Multimodal Quasi-AutoRegression: Forecasting the visual popularity of new fashion products

Stefanos-Iordanis Papadopoulos*, Christos Koutlis, Symeon Papadopoulos and Ioannis Kompatsiaris

Information Technology Institute, CERTH, Thessaloniki, Greece .

*Corresponding author(s). E-mail(s): stefpapad@iti.gr;
Contributing authors: ckoutlis@iti.gr; papadop@iti.gr; ikom@iti.gr;

Abstract

Estimating the preferences of consumers is of utmost importance for the fashion industry as appropriately leveraging this information can be beneficial in terms of profit. Trend detection in fashion is a challenging task due to the fast pace of change in the fashion industry. Moreover, forecasting the visual popularity of new garment designs is even more demanding due to lack of historical data. To this end, we propose MuQAR, a Multimodal Quasi-AutoRegressive deep learning architecture that combines two modules: (1) a multimodal multilayer perceptron processing categorical, visual and textual features of the product and (2) a quasi-autoregressive neural network modelling the "target" time series of the product's attributes along with the "exogenous" time series of all other attributes. We utilize computer vision, image classification and image captioning, for automatically extracting visual features and textual descriptions from the images of new products. Product design in fashion is initially expressed visually and these features represent the products' unique characteristics without interfering with the creative process of its designers by requiring additional inputs (e.g manually written texts). We employ the product's target attributes time series as a proxy of temporal popularity patterns, mitigating the lack of historical data, while exogenous time series help capture trends among interrelated attributes. We perform an extensive ablation analysis on two large scale image fashion datasets, Mallzee-P and SHIFT15m to assess the adequacy of MuQAR and also use the Amazon Reviews: Home and Kitchen dataset to assess generalisation to other domains. A comparative study on the VISUELLE dataset, shows that MuQAR is capable of competing and surpassing the domain's current state of the art by 4.65% and 4.8% in terms of WAPE and MAE respectively.

Keywords: Popularity Forecasting, Trend Detection, Quasi Autoregression, Multimodal learning, Computer Vision, Fashion

1 Introduction

Fashion is a dynamic domain, and fashion trends and styles are highly time-dependent. Systematic analysis of fashion trends is not only useful for consumers who want to be up-to-date with current trends, but is also vital for fashion designers, retailers and brands in order to optimize production cycles and design products that customers will find appealing when they hit the shelves. Moreover, it could potentially help mitigate the problem of unsold inventory in fashion that is caused by a mismatch between supply

and demand [8] and has significant environmental impact; with millions of tonnes of garments ending up in landfills or being burned every year [21].

At the same time, fashion is a primarily visually-driven domain. As a result, computer vision has successfully been utilized to assist fashion recommendations and trend forecasting [6]. Recent studies have utilized visual features - extracted by computer vision models - in order to identify fashion styles [2] or attributes [19] and then detect and analyse trends in fashion. However, such approaches are limited to detecting coarse-level trends and can not work for specific garment designs. They can forecast whether "chunky trainers" will be trending this season but all "chunky trainers" will receive the same popularity score. Specific visual differences in individual garments are not taken into consideration. Autoregressive (AR) neural networks have been used for forecasting the popularity of specific garments based on their past popularity [16]. However, new products by definition lack historical data, which renders the use of conventional AR networks impracticable. Few recent research works have addressed sales forecasting of new garments, by utilizing KNN-based (nearest neighbors) [7], auto-regressive networks with auxiliary features (images, fashion attributes and events) [8] or non-AR Transformers modelling images and fashion attributes along with the "target" time series of those attributes collected from Google Trends[1] [25]. However, fashion attributes are not always independent of each other. Trends in certain attributes may affect other interdependent attributes. If for example "warm minimalism" was trending in fashion, a series of light, neutral and pastel colors would show an increase in popularity while bold graphics and patterns would decrease.

