```
In [1]: import logging
        from utils import logging as lg
        lg.set_logging(logging.ERROR)
        from skimage.measure import block_reduce
        import numpy as np
        import logging
        import pickle
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        sns.set(color_codes=True, font_scale=2, style="whitegrid", palette="mute
        d")
        from notebook_utils import plot
        from model import base, provider, heatmap_evaluation
        import config
        import tensorflow as tf
        tf.logging.set verbosity(tf.logging.ERROR)
        %matplotlib inline
In [2]: from utils import data provider
In [3]: dataset_loader = data_provider.DatasetLoader(data_dir='../data')
In [4]: | def plot_heatmaps(_model, dataset, seq):
            if model == 'shallow':
                model = 's2'
            elif model == 'deep':
                model = 's3'
            elif model == 'deepv2':
                model = 'deep_41'
            elif model == 'convdeep':
                model = 'convdeep 41'
            model_path = '.%s' % provider._model_path(model, dataset, seq)
            print(model path)
            plot.plot_relevance_methods(model_path, dataset_loader, only_positiv
        e_rel=True, methods=['sensitivity', 'guided_backprop', 'lrp_alpha2_beta
        1', 'lrp deep taylor'])
```

```
In [5]: def plot_relevance_dist_in_middle_region(datasets=['mnist-3-digits'], se
        q=12, methods=['sensitivity', 'simple_taylor', 'guided_backprop', 'lrp_a
        lpha2_beta1', 'lrp_deep_taylor']):
            results = []
            print(results)
            for dataset in datasets:
                for model in ['s2', 's3', 'deep_41', 'convdeep_41']:
        #
                      print(model)
                    file = "../stats/rel-dist-%s-seq-%d-%s.pkl" % (dataset, seq,
         model)
                    try:
        #
                          print('getting data from %s' % file)
                        results = results + pickle.load(open(file, "rb"))
                    except:
                        print('%s not found' % file)
        #
              print(results)
            df = None
            df = pd.DataFrame(results)
            df = df[df.method.isin(methods)]
            def get_marker_linestyle(method):
                if method == 'guided_backprop':
                    mk = 's'
                    ls = '-'
                elif 'lrp' in method:
                    mk = '^'
                    ls = '-'
                else:
                    mk = '.'
                    ls =':'
                return mk, ls
            marker linestyles = [get marker linestyle(m) for m in methods]
            markers = [ m[0] for m in marker linestyles ]
            linestyles = [ m[1] for m in marker linestyles ]
            df['architecture_idx'] = df['architecture'].apply(plot.architecture_
        idx)
            col name = 'Percentage of relevance \n in data region'
            df[col_name] = df['rel_dist_in_data region']
            for c in [col_name]:
                g = sns.factorplot(x="architecture idx", y=c, col='dataset', hue
        ="method",
                                    data=df, size=5, markers=markers,
                                    linestyles=linestyles)
                g.set xticklabels(['Shallow', 'Deep', 'DeepV2', 'ConvDeep'], rot
        ation=15)
                g.set(xlabel='')
            return df
        # plt.savefig('rel-dist-3digits.pdf')
```

Experiment concatenated mnists with correct class having 2 digits

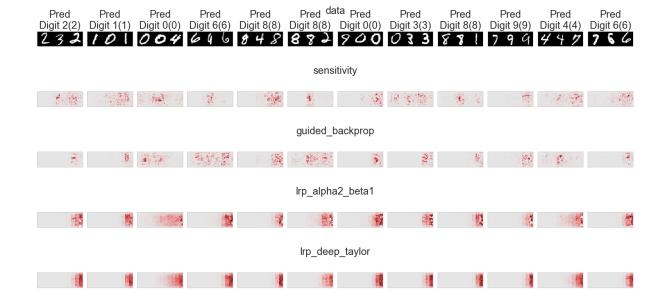


Heatmaps of MNIST

In [6]: plot_heatmaps('shallow', 'mnist-3-digits-maj', 12)

