The multi-modal universe of fast-fashion: the Visuelle 2.0 benchmark Supplementary Material

Geri Skenderi[1] Christian Joppi[2] Matteo Denitto[2] Berniero Scarpa[3] Marco Cristani[1,2]

[1] Università degli Studi di Verona [2] Humatics s.r.l. [3] Nuna Lie s.r.l.

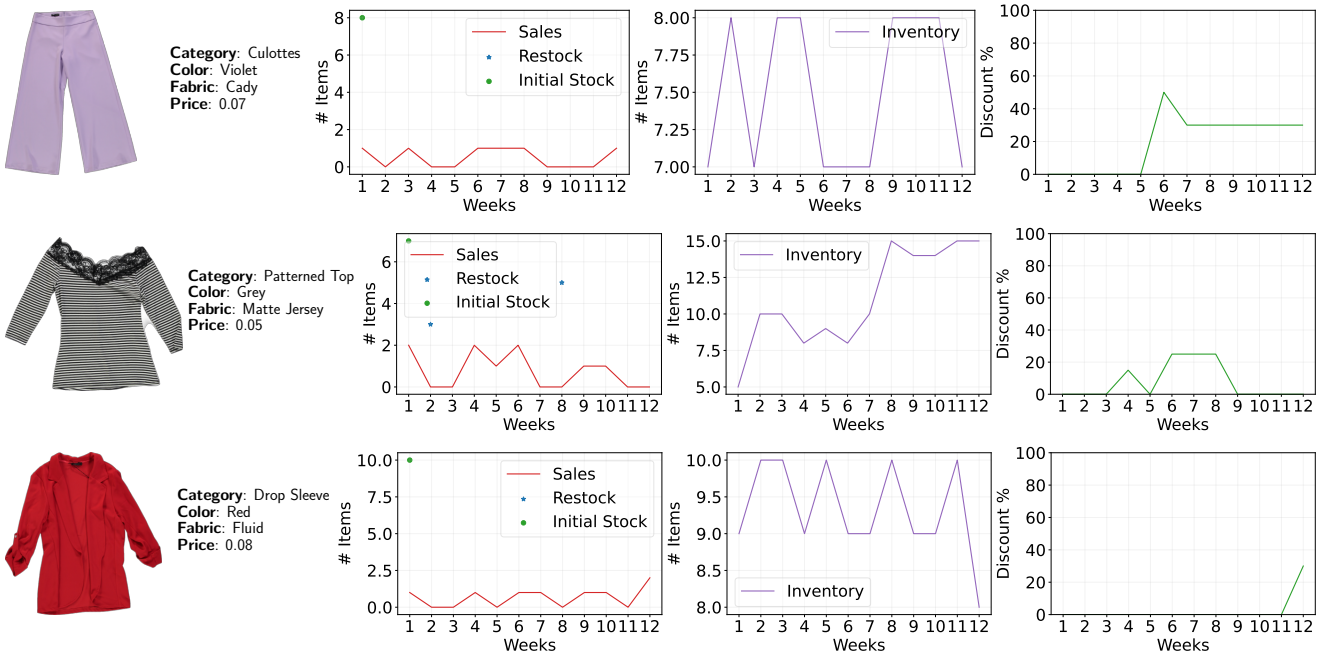

Figure 1. Examples of different products that have been sold in the shop SHOP11 during the AW19 season. The figure reports (from left to right): the product's image; textual tags; sales time series; restocking information; inventory time series; and discount time series.

1. Outline of this Supplementary Material

The main paper discusses two contributions: 1) the novel Visuelle 2.0 dataset and 2) the advantage of utilising a multi-modal approach for short-observation new product forecasting (SO-fore). This supplementary material adds to each one of these main topics, as follows:

The Visuelle 2.0 dataset The following topics will be faced in Sec. 2:

- How Visuelle 2.0 compares with respect to the current datasets of forecasting for fast-fashion, exploring the literature of time-series forecasting and computer vision in this area (Sec.2.1);
- A showcase of the data contained in Visuelle 2.0, including examples of products, the associated *time series data*, *image data* and the *text data*. An excerpt of these data is reported in Fig. 1. Subsequently, we will show some examples of *exogenous data* i.e., the Google Trends associated to a given product, and the weather reports associated to a given shop; finally, *customer purchase data* will be presented (Sec. 2.2);
- A list of possible challenges that can be studied on Visuelle 2.0 will be presented, and specifically: more tasks related to *new product short observation forecasting*; the *new product demand forecasting*; the *product recommentation* (Sec. 2.3).

