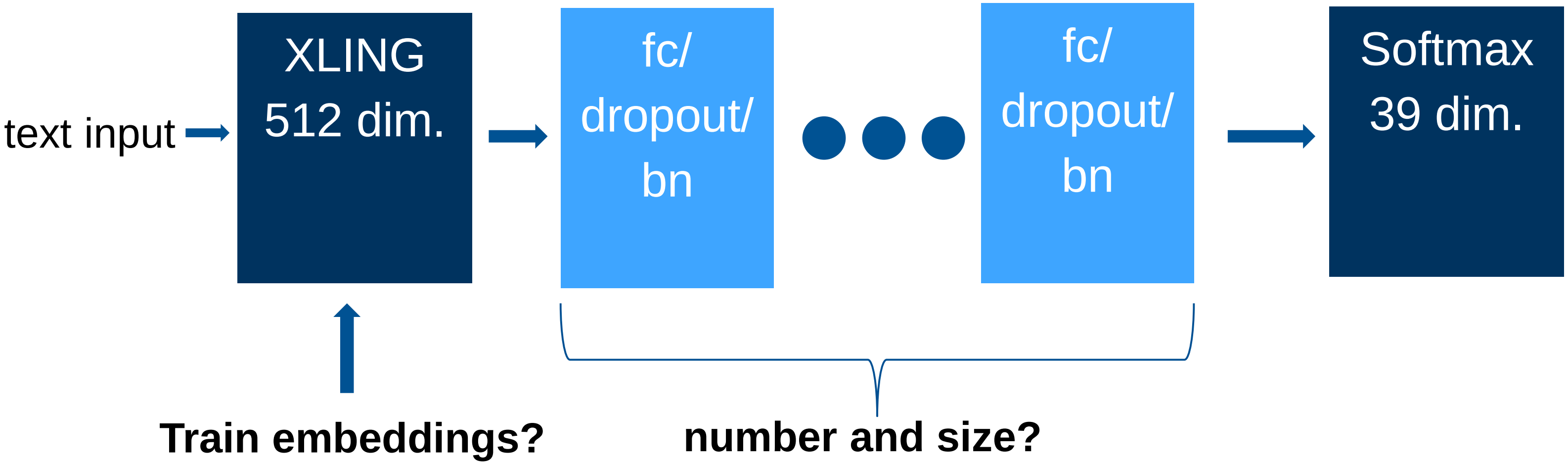
Michael Sorg

19.06.2019



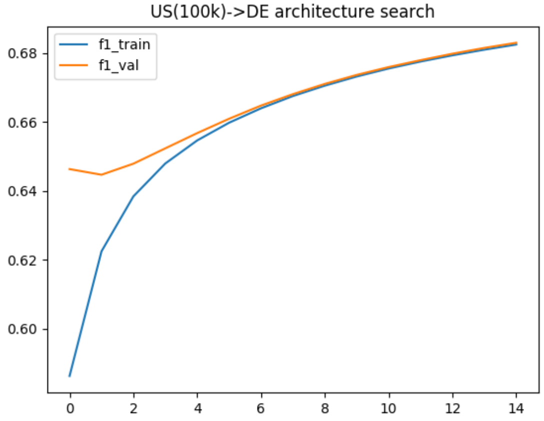
Network search

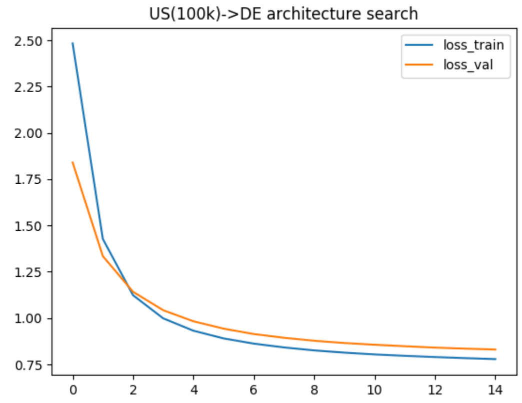
- Task: train on english data only – test on German data