../final-models/s2_network-mnist-3-digits-maj-seq-12

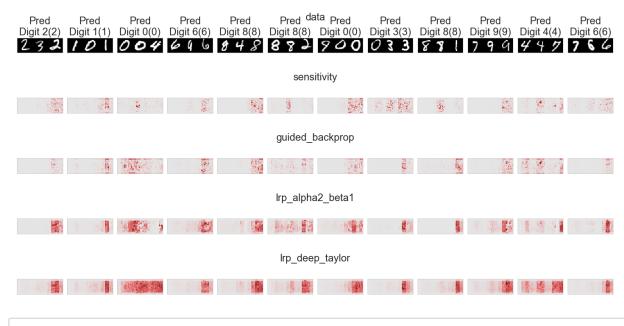
Heatmaps from different explaination methods s2_network:hidden:256[out:10 -recur:256 s2_network-minist-3-digits-maj-seq-12---2018-02-20-21-57-02 (no. variables 184330) (opt AdamOptimizer, acc 0.9806, keep prob 0.80)



In [7]: plot_heatmaps('deep', 'mnist-3-digits-maj', 12)

../final-models/s3_network-mnist-3-digits-maj-seq-12

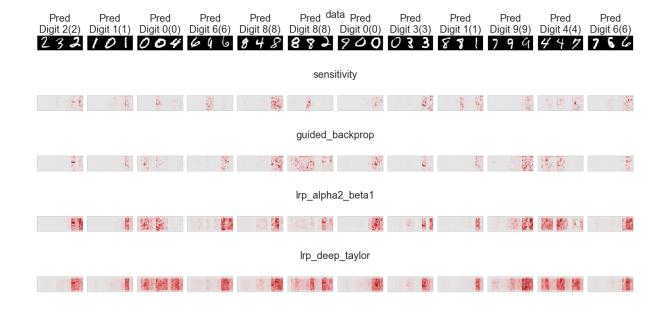
Heatmaps from different explaination methods s3_network.in1:256|inident:128|out1:30_out1:01-recur:256 s3_network-minist-3-digits-maj-seq-12---2018-02-20-21-57-01 (no. variables 153578) (opt AdamOptimizer, acc 0.9839, 8eep_prob 0.80)



In [8]: plot_heatmaps('deepv2', 'mnist-3-digits-maj', 12)

../final-models/deep_4l_network-mnist-3-digits-maj-seq-12

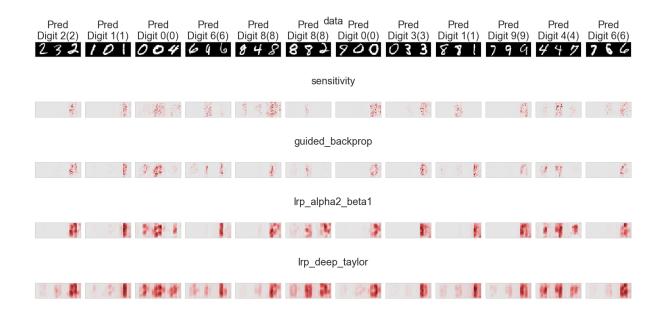
Heatmaps from different explaination methods deep_4l_network.ir128|in2-256|iniden:128|ou1:32|ou12:10-ercur:256 deep_4l_network-misk-3-digits-maj-seq-12--2018-02-20-21-58-31 (no. variables 161386) (opt AdamOptimizer, aco. 0.9809, keep_prob. 0.80)



In [9]: plot_heatmaps('convdeep', 'mnist-3-digits-maj', 12)

../final-models/convdeep_4l_network-mnist-3-digits-maj-seq-12

convdeep_4I_network.conv1:5x5x24=>2x[2]_gloruv2:3x322=>2x2[2]gln1:128|hidden:128|out1:32|out2:10-recur:256 convdeep_4I_network.mist3-digits-maj-seq-12-2018-02:21-01:20-50 (nc. variables 151802) (opt AdamOpthimeze, acc 0 9903, keep_prob 0 80)

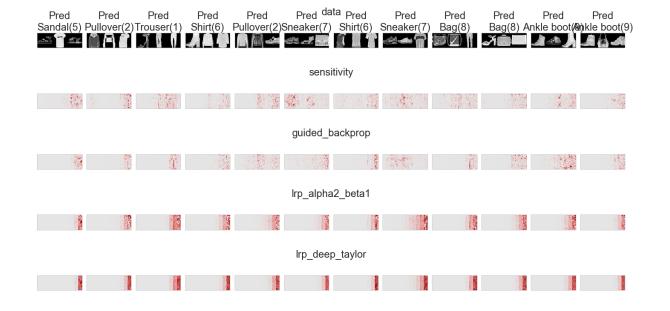


Heatmaps of FashionMNIST

In [10]: plot_heatmaps('shallow', 'fashion-mnist-3-items-maj', 12)