The SO-fore problem In Sec.3, the Cross-Attention RNN will be detailed, with a graphical representation that illustrates its components (Sec. 3.1).

2. The Visuelle 2.0 dataset

2.1. Datasets for forecasting in fast fashion

The Visuelle 2.0 dataset can be related to two scientific fields: 1) the one of forecasting for (fast) fashion [7, 17, 25], with particular emphasis on those works that exploit deep learning techniques [8, 15]; 2) the recent field at the intersection between computer vision and fashion [6], with special emphasis on the task of *popularity prediction* and *fashion forecasting*. In both cases, Visuelle 2.0 innovates for specific reasons which will be detailed in the following, in two separate sections. In both the cases, Visuelle 2.0 has an unprecedented richness of data which other datasets do not possess, such as customer purchase information, and the different exogenous data.

Forecasting for (fast) fashion The most used forecasting techniques for fast fashion incorporate classical ARIMA, SARIMA, exponential smoothing [5], regression [18], Box & Jenkins [4] and Holt Winters [28], as reported in the recent review of [15]; machine learning approaches (decision trees, random forests, SVMs, neural networks) are at their infancy on this topic and most importantly they are not considering multi-modal data, but only time-series. This has naturally created an abundance of datasets for time-series analysis for sale forecasting [20] and demand forecasting [19], and the absence of datasets with images included, which we are filling with Visuelle 2.0. As a notable exception, the work of [8] proposes a set of techniques for demand forecasting, in which images are taken into account by an Attention-based RNN framework, which we also utilise for SO-fore as explained in Sec. 3.1. Unfortunately, the dataset on which they perform their experiments is not publicly available, while Visuelle 2.0 will be made publicly available.

Intersection between computer vision and fashion The peculiarity of this field is the exploration of multi-modal data (images, text, time-series) for prediction tasks. In general, computer vision approaches have been considered for the task of *popularity prediction* or *fashion forecasting* [6] In both the cases, the ground truth signal is built on top the public ratings obtained on online platforms such as Chictopia.com [22, 29], Lookbook.nu [14]or Amazon [1], which consider outfits [14, 22, 22, 29], or several outfits exhibiting the same style [1]: an outfit or a style is popular if it receives a high rating in terms of number of "likes" or "stars". In the case of Visuelle 2.0, we can assume that a product is more popular than others of the same category, if in the same season, it has sold more. In this setup, our dataset allows to be more fine-grained, since one can predict the popularity, in terms of sales, of a single product. Also, Visuelle 2.0 represents the very first dataset which permits multi-modal analysis on the data of a real fast fashion company, meaning that approaches which succeed on this benchmark can be directly applicable on the fast fashion market.

2.2. Visuelle 2.0: a showcase

Example of products In Sec.2 of the main paper we have given some statistics about the products which are in the Visuelle 2.0 dataset. Here we will give some qualitative examples of their *image data*, *text attributes* and associated time series: *product sales*, *inventory position*, *Restock flag*, and *discount*, the latter omitted in the main paper due to the lack of space. Formally, following the notation of the main paper, given a product i at a retail store r, we refer to its *discount* signal as $D(i, r, t)$ where t refers to the t-th week of market delivery, with $i = 1, ..., N$, $r = 1, ..., L$ and $t = 1, ..., K$. The discount signal is expressed as a percentage, describing how much a particular item is discounted; for example, $D(i, r, t) = 20$ indicates that the initial price defined for a product i in the retail store r was discounted by a 20% at time t. Fig. 1 showcases all of these data for three products of season AW19. Other figures can be found at the end of this additional material, reporting products of other seasons (Fig. 6, Fig. 7 and Fig. 8).

Exogenous data Exogenous data is often neglected as a resource within datasets, especially in forecasting. This is due to their nature, since they are, by definition, coming from an external phenomenon that is not directly related with the data being analysed. Nevertheless, adding exogenous variables such as weather data [3, 24] or popularity data [21, 23] to forecasting models has proven extremely beneficial in terms of forecasting performance. For this reason, we provide in Visuelle 2.0 multivariate exogenous data both for the weather and popularity, in the form of detailed weather reports (Fig. 2) and Google Trends (Fig. 3). An more profound explanation for both examples is provided in the respective captions.