The objective of this study is to accurately forecast the visual popularity of new garment designs that lack historical data. To this end, we propose MuQAR, a Multimodal Quasi-AutoRegressive neural network architecture and in Figure 1 we illustrate its high level workflow. MuQAR combines two modules: (1) a multimodal multilayer perceptron (FusionMLP) representing the visual, textual, categorical and

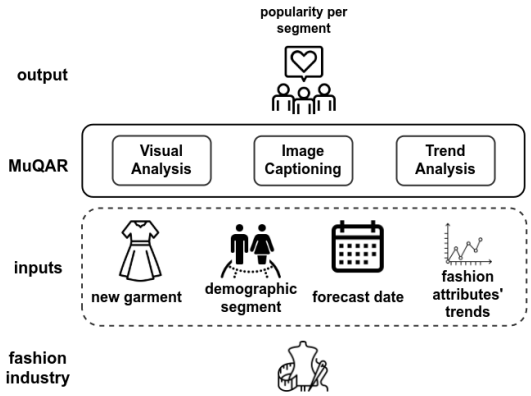


Fig. 1: High level workflow of MuQAR. The new garment's image and the trends of fashion attributes are analysed by the modules of MuQAR, predicting the garment's popularity for a given demographic segment and target date.

temporal aspects of a garment and (2) a quasi-autoregressive (QAR) neural network modelling the time series of the garment's fashion attributes (target time series) along with the time series of all other fashion attributes (exogenous time series). We expand the QAR framework by employing the logic of nonlinear autoregressive network with exogenous inputs (NARX). Our rationale is that modelling the target time series of the garment's attributes will work as an informative proxy of temporal patterns mitigating the lack of historical data while exogenous time series will help the model identify relations among fashion attributes.

The aim of this study is to provide fashion designers with real-time feedback for their new designs without interfering in their creative process. In fashion, design usually begins with sketching and visual prototyping - in 2D or 3D programs - expressing the silhouette, fit, colors and fabrics of the new garment. Previous works have relied on computer vision to extract relevant visual features from the garment's images along with its fashion attributes in order to forecast its popularity without interfering in the creative process (e.g by requiring textual descriptions of the garment). We expand upon this idea by utilizing image captioning (IC) for automatically extracting textual descriptions of the new garment that could provide richer descriptions and useful contextual information about the attributes of the

[1] https://trends.google.com

garment. For example, while an attribute detection model may recognise that a "varsity college jacket" has a graphic design, it is black and green and is made out of leather and jersey, an IC model could also describe the position of the graphic design (e.g across the chest) colors and fabrics (e.g black leather sleeves, green jersey body).

To the best of the authors' knowledge, no previous works have utilized NARX or IC in order to forecast the popularity of new products. The main contributions of our work are:

- A novel deep learning architecture that employs the logic of NARX models in QAR for forecasting the popularity of new products that lack historical data. We compare various QAR models, including: CNN, LSTM, ConvLSTM, Feedback-LSTM, Transformers and DA-RNN.
- Integration of image captioning in the multimodal module for capturing contextual and positional relations of the products' attributes.
- A new large-scale fashion dataset that includes popularity scores in relation to demographic groups allowing specialised forecasts for different market segments.
- An extensive ablation study on three datasets (two fashion and one home decoration) to assess the validity and generalisation of the proposed methodology. A comparative study on a fourth fashion-related dataset shows that our model surpasses the domain's state of the art by 4.65% and 4.8% in terms of WAPE and MAE respectively.

2 Related Work

Deep learning has been used for time series forecasting in numerous domains including climate modelling, biological sciences, medicine [15], music, meteorology, solar activity, finance [13] and among other industries, in fashion [4]. Additionally, researchers have been experimenting in recent years with the inclusion of visual information for detecting fashion trends and forecasting the popularity of garments or complete outfits [6].

Visual features extracted by computer vision models have been used to discover fashion styles [2] or fashion attributes [19] and neural networks to detect fashion trends. More recent works have examined how fashion trends from one city may influence other cities [1] or how knowledge enhanced neural networks can improve fashion trend detection [18]. The aforementioned studies extracted fashion styles or attributes from fashion imagery in order to forecast fashion trends. A significant limitation of these approaches however is that they can only work on a coarse level but not a finer level. Trends may show that "floral dresses will be trending next spring" but all new "floral dress" designs will receive the same score and cannot produce informed and specialized predictions for individual garments or outfits. Autoregressive (AR) networks have also been used to forecast the visual popularity for specific garments or outfits [16]. Nevertheless, new garment design, by definition, lack historical data and therefore conventional AR networks - that rely on past sequences to predict future outcomes - cannot be utilized.

