Factorized Variational Autoencoders for Modeling Audience Reactions to Movies

Supplementary Material

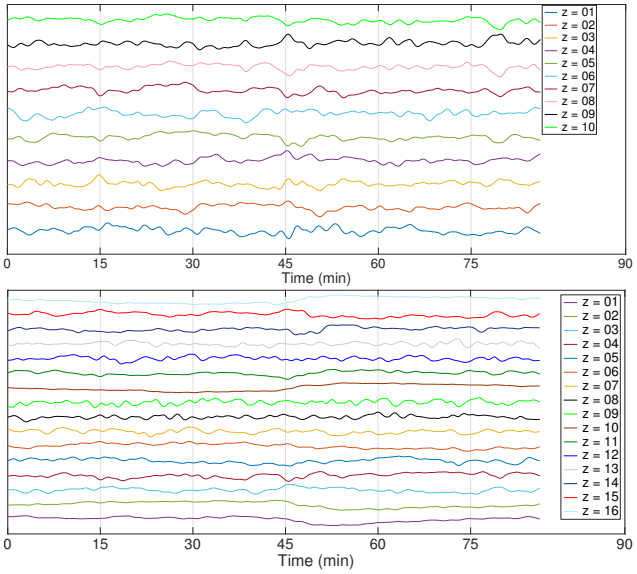


Figure 1. V_t **matrix visualization**. Visualization of the V_t as a temporal signal for each dimension for (top row) TF and (bottom row) NLTF for the movie *Inside Out*.

Movie	Latent Size K	Reconstruction MSE TF	NLTF	FVAE	Prediction MSE TF	NLTF	FVAE
Inside Out	2	1448.1	1297.4	853.8	4065.1	1638.8	1702.9
	4	1320.5	1211.8	597.5	1789.3	1429.7	1376.6
	8	1255.41	1166.8	347.4	1862.1	1384.4	1183.5
	10	1214.87	1145.6	272.7	1977.9	1292.9	1223.4
	16	1194.6	1132.7	262.1	2010.3	1240.	**1161.4**
	32	1168.5	1148.4	202.2	2666.8	1269.1	1176.9
Zootopia	2	1417.7	1217.5	847.3	2222.5	1709.9	2150.0
	4	1269.0	1158.1	610.8	1521.0	1539.7	1422.1
	8	1202.9	1144.4	340.4	1414.8	1420.0	1193.0
	10	1181.0	1163.8	266.2	1277.4	1367.9	1178.1
	16	1161.1	1153.8	189.9	1357.4	1407.5	**1153.7**
	32	1148.4	1132.6	169.9	2515.2	1479.6	1164.6
Good Dinosaur	2	1429.0	1219.0	825.7	1797.9	2034.5	1682.6
	4	1235.8	1184.4	589.1	1421.6	1621.4	1440.6
	8	1177.6	1160.7	357.6	1410.1	1505.6	1308.0
	10	1156.5	1139.9	329.7	1328.4	1453.5	1261.5
	16	1147.3	1132.1	275.4	1454.5	1416.1	**1244.7**
	32	1134.7	1131.3	237.2	1380.1	1417.7	1287.7
The Jungle Book	2	1383.1	1255.1	959.2	2154.2	1487.3	1515.8
	4	1239.1	1124.2	619.7	2262.1	1365.9	1607.5
	8	1185.5	1153.1	333.0	2065.0	1332.5	1181.6
	10	1150.0	1117.8	252.2	1622.1	1307.8	1114.3
	16	1137.2	1115.3	186.4	1715.3	00.2	1099.8
	32	1119.1	1111.7	182.8	2439.9	1330.8	**1098.2**

Table 1. **Performance:** The training and testing performance error for all the models with varying sized latent spaces

1. Implementation

We used minibatch of size 10 and perform the stochastic gradient descent algorithm to train TF, NLTF and FVAE models. For the multi-layer neural network, that we used in nonlinear tensor factorization and encoder/decoder of Factorized VAE, we used three stacked fully connected layers with *ReLU* as activation function. The models are trained for 16 epochs. We implemented these models in *Torch*[1] and conduct the experiment with Tesla K40 GPU.

1.1. Error Visualisation

We report the error values (RMSE/MSE) in the the paper for each audience member rather than for each landmark on the face. Figure 2 shows the comparisons of ground-truth landmarks with the generated landmarks for different RMSE values for better visualisation to understand the values reported in the paper.