Customer purchase data Quoting Sec.2 and Fig.6 of the main paper, among the 667086 total registered users of Nuna Lie, 6k users have bought continuously a total of 25 products over 4 seasons. In Fig. 4 we have a random excerpt of 10 of these users, with a random subset of 9 purchased items each (no cherry picking); in many cases, personal styles do emerge, showing systematic preferences on diverse attributes, as written in the caption of the figure.

2.3. Challenges on Visuelle 2.0

In this paper, we explored the problem of short-observation forecasting on Visuelle 2.0, with the precise focus of showing the benefit of the image data on this task. Obviously, we are far from saturating the performance, encouraging further improvements. These could be provided by diverse techniques (exploiting LSTM or transformer-based architectures), or including additional

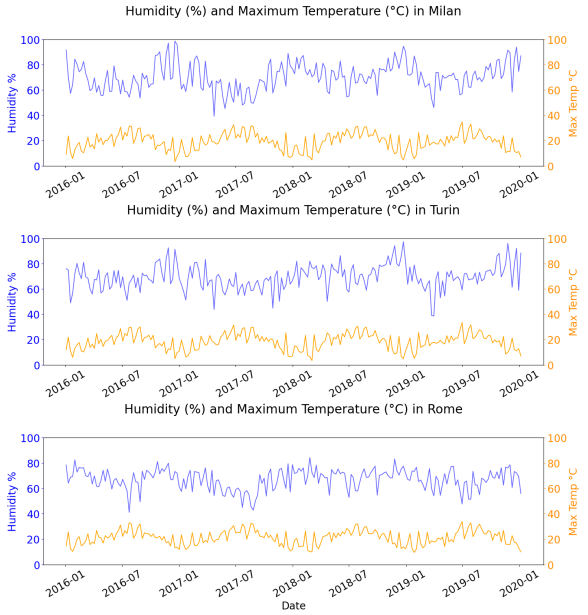

Figure 2. Excerpt of exogenous weather data (humidity and maximum temperature) for three major Italian municipalities: Milan, Turin and Rome. Humidity in these areas tends to be negatively correlated with maximum temperature. Rome tends to be much hotter than Milan and Turin throughout the year. An interesting observation which can prove beneficial to forecasting is the seasonal nature of weather phenomena, which can induce information as to which clothing products may sell more in a particular period.

exogenous data, available in Visuelle 2.0. Google Trends and weather reports, in fact, are signals which have been shown elsewhere to be predictive [3, 21, 23, 24], so this should be a natural next step.

Other challenges which can be experimented on Visuelle 2.0 are listed in the following.

- **Demand forecasting.** Forecasting demand is a crucial issue for driving efficient operations management plans [10, 17, 26]. This is especially the case in the fast fashion industry, where demand uncertainty, lack of historical data, variable ultra-fast life-cycle of a product and seasonal trends usually coexist [13, 16]. In rough terms demand forecasting outputs the amount of goods to buy from the suppliers. This amount is then distributed among the different retailers, with the aim of avoiding zero-stocks or excessive unsold inventory. In this paper we show a glimpse of demand forecasting on Visuelle 2.0, at the level of single shop (i.e. predicting how much a single shop will need during the next season), adopting the recent RNN-based approach of [8] on time series and time series + image. In [23] we report some results on the aggregated signal in the old Visuelle dataset; it is worth remembering that in that case the signal about the single shops was missing, less products were available, and the only important result was to show how Google Trends data are beneficial. In this case, a deep analysis on demand forecasting on Visuelle 2.0 needs to be carried out, including discounts and exogenous signals like weather reports (Fig. 2) and Google Trends (Fig. 3).

- **Product recommendation.** An important feature of Visuelle 2.0 is the presence of customer purchase data; 667086 customers have bought along 8 seasons a total amount of 3253876 items, which cover a consistent percentage (84%) of the total purchases collected within the dataset. A graphical representation of these data is reported in Fig. 4, where it is visible that some users have marked preferences.

