Fashion Forward: Forecasting Visual Style in Fashion Supplementary Material

Ziad Al-Halah[1] Rainer Stiefelhagen[1] Kristen Grauman[2]
[1]Karlsruhe Institute of Technology [2]The University of Texas at Austin
{ziad.al-halah, rainer.stiefelhagen}@kit.edu, grauman@cs.utexas.edu

This document provides additional information for:

- The deep attribute and the ClothingNet architectures.
- The forecast baseline models.
- The discovered topics on the Shirts dataset (see Fig. 1).
- Forecast examples of our model in comparison to the baselines on the three datasets (see Fig. 2).
- Analysis of varying the number of styles K on the forecast performance.

1. The deep attribute model

Fig. 3 shows the details of the network architecture for our attribute prediction model. The model is composed of 5 convolutional layers with decreasing filter sizes from 11×11 to 3×3 followed by 3 fully connected layers and 2 dropout layers with probability of 0.5. Additionally, each convolutional layer and the first two fully connected layers in our model are followed by a batch normalization layer and a rectified linear unit (ReLU). For information on the training procedure and the hyperparameters see Section 3.1 in the main submission.

2. ClothingNet

The ClothingNet model is similar to our attribute model architecture with the last sigmoid layer replaced with a softmax. The network is trained to distinguish 50 categories of garments (*e.g.* *Sweater*, *Skirt*, *Jeans* and *Jacket*) from the DeepFashion dataset. The model is trained for 45 epochs using Adam [3]. On a held-out test set on DeepFashion, the ClothingNet achieves 86.5% Top-5 accuracy.

3. Forecast models

Naïve which includes three simple models:

1) *mean*: the future values are forecasted to be equal to the mean of the observed series, *i.e.* $\hat{y}_{n+1|n} = \frac{1}{n}\sum_{t=1}^{n} y_t$.
2) *last*: the forecast is equal to the last observed value, *i.e.* $\hat{y}_{n+h|n} = y_n$.

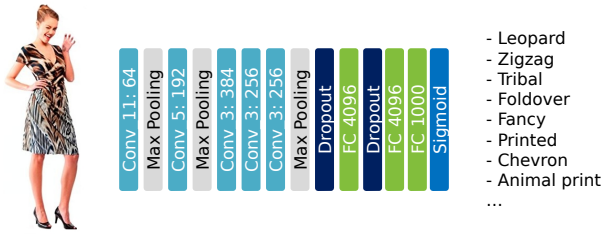


Figure 3: The architecture of our deep attribute CNN model.

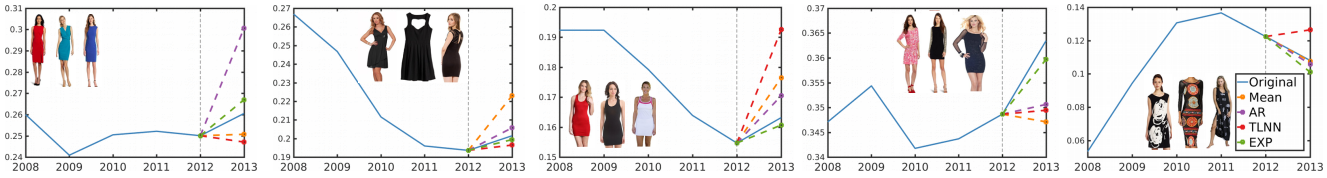
3) *drift*: the forecast follows the general trend of the series, *i.e.* $\hat{y}_{n+h|n} = y_n + \frac{h}{n-1}(y_n - y_1)$ where h is the forecast horizon.

Autoregressors these linear regressors assume the current value to be a linear function of the last observed values "lags", *i.e.* $\hat{y}_n = b + \sum_i^P \alpha_i y_{n-i} + \epsilon$ where b is a constant, $\{\alpha_i\}$ are the lag coefficients, P is the maximum lag (set by cross validation in our case) and ϵ an error term. We consider several variations of the model [1]:

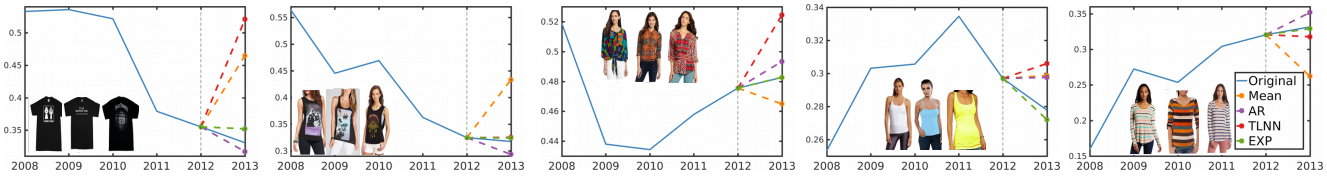
1) *AR*: the autoregressor in its standard form.
2) *AR+S*: which further incorporates seasonality, *e.g.* for a series with 12 months seasonality the model will also consider the lag at $n-12$ along with most recent lags to predict the current value.
3) *VAR*: the vector autoregoressor considers the correlations between the different styles trajectories when predicting the future.
4) *ARIMA*: the autoregressive integrated moving average model which models the temporal trajectory with two polynomials, one for autoregression and the other for the moving average. In addition it can handle non-stationary signals through differencing operations (integration).

Neural Networks (NN) Similar to the autoregressor, the neural models rely on the previous lags to predict the current value of the signal; however these models incorporate non-linearity which make them more suitable to model complex

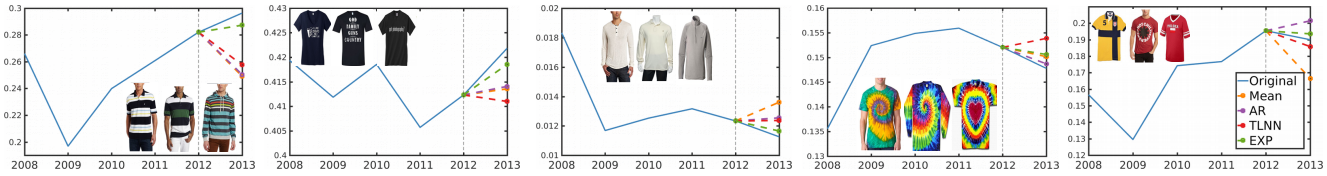
Figure 1: The discovered visual styles on the Shirts dataset with their visual signature on top defined by semantic attributes. For discovered styles in Dresses and Tops&Tees see Figure 3 in the main submission.



(a) Dresses



(b) Tops&Tees



(c) Shirts

Figure 2: The forecasted popularity of the visual styles in (a) Dresses, (b) Tops&Tees and (c) Shirts. Our model (EXP) successfully captures the popularity of the styles in year 2013 with minor errors in comparison to the baselines.

time series. We consider two architectures with sigmoid non-linearity:

1) *TLNN*: the time lagged neural network [2].

2) *FFNN*: the feed forward neural network.

Fig. 2 shows the style popularity forecasts estimated by baselines from the three previous groups in comparison to our approach. The Naive and NN based forecast models seem to produce larger prediction errors. Our model performs the best followed by the Autoregressor (AR). For quantitative comparisons and more detailed discussion see Section 4.2 in the main submission.

4. Number of Styles

Table 1 shows the performance of our model in terms of forecasting error when varying the number of styles K between 15 and 85. We notice that increasing K results in introducing more noise in the time line of the style as some

#Styles	Dresses	Tops & Tees	Shirts
15	7.70	6.71	**3.03**
30	**6.54**	**5.36**	3.16
45	8.15	5.98	3.78
70	8.22	5.60	4.10
85	10.66	5.62	4.14

Table 1: The forecast error (MAPE) of our approach using varying number of styles.

of them doesn't capture a consistent style in the data and the forecasting error increases. Nonetheless, the variance in performance is still acceptable for the tested K values.

From the visual perspective, we see that at K=30 the styles have a coherent visual appearance of mid-level granularity. However, capturing the visual quality of the discovered styles in a quantitative manner is not a trivial task. We believe this is an interesting and important question for future investigation.

References

[1] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015. 1

[2] J. Faraway and C. Chatfield. Time series forecasting with neural networks: a comparative study using the airline data. *Applied statistics*, pages 231–250, 1998. 2

[3] D. P. Kingma and J. L. Ba. ADAM: A Method for Stochastic Optimization. In *ICLR*, 2015. 1