While no works have directly addressed popularity forecasting of new garments, few recent research works have addressed sales forecasting of new garments, a closely related problem. Loureiro et al. [17] and Singh et al. [24] were two of the first works to utilize deep learning-based regression models for sales forecasting of new garments but did not utilize visual features. Craparotta et al. [7] proposed a KNN-based approach relying on an image similarity network connected to a siamese network for forecasting sales of new garments. Ekambaram et al. [8] utilized an AR multimodal RNN that combines visual and textual features with exogenous factors (holidays, discount season, events, etc.). Since new products do not have historical data, the authors bootstrap the network by adding two "start delimiters" and then utilize a "teacher forcing" method for next time steps. Skenderi et al. [25] criticised the reliance on purely AR networks for new product sale forecasting because of the compounding effect caused by first-step errors. Instead, they propose GTM-Transformer, a multimodal, non-AR Transformer that utilizes images, text and time series of the garment's attributes collected from Google Trends. Their work can be considered the first to utilize a version of quasi-autoregression using the "target" attributes time series and it was capable of outperforming both KNN-based approaches [7] and multimodal AR [8].

However, fashion attributes are not always independent of each other. Trends in some

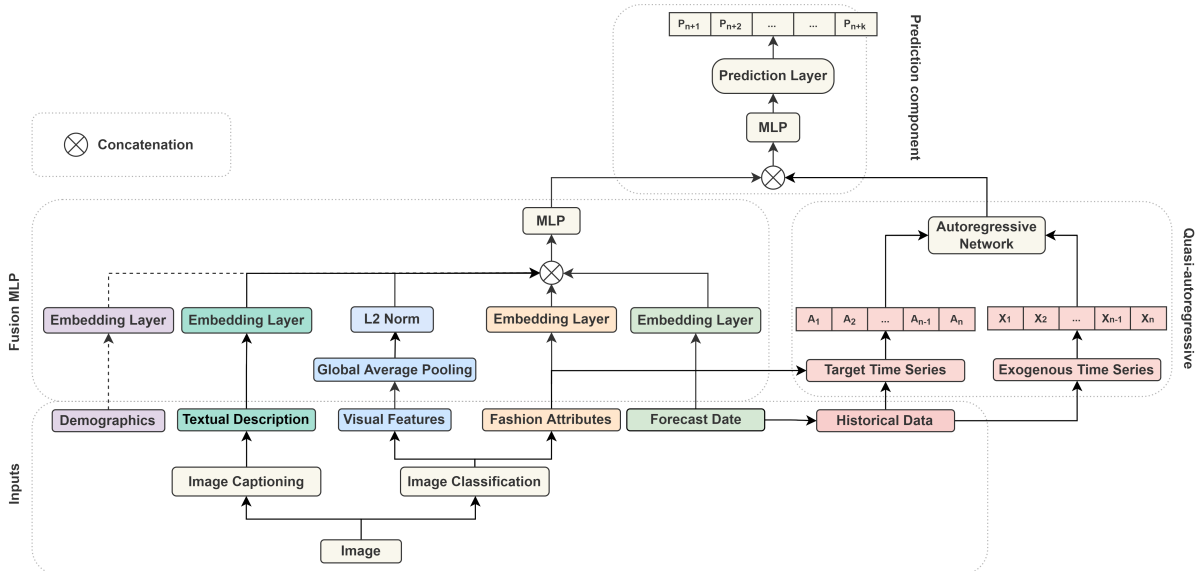


Fig. 2: MuQAR architecture. Intermittent arrows are optional and applicable only to datasets that provide demographic information.

attributes may positively or negatively affect others, e.g. complementary colors or matching categories. To this end, we employ the logic of nonlinear autoregressive network with exogenous inputs (NARX) [3] within QAR by integrating the "exogenous" fashion attributes along with the "target" attributes of a new garment in order to forecast its popularity.