  Product recommendation on these data would consist in defining a particular time index t_{rec}, when the historical data of all the past purchases (older than t_{rec}) of all the customers will be taken into account. Therefore, two types of inferences will be possible: 1) to suggest which product (or category, or attribute) z_k a specific customer u_i could be interested in; a positive match will be in the case of an effective purchase of z_k (or some item which is in the category z_k or that expresses the attribute z_k) by u_i after time t_{rec}; 2) same as before, but including a specific time interval T_{buy} within which the customer will buy. In practice, a positive match will be in the case of an effective purchase of z_k (or some item which is in the category z_k or that expresses the attribute z_k) by u_i *in the time interval* $]t_{\text{rec}}, t_{\text{rec}} + T_{\text{buy}}]$. In general, product recommendation can be carried out by standard collaborative-filtering based techniques, but also considering recommendation as an instance of forecasting [9, 12] and viceversa: this interplay could be certainly explored with the Visuelle 2.0 dataset.

3. The SO-fore problem

3.1. Cross-Attention RNN

The Cross-Attention RNN [8], can be described as an autoregressive, sequence-to-sequence neural network that tries to understand the different, non-linear relationships in the various data modalities and then perform predictions by understanding which part of the data is most important for the forecasting task. The attention modules constitute a large part of the model and are exactly as in [2], where at each decoding step we try to attend to the encoder features based on the current decoder hidden state.