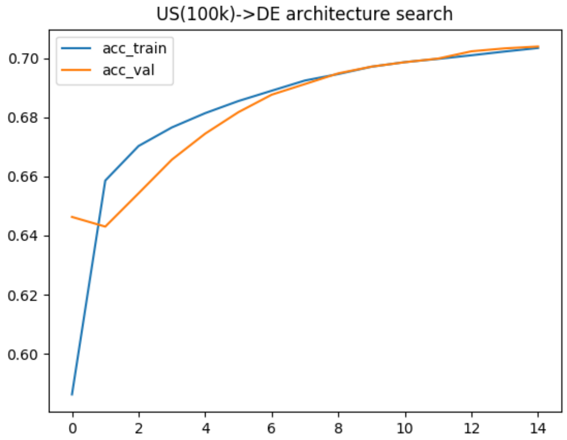


Baseline experiments

- One hidden layer (only xling + softmax)


US(100k)->DE architecture search

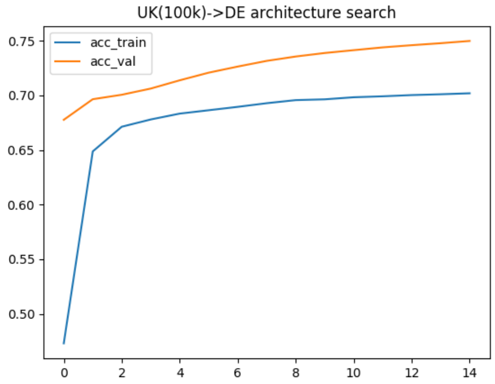

US(100k)->DE architecture search

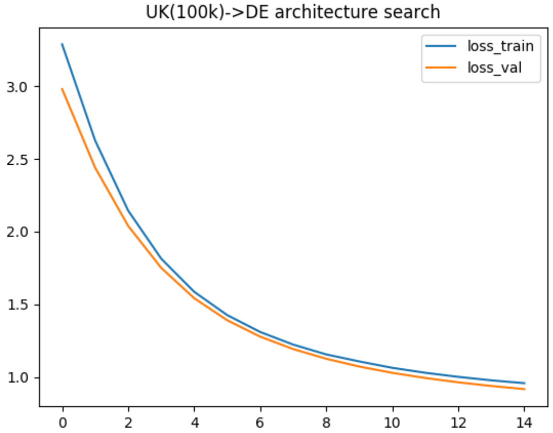

US(100k)->DE architecture search

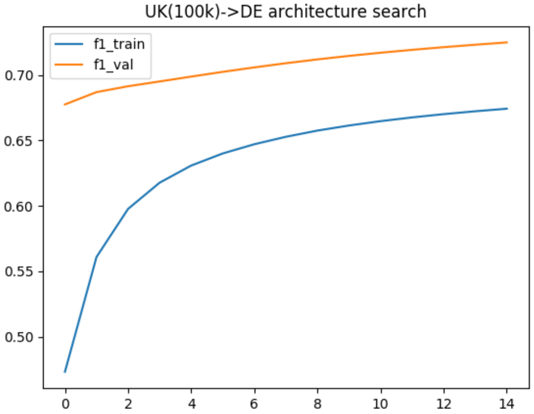
Train data	US (100k)
Test data	DE (100k)

Baseline experiments

- One hidden layer (only xling + softmax)