../final-models/s2_network-fashion-mnist-3-items-maj-seq-12

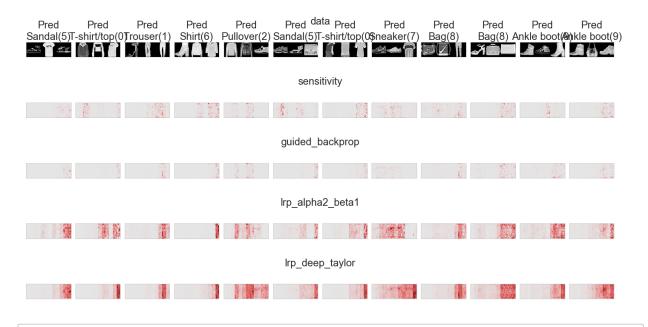
Heatmaps from different explaination methods £2_network: hidden_256[out.10-recur_256] \$2_network-fashion-mnist-3-ifems-maj-seq-12—2018-02-21-01-09-24 (no. variables 184330) (opt AdamOptimizer, acc 0.9020, keep prof



In [11]: plot_heatmaps('deep', 'fashion-mnist-3-items-maj', 12)

../final-models/s3_network-fashion-mnist-3-items-maj-seq-12

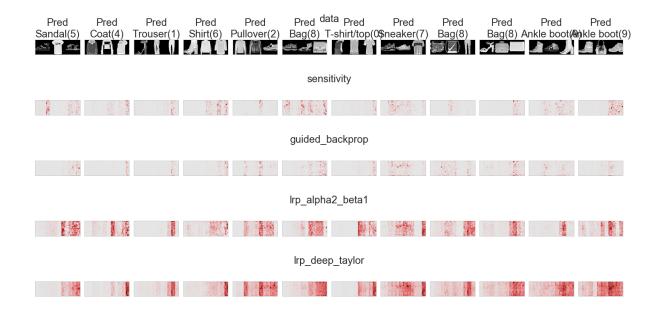
Heatmaps from different explaination methods \$3_network.i125(l)id.126(l)id.212(10-recur256 \$3_network.fashion-minist.3-items_maj.seq.12.—2018.02.21.—01.15.18 (no. variables 153578) (opt AdamOptimizer, acc 0.9058, keep_prob.0.80)



In [12]: plot_heatmaps('deepv2', 'fashion-mnist-3-items-maj', 12)

../final-models/deep_4l_network-fashion-mnist-3-items-maj-seq-12

Heatmaps from different explaination methods deep_4I_network:n1:128ii_2256|hidden:128|out1:312out2:10-recur:256 deep_4I_network-fashion-mnist-3-tems-maj-seq-12--2018-02-21--01-16-19 (no. variables 161386) (opt_4damOptimizer, acc. 0.9071, keep_prob_0.80)

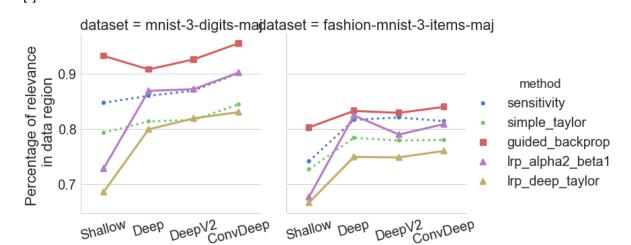


.../final-models/convdeep_4l_network-fashion-mnist-3-items-maj-seq-12

Predictions Indianal Predictions** Indi

plot_heatmaps('convdeep', 'fashion-mnist-3-items-maj', 12)

Relevance Distributions



Model Accuracy

Out[15]:

	seq	Shallow	Deep	DeepV2	ConvDeep
0	12	0.9806	0.9839	0.9809	0.9903

Out[16]:

	seq	Shallow	Deep	DeepV2	ConvDeep
0	12	0.902	0.9065	0.9071	0.9238