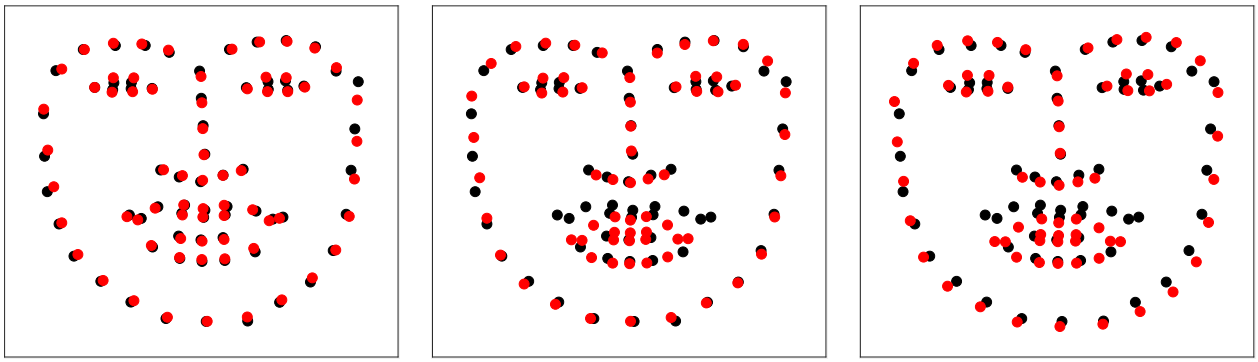
[1] http://torch.ch

2. Full Breakdown of Experiments

In this section we show the further explanation/breakdown of our experimental results in Section 5.1, 5.2 and 5.4 in the paper.

2.1. K values exploration (Section 5.1)

Here we show the exploration experiments of TF, NLTF, and FVAE for different K values. We compare the reconstruction error during training and prediction error during testing on $K = \{2, 4, 8, 10, 16, 32\}$. Note that we report the error in terms of mean square error (MSE) for each audience member rather than for each landmark on the face. Figure 5 in the paper shows the average results across all movies (Refer Table 1).



(a) (b) (c)

Figure 2. **Error (RMSE/MSE) Visualisation**. Visualization of RMSE/MSE error. Black points indicate the ground-truth landmark values while red points indicate the generated landmarks. (a) FVAE (19.9 RMSE) (b) NLTF (50.3 RMSE) and (c) TF (53.2 RMSE)

2.2. Visualizing the Latent Factors (Section 5.2)

Visualization of the V_t as a temporal signal for each dimension TF and NLTF models is given in Figure 1 for movie *Inside Out*. TF has 10 dimensions and NLTF has 16 dimensions. Here, we can see that since there's no semantically meaningful latent factors, the signals across each dimension for TF and NLTF doesn't show any strong correspondence with the meaningful audience reaction.

The FVAE learned for *Zootopia* exhibits a similar 'smile' latent factor. As Fig. 3 shows, humorous scenes in the movie again correspond with significant peaks in the time series.

2.3. Time Analysis (Section 5.4)

In this section we report the long-term prediction of audience reactions from different models by observing. We investigate different fractions $\lambda \in \{0.001, 0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8\}$ of the test data to use as observations to estimate $\mathbf{U}_i$. Figure 9 depicts the average results for each genre (i.e *Animation, Drama* and *Action & Adventure*.