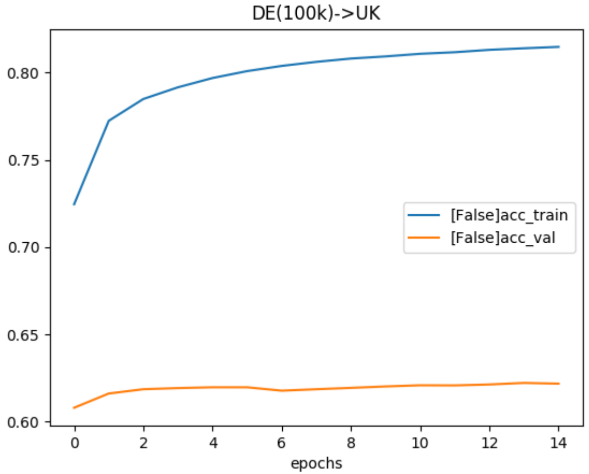


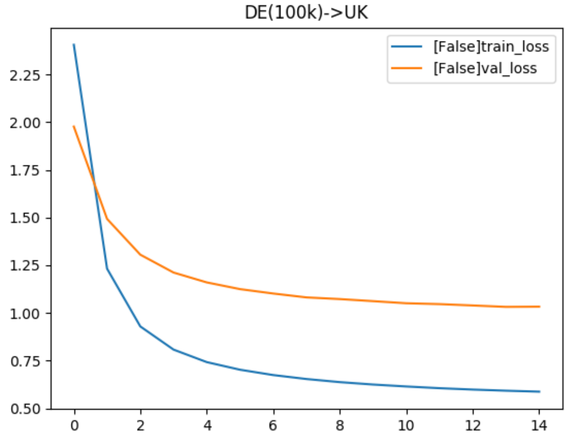


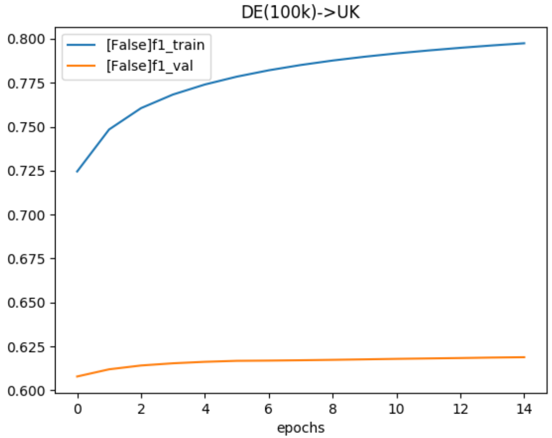
Train data	**UK** (100k)
Test data	DE (100k)

Baseline experiments

- One hidden layer (only xling + softmax)



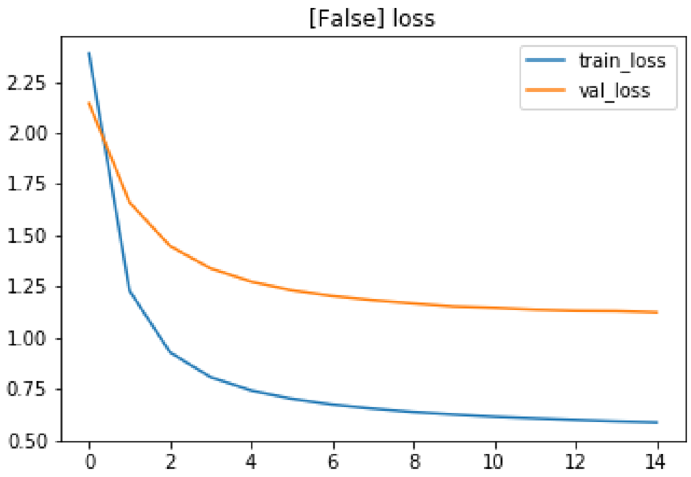


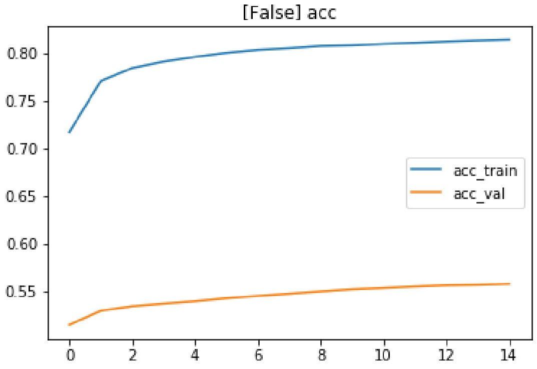


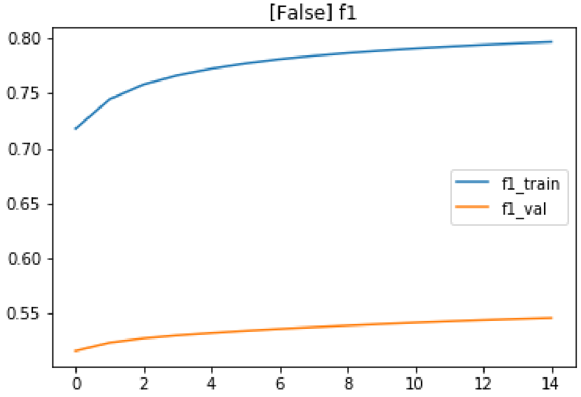
Train data	**DE** (100k)
Test data	UK (100k)