Apart from the visual features of the garment, both Ekambaram et al. [8] and Skenderi et al. [25] used fashion attributes as textual information. Fashion attributes can be extracted from images by specialised classification models without requiring manual annotation from the fashion designers and can offer valuable information to the overall neural network. The advantage of such approaches is that, in real world settings, a forecasting model utilizing information about attributes would not interfere with the creative process of the designers. We expand upon this idea by utilizing an image captioning model (IC) for automatically extracting full textual descriptions from the images of new garment designs. Our hypothesis is that IC could create richer descriptions that capture useful contextual and positional information about the garment's attributes that will help improve the performance of multimodal forecasting models.

3 Methodology

In this study we attempt to forecast the visual popularity of new garment designs. Conventional autoregressive (AR) forecasting models can not be utilized since new products lack historical data. On the other hand, conventional regression models are not well equipped to detect temporal trends, which play an important role in the domain of fashion. To this end, we propose MuQAR, a multimodal quasi-autoregressive neural architecture that consists of two modules: FusionMLP and QAR as well as a final prediction component that combines them. Figure 2 illustrates the details of its architecture.

3.1 FusionMLP

The first module consists of a multimodal multilayer perceptron that processes the visual, textual, categorical and temporal features of a product. The visual features vector $F_v \in \mathbb{R}^V$, is extracted by the last convolutional layer of CNN-based networks. As will be discussed in Section 4.2, we use features extracted from pre-trained networks on ImageNet for SHIFT15m, Amazon Reviews and VISUELLE to ensure comparability. We also utilize a hierarchical deep learning network, fine tuned on a fashion image dataset [22]. After the

extraction, global average pooling and L2 normalisation is applied.

We utilize OFA [28] - a state-of-the-art IC model on COCO Captions[2] - for automatically extracting textual descriptions from fashion imagery. OFA does not specialise on fashion imagery but it has been trained on a large-scale e-commerce dataset that also included numerous fashion products. We manually examined hundreds of inference texts and deemed its predictions to be very accurate. We pre-process the extracted captions by lower-casing, removing punctuation and stop-words and then tokenizing them. We use a word embedding layer that produces the $F_w \in \mathbb{R}^W$ vector based on one integer index per token in relation to the whole vocabulary W. We apply IC on the Mallzee-P and VISUELLE datasets since they provide the garment images but not on SHIFT15m and Amazon Reviews that only provide pre-computed visual features.

The catecorical features vector $F_c \in \mathbb{R}^{c_p \cdot d_c}$, is the concatenation of c_p learnable embeddings of size d_c each corresponding to a fashion label assigned to product p, defined by:

$$F_c = [f_1 \mathrm{E}_c;\ f_2 \mathrm{E}_c;\ \ldots;\ f_{c_p} \mathrm{E}_c] \tag{1}$$

where [;] denotes concatenation, $E_c \in \mathbb{R}^{C \times d_c}$ is the embedding matrix for fashion labels, C is the total number of fashion labels and $f_i \in \mathbb{R}^C$ are one-hot encoding vectors with 1 at the index of the corresponding fashion label and zero elsewhere. The temporal feature vector $F_t \in \mathbb{R}^{4 \times d_t}$, is the concatenation of 4 learnable embeddings of size d_t corresponding to the day, week, month and season of the target date, defined by:

$$F_t = [d\mathrm{E}_d;\ w\mathrm{E}_w;\ m\mathrm{E}_m;\ s\mathrm{E}_s] \tag{2}$$

where $E_d \in \mathbb{R}^{366 \times d_t}$, $E_w \in \mathbb{R}^{52 \times d_t}$, $E_m \in \mathbb{R}^{12 \times d_t}$, $E_s \in \mathbb{R}^{4 \times d_t}$ are embedding matrices for day, week, month and season of the year respectively and $d \in \mathbb{R}^{366}$ (leap year provision), $w \in \mathbb{R}^{52}$, $m \in \mathbb{R}^{12}$, $s \in \mathbb{R}^4$ are the corresponding one-hot encoding vectors. For the demographic group input, that is optional and considered only in one dataset here, we also consider a learnable embedding $F_g = g\mathrm{E}_g \in \mathbb{R}^{d_g}$, accordingly. Finally, a standard MLP network with n_{mlp} dense layers of u_{mlp} relu activated units processes the concatenation of all features resulting in $F_F \in \mathbb{R}^f$:

$$F_F = \mathrm{MLP}([F_v;\ F_w;\ F_c;\ F_t;\ F_g]) \tag{3}$$

3.2 QAR

The second module utilizes the product attributes' time series ("target") $\{A_t\}$ along with all other attributes time series ("exogenous") $\{X_t\}$ as input in order to predict the product popularity time series $\{P_t\}$. More precisely, we feed QAR module with two matrices $\mathbf{A} = \{A_1, \ldots, A_n\} \in \mathbb{R}^{n \times c_p}$ that contains n time steps prior to the forecast date for c_p target fashion labels assigned to product p and $\mathbf{X} = \{X_1, \ldots, X_n\} \in \mathbb{R}^{n \times c_x}$ for all "exogenous" fashion labels c_x. X includes the time series for all available fashion attributes within the time period, but we set the target attributes of c_p to zero in c_x to avoid information leakage.

The proposed methodology, MuQAR, is a modular architecture meaning that it can integrate any AR network to the QAR module. This allows for identifying the optimal AR network for a given task. In this study we experimented with multiple AR architectures that have been used for time series forecasting, namely: Long Short Term Memory network (LSTM) [10], (2) Feedback LSTM (F-LSTM) [9], (3) Convolutional Neural Network (CNN) [30], (4) Convolutional LSTM (ConvLSTM) [29], (5) Transformer [27] for experiments that only utilize $\{A_t\}$ and (6) Dual-Stage Attention-Based Recurrent Neural Network (DA - RNN) [23] adapted for multivariate time series, (7) Convolutional LSTM with exogenous time series (ConvLSTM + X) inspired by the encoder proposed in [5] for experiments that utilize both $\{A_t\}$ and $\{X_t\}$. DA-RNN is a dual-stage architecture that first processes $\{X_t\}$ and then $\{A_t\}$ while ConvLSTM+X processes $\{A_t\}$ and $\{X_t\}$ in parallel - with two separate ConvLSTM neural networks - and then concatenates the resulting representation vectors. After processing the input time series, QAR produces a vector representation $F_Q \in \mathbb{R}^q$ pertinent to the forecast.

3.3 Prediction component

The concatenated vector $F = [F_F;\ F_Q] \in \mathbb{R}^{f+q}$ is further processed by another dense layer on top of which a linear layer forecasts the product's next

[2] https://paperswithcode.com/sota/image-captioning-on-coco-captions

k popularity time steps $\{P_{n+1}, \ldots, P_{n+k}\}$, as can be seen in Figure 2.

4 Experimental Setup

4.1 Evaluation Protocol

To correctly assess the task of forecasting the popularity of new products, we propose an evaluation protocol where models are trained on *established products* and evaluated on *new products*. We consider as established those products that have multiple records in the dataset in different times, while new products are those that make only a single appearance in the dataset and the model has not previously encountered. We therefore split the data into established products, which are used as the training set, and new products, which are split in half for the validation and testing sets. We apply this protocol on Mallzee-P, SHIFT15m and Amazon Reviews datasets but for VISUELLE we follow the experimental protocol described in [25].

For evaluation, we used multiple evaluation metrics. For regression tasks we used the Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC) and Binary Accuracy (BA). For classification tasks we used Accuracy and the Area under the ROC Curve (AUC). We selected the best performance per model based on TOPSIS, a multi-criteria decision analysis method [11].

We perform an extensive grid-search for tuning the hyper parameters of FusionMLP and the QAR networks. We integrate the best performing QAR network with FusionMLP's best performing hyper-parameter combinations on each dataset separately to create MuQAR. QAR models are trained with weekly aggregated time series using 12 weeks as input and 1 week as output.