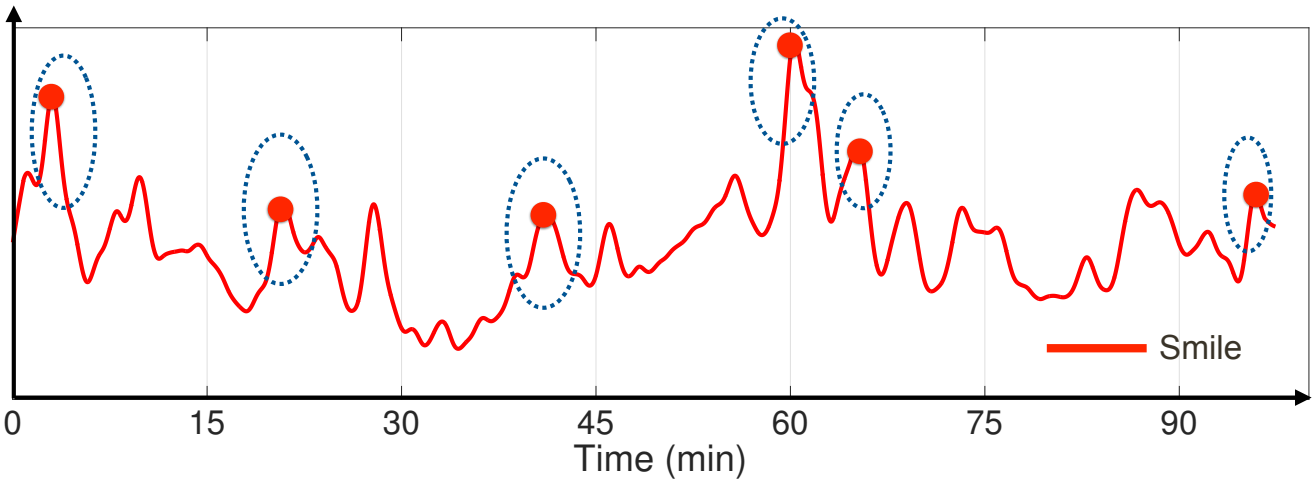


Figure 3. **Latent Temporal Factors**. The $\mathbf{V}_t$ elements that modulate the smiling correlate with humorous scenes in the movie *Zootopia*.

Movie	Model	Prediction Error for Observed Percentage λ (MSE)							
		0.001	0.01	0.05	0.1	0.2	0.4	0.6	0.8
Ant Man	TF	2056.3	1976.5	1987.4	2038.1	1949.9	2214.3	1910.1	1948.7
	NLTF	1971.3	1653.2	1563.3	1534.6	1519.5	1568.8	1530.1	1528.3
	FVAE	1717.3	1544.2	1482.2	1428.9	1422.6	1377.8	1406.6	1370.7
Big Hero 6	TF	8085.0	2026.4	1977.5	1939.3	1915.8	1945.8	1913.8	1945.3
	NLTF	2342.6	1702.5	1611.5	1590.9	1564.1	1538.0	1534.2	1526.0
	FVAE	2754.4	1628.1	1562.6	1517.8	1525.5	1561.2	1563.6	1513.7
Bridge Of Spies	TF	1903.1	1590.2	1435.8	1422.2	1656.1	1414.5	1304.5	1391.6
	NLTF	1907.7	1454.8	1395.3	1428.7	1198.2	1208.6	1275.0	1220.4
	FVAE	1490.9	1256.8	1098.7	1072.5	1052.9	1043.2	1053.2	1055.8
The Good Dinosaur	TF	2694.6	1817.0	1569.9	1626.1	1673.9	1680.2	1588.6	1721.4
	NLTF	2882.6	2233.1	1728.9	1831.8	1652.7	1595.1	1465.3	1443.4
	FVAE	1547.0	1480.5	1374.9	1377.8	1317.6	1308.5	1296.8	1254.2
Inside Out	TF	1910.9	1905.8	1900.3	1752.4	1682.2	1589.6	1690.5	1649.2
	NLTF	1835.9	1477.7	1357.3	1370.2	1287.7	1406.1	1312.8	1308.8
	FVAE	1368.7	1265.7	1205.9	1165.7	1129.7	1140.1	1133.8	1086.7
Jungle Book	TF	2131.2	2471.0	2507.3	2046.1	2051.9	2134.0	1696.8	1703.3
	NLTF	2407.7	1546.3	1415.7	1495.4	1583.4	1315.8	1342.7	1324.1
	FVAE	1470.1	1363.0	1241.6	1204.4	1180.7	1182.5	1199.9	1128.6
Star Wars: TFA	TF	3053.0	2098.6	2253.7	2002.9	1963.6	1927.7	1720.4	1608.5
	NLTF	1731.5	1448.4	1424.9	1314.1	1341.6	1207.7	1252.5	1114.1
	FVAE	1407.7	1287.6	1164.5	1139.0	1123.4	1102.8	1089.3	1015.9
The Finest Hour	TF	2428.8	1981.1	1799.6	1970.5	1902.0	1984.8	1921.2	1821.5
	NLTF	1903.1	1590.2	1435.8	1422.2	1656.1	1414.5	1304.5	1391.6
	FVAE	1662.0	1381.2	1154.3	1097.8	1112.4	1087.2	1035.5	1013.0
Zootopia	TF	2007.9	1782.8	1743.9	1524.7	1682.1	1537.5	1580.9	1596.9
	NLTF	2297.4	1958.9	1653.7	1598.8	1499.6	1521.1	1580.4	1512.4
	FVAE	1841.3	1656.7	1474.2	1333.4	1327.6	1328.6	1302.2	1278.6

Table 2. **Predicting reactions:** We predict the future facial landmarks of each audience member using TF, NLTF and FVAE models based on varied proportion of observations in a given movie.