Baseline experiments

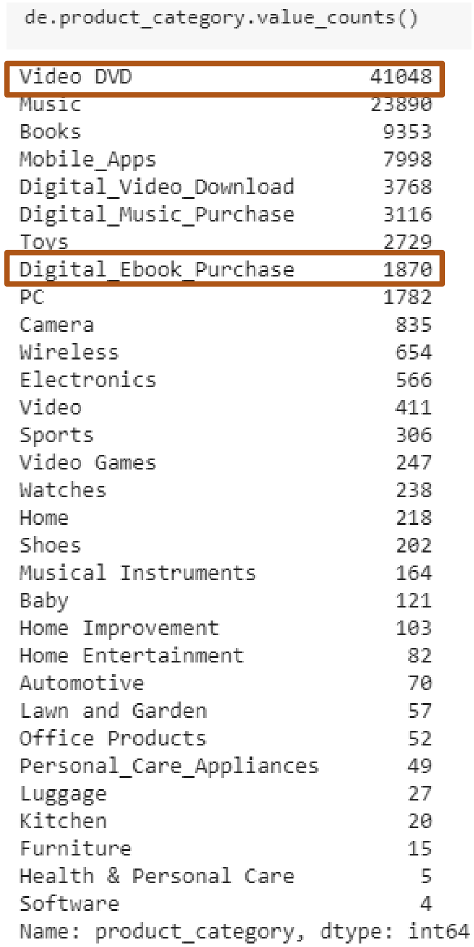
- One hidden layer (only xling + softmax)





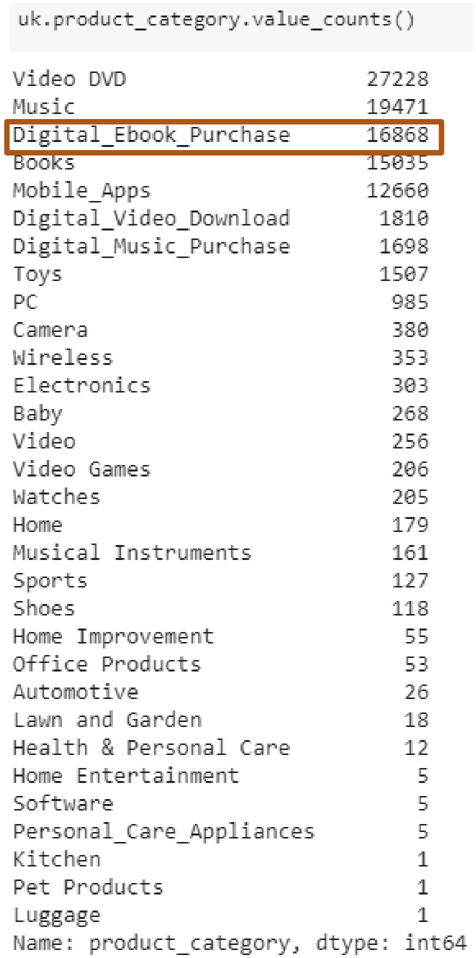


Train data	**DE** (100k)
Test data	US (100k)