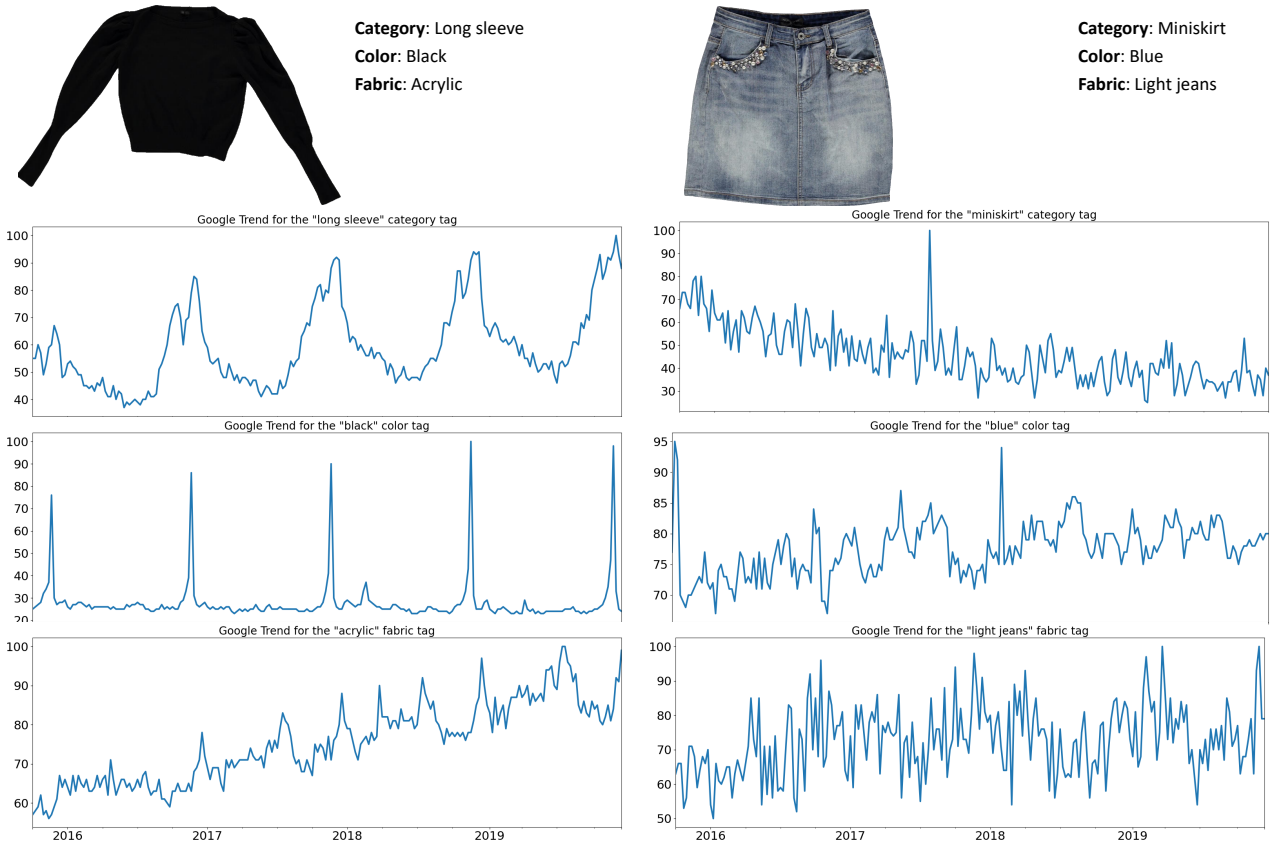

Figure 3. Example of the exogenous Google Trends data available in Visuelle 2.0 for two completely different products. The signals can have regular trend and seasonality (left) or be stationary and seem noisy (right). This can prove helpful for forecasting models, because it allows to understand both global and cyclical popularity and therefore anticipate sales.

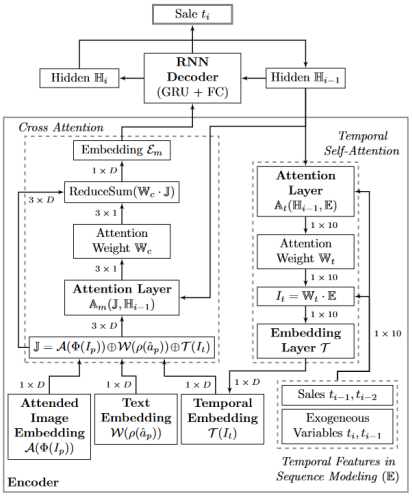

Figure 5. A visual description of the Cross-Attention RNN model, the neural network architecture used in our SO-fore experiments. Taken from [8].

The encoder starts by embedding each input modality into a common feature space R^D. The input observation sales (time series) are first passed through an additional self-attention layer [27], differently from the original work in [8] and then projected through a fully connected layer. This helps filter out initial noise from past sales observations. The image and textual tags are embedded processed using a ResNet-101 [11] and learnable embedding layers respectively. The temporal features extracted from the product's release date are also embedded using learnable embedding layers.

Cross-attention RNN works, by default, in an autoregressive manner, therefore at each decoding step three different additive attention modules are applied. These modules allow the decoder hidden state to attend to the time series embedding, the image embedding and most importantly to the concatenated, multi-modal embedding. A residual learning approach [11] is applied to allow the network to scale bet-

Figure 4. A random sampling of users/ purchases. Personal styles do emerge: users 1 and 10 have no trousers in their logs, user 6 has bought almost short sleeves and no trousers, while user 7 seems to prefer long sleeves and several trousers; user 10 has a marked preference for light yellow-grayish colors.

ter with the number of hidden layers and also learn to ignore null contributions from the attention mechanism. After each decoding step, the GRU hidden state is updated based on the last processed input. After performing a prediction based on the GRU hidden state, we concatenate that prediction to the decoder input features and recurrently perform all these operation until the desired output length is reached. For extensive details on the architecture, we refer to [8].

The model is trained with a batch size of 128 and MSE (Mean Squared Error) loss function, using the Adafactor optimizer, on two NVIDIA RTX TITAN GPUs. During training, we apply dropout after each embedding module and also apply teacher forcing at random with a probability $p_{tf} = 0.5$.

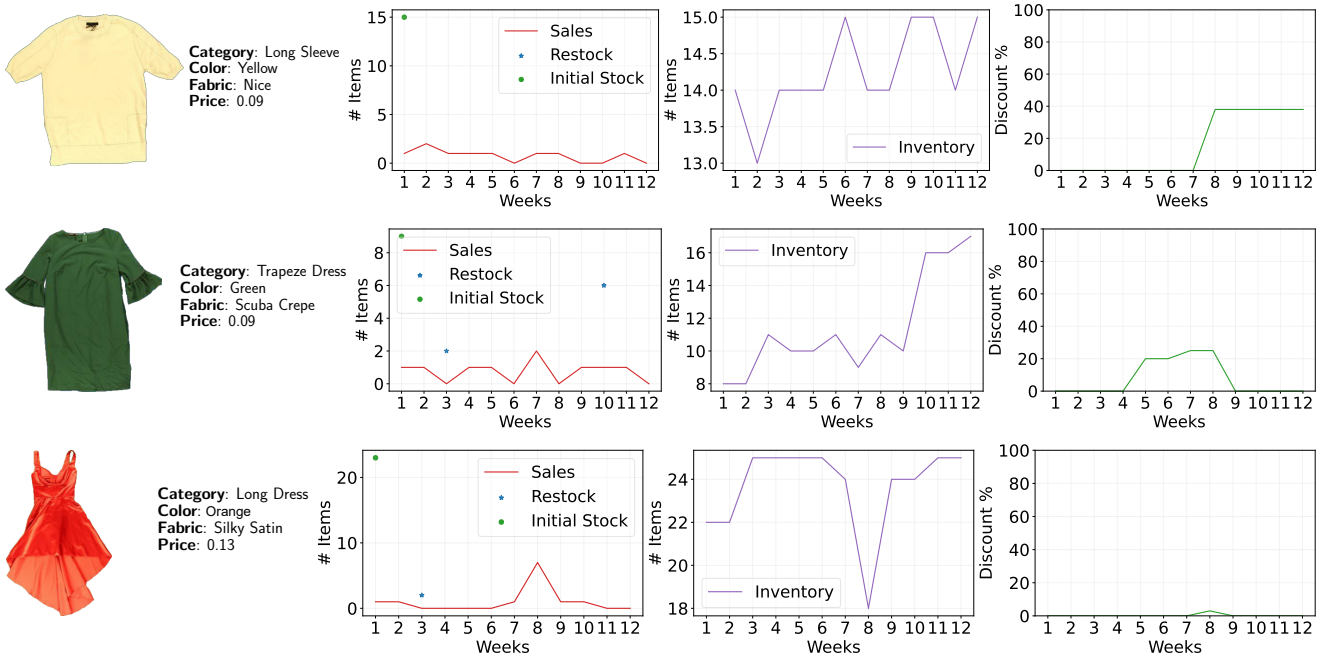

Figure 6. Examples of different products that have been sold in the shop SHOP3 during the SS17 season. The figure reports (from left to right): the product's image; textual tags; sales time series; restocking information; inventory time series; and discount time series.