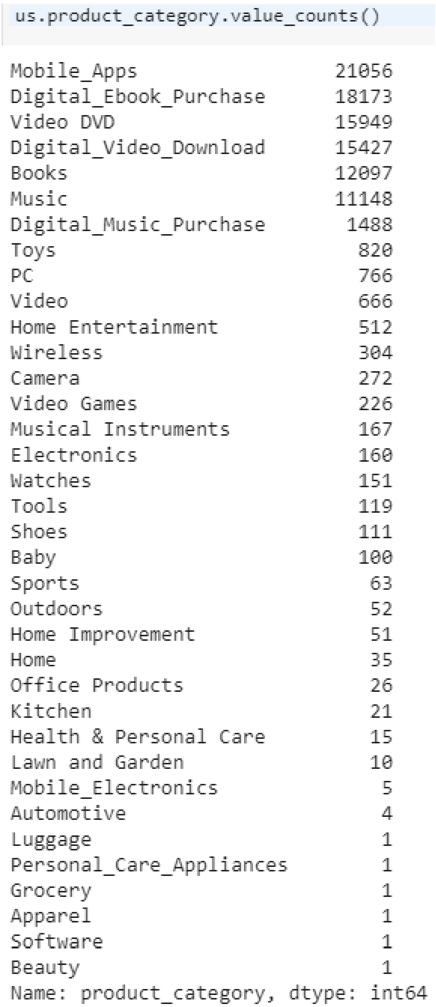
```
de.product_category.value_counts()
```

```
Video DVD                  41048
Music                      23890
Books                       9353
Mobile_Apps                 7998
Digital_Video_Download      3768
Digital_Music_Purchase      3116
Toys                        2729
Digital_Ebook_Purchase      1870
PC                          1782
Camera                       835
Wireless                     654
Electronics                  566
Video                        411
Sports                       306
Video Games                  247
Watches                      238
Home                         218
Shoes                        202
Musical Instruments          164
Baby                         121
Home Improvement             103
Home Entertainment            82
Automotive                    70
Lawn and Garden               57
Office Products               52
Personal_Care_Appliances      49
Luggage                       27
Kitchen                       20
Furniture                     15
Health & Personal Care         5
Software                       4
Name: product_category, dtype: int64
```



```
uk.product_category.value_counts()
```

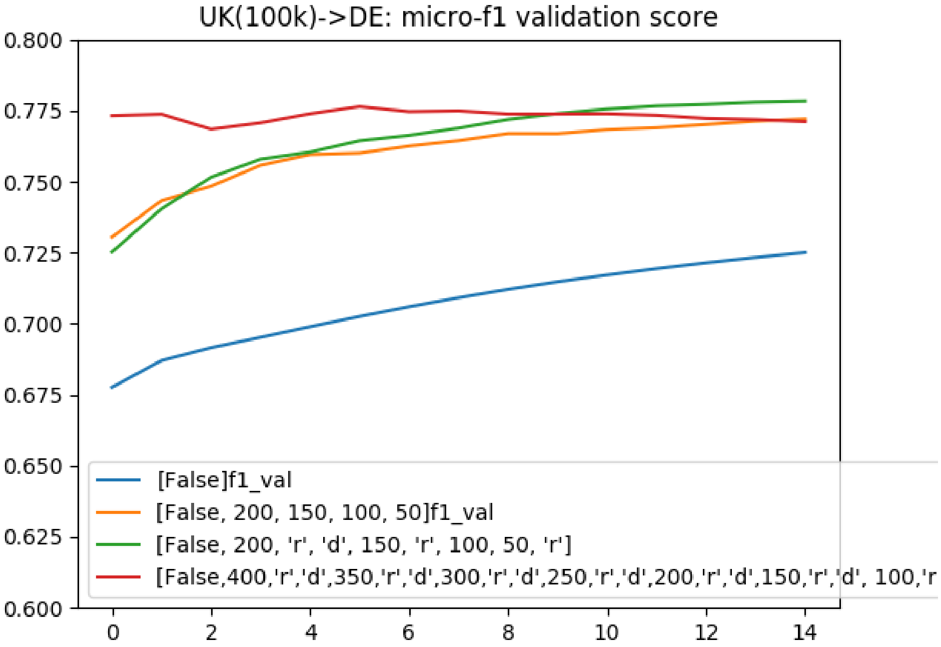
```
Video DVD                   27228
Music                       19471
Digital_Ebook_Purchase      16868
Books                       15035
Mobile_Apps                 12660
Digital_Video_Download       1810
Digital_Music_Purchase       1698
Toys                         1507
PC                            985
Camera                        380
Wireless                      353
Electronics                   303
Baby                          268
Video                         256
Video Games                   206
Watches                       205
Home                          179
Musical Instruments           161
Sports                        127
Shoes                         118
Home Improvement               55
Office Products                53
Automotive                     26
Lawn and Garden                18
Health & Personal Care         12
Home Entertainment              5
Software                        5
Personal_Care_Appliances        5
Kitchen                         1
Pet Products                    1
Luggage                         1
Name: product_category, dtype: int64
```



```
us.product_category.value_counts()
```

```
Mobile_Apps                 21056
Digital_Ebook_Purchase      18173
Video DVD                   15949
Digital_Video_Download      15427
Books                       12097
Music                       11148
Digital_Music_Purchase       1488
Toys                          820
PC                            766
Video                         666
Home Entertainment            512
Wireless                      304
Camera                        272
Video Games                   226
Musical Instruments           167
Electronics                   160
Watches                       151
Tools                         119
Shoes                         111
Baby                          100
Sports                         63
Outdoors                       52
Home Improvement               51
Home                           35
Office Products                26
Kitchen                        21
Health & Personal Care         15
Lawn and Garden                10
Mobile_Electronics              5
Automotive                      4
Luggage                         1
Personal_Care_Appliances        1
Grocery                         1
Apparel                         1
Software                        1
Beauty                          1
Name: product_category, dtype: int64
```