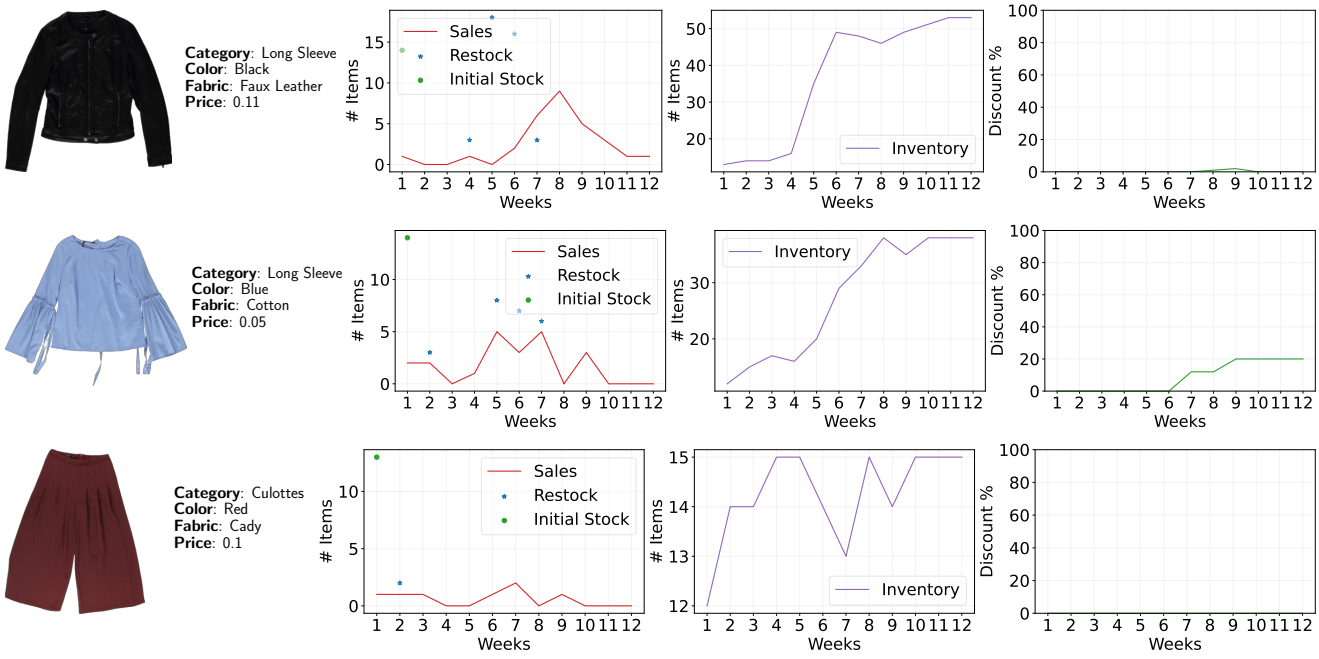

Figure 7. Examples of different products that have been sold in the shop SHOP28 during the AW17 season. The figure reports (from left to right): the product's image; textual tags; sales time series; restocking information; inventory time series; and discount time series.

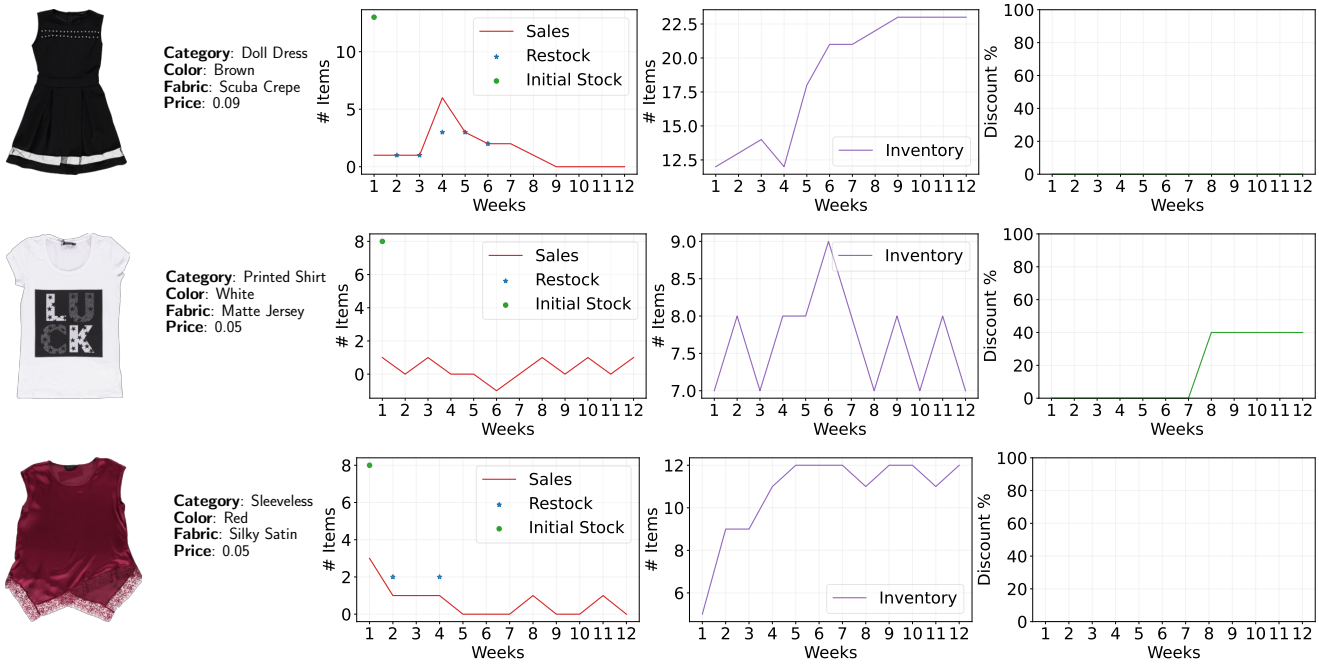

Figure 8. Examples of different products that have been sold in the shop SHOP51 during the SS18 season. The figure reports (from left to right): the product's image; textual tags; sales time series; restocking information; inventory time series; and discount time series.

References

[1] Ziad Al-Halah, Rainer Stiefelhagen, and Kristen Grauman. Fashion forward: Forecasting visual style in fashion. In *Proceedings of the IEEE international conference on computer vision*, pages 388–397, 2017. 2

[2] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. Jan. 2015. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015. 3

[3] Samaneh Beheshti-Kashi, Hamid Reza Karimi, Klaus-Dieter Thoben, Michael Lütjen, and Michael Teucke. A survey on retail sales forecasting and prediction in fashion markets. *Systems Science & Control Engineering*, 3(1):154–161, 2015. 2, 3

[4] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015. 2

[5] Robert Goodell Brown. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation, 2004. 2

[6] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion meets computer vision: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–41, 2021. 2

[7] Tsan-Ming Choi, Chi-Leung Hui, Na Liu, Sau-Fun Ng, and Yong Yu. Fast fashion sales forecasting with limited data and time. *Decision Support Systems*, 59:84–92, 2014. 2

[8] Vijay Ekambaram, Kushagra Manglik, Sumanta Mukherjee, Surya Shravan Kumar Sajja, Satyam Dwivedi, and Vikas Raykar. Attention based multi-modal new product sales time-series forecasting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3110–3118, 2020. 2, 3, 4, 5

[9] Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim, and Rasha Kashef. Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *applied sciences*, 10(21):7748, 2020. 3

[10] Javad Feizabadi. Machine learning demand forecasting and supply chain performance. *International Journal of Logistics Research and Applications*, 25(2):119–142, 2022. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[12] Yan Hu, Qimin Peng, Xiaohui Hu, and Rong Yang. Web service recommendation based on time series forecasting and collaborative filtering. In *2015 ieee international conference on web services*, pages 233–240. IEEE, 2015. 3

[13] Rajasekhar Kalla, Saravana Murikinjeri, and R Abbaiah. An improved demand forecasting with limited historical sales data. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. IEEE, 2020. 3

[14] Ling Lo, Chia-Lin Liu, Rong-An Lin, Bo Wu, Hong-Han Shuai, and Wen-Huang Cheng. Dressing for attention: Outfit based fashion popularity prediction. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3222–3226. IEEE, 2019. 2

[15] Ana LD Loureiro, Vera L Miguéis, and Lucas FM da Silva. Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, 114:81–93, 2018. 2

[16] Dennis Maaß, Marco Spruit, and Peter de Waal. Improving short-term demand forecasting for short-lifecycle consumer products with data mining techniques. *Decision analytics*, 1(1):1–17, 2014. 3

[17] Maria Elena Nenni, Luca Giustiniano, and Luca Pirolo. Demand forecasting in the fashion industry: a review. *International Journal of Engineering Business Management*, 5:37, 2013. 2, 3

[18] Alex D Papalexopoulos and Timothy C Hesterberg. A regression-based approach to short-term system load forecasting. *IEEE Transactions on power systems*, 5(4):1535–1547, 1990. 2

[19] Bohdan Pavlyshenko. Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, pages 255–258. IEEE, 2018. 2

[20] Bohdan M Pavlyshenko. Linear, machine learning and probabilistic approaches for time series analysis. In *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, pages 377–381. IEEE, 2016. 2

[21] Emmanuel Sirimal Silva, Hossein Hassani, Dag Øivind Madsen, and Liz Gee. Googling fashion: forecasting fashion consumer behaviour using google trends. *Social Sciences*, 8(4):111, 2019. 2, 3

[22] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 869–877, 2015. 2

[23] Geri Skenderi, Christian Joppi, Matteo Denitto, and Marco Cristani. Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends. *arXiv preprint arXiv:2109.09824*, 2021. 2, 3

[24] Donald Sull and Stefano Turconi. Fast fashion lessons. *Business Strategy Review*, 19(2):4–11, 2008. 2, 3

[25] Ian Malcolm Taplin. Global commodity chains and fast fashion: How the apparel industry continues to re-invent itself. *Competition & Change*, 18(3):246–264, 2014. 2

[26] S Thomassey et al. Intelligent demand forecasting systems for fast fashion. In *Information systems for the fashion and apparel industry*, pages 145–161. Elsevier, 2016. 3

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4

[28] Peter R Winters. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342, 1960. 2

[29] Kota Yamaguchi, Tamara L Berg, and Luis E Ortiz. Chic or social: Visual popularity analysis in online fashion networks. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 773–776, 2014. 2