Deeper architectures



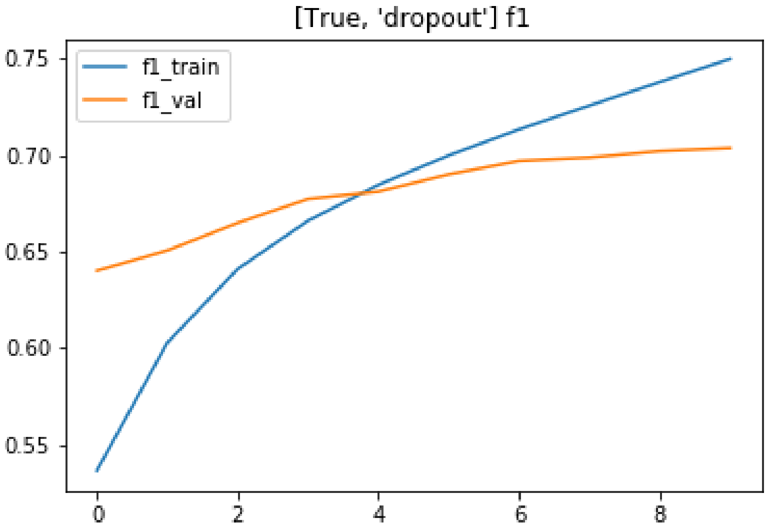
Train data	**UK** (100k)
Test data	DE (100k)

Train Embeddings

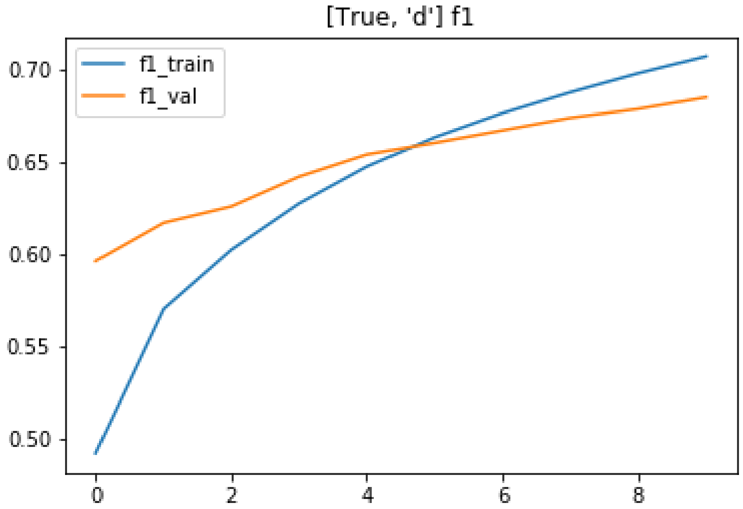
- Quickly overfits
- Even with dropout and reduced learning rate

Train data	**US** (100k)
Test data	DE (100k)

lr: 0.001



lr: 0.0005



Progress

- Colab issues –> switched to google cloud compute engine
- Training 100k examples takes 5 min per epoch (15 epochs ~45 min.)
- Training 1 Million examples takes 30 min per epoch (15 epochs ~7-8 hours)

- Problem: balance between training size and computation time

Roadmap

- Continue with architecture search
- Balanced training

- Run baseline architecture on organic dataset
- Fine-tune on organic dataset