4.2 Datasets

4.2.1 VISUELLE

VISUELLE[3] is a public fashion image dataset that contains 12 week long sales sequences for 5577 garments spanning from October 2016 to December 2019 [25]. For each garment, it provides an image, textual information (categorical labels related to fashion categories, fabrics, colors) and time series related to the categorical labels collected from Google Trends. The dataset is sorted by date and split into 5080 garments for training and 497 for evaluation.

4.2.2 SHIFT15m

Due to the fact that VISUELLE is a relatively small scale dataset we also experiment with larger-scale image fashion datasets. SHIFT15m[4] is a public, large scale and multi-objective fashion dataset that consists of 15,218,721 outfits posted by users in a social network platform among 193,574 users and 2,335,598 garments between 2010 and 2020 [12]. The dataset provides the user ID, the number of likes that the outfit has received, the date it was published and the items that constitute the outfit, including the item IDs and two types of fashion categories (comprising 7 and 43 unique categories respectively). The outfits' images are not available but SHIFT15m provides the visual features extracted by a VGG network pre-trained on the ImageNet dataset.

We re-purpose SHIFT15m for new garment popularity forecasting. We remap the initial outfit-level onto the garment-level by splitting each outfit into the individual garments that make it up and by defining the number of likes as the target variable. We assume that each garment has an equal contribution to the overall popularity of the outfit. The number of likes follows a skewed distribution. We therefore normalise it with the logarithmic transform, namely $log(\text{likes} + 1)$ and scale it within the range of $(0, 1)$ with the use of min-max scaling. To create the time series used in QAR, we simply compute the weekly mean for each fashion category. The time series exhibit 10.72% sparsity, so we apply linear interpolation to fill the missing values. The dataset is split into 14,342,771 samples for *established products* (training set) and 875,950 samples for *new products* (validation and testing sets).

4.2.3 Mallzee-P

One limitation of SHIFT15m is that it provides no user-side information. Forecasting networks trained on SHIFT15m would learn to forecast the general popularity of a new garment design but the predictions would be identical regardless of age

[3] https://github.com/HumaticsLAB/GTM-Transformer

[4] https://github.com/st-tech/zozo-shift15m/tree/main/benchmarks

Fig. 3: Monthly aggregated time series for fashion categories from the Mallzee-P dataset.

group or gender. For example, a newly designed "floral dress" would have the same popularity score for "women, age: 18-25" and "men, age: 40-50". To alleviate this issue and perform more targeted and useful forecasts for different segments of the market we collect the Mallzee-P dataset.

Our partner Mallzee collected 5,412,193 records from their databases between 14 demographic groups and 571,333 unique products. The selected demographics consist of two gender groups (men, women) and 7 age-groups including 0-18, 18-25, 25-30, 30-35, 35-45, 45-55, ¿55. The data span 1081 days between 2017-10-16 and 2020-09-30. We extract the visual features from a three-stage convolutional neural network consisting of an object detector, a garment category classifier and a fine-grained attribute detector, proposed in [22][5]. We use the object-detector to identify individual garments in an image and the other two modules to extract the predicted fashion labels and the visual features from their last convolutional layers. The Mallzee dataset is classified into 22 garment categories such as blouses, dresses, jackets etc. and 109 fine-grained attributes including patterns (e.g. checked, quilted), prints (e.g. floral, graphic) and styles (e.g. bomber jacket, puffer jacket). The dataset is split into 5,320,076 samples for *established products* (training set) and 92,117 samples for *new products* (validation and testing sets).

As the target variable, for each product p we consider the popularity metric P (Eq.4a) that expresses both *likability* L (Eq.4b) and *reachability* R (Eq.4c), given the set A of the product's

[5]The computer vision neural network was trained on an image dataset collected by project partner Mallzee that consists of images with category and attributes annotations. Mallzee-P is a different dataset that includes images with demographic-based popularity scores.

attributes, the target demographic group G and the target time t:

$$P(p \mid A, G, t) = L(p \mid G, t) \cdot R(A, G, s) \tag{4a}$$

$$L(p \mid G, t) = \frac{I(\oplus, G, p, t)}{I(\oplus, G, p, t) + I(\ominus, G, p, t)} \tag{4b}$$

$$R(A, G, s) = \prod_{a \in A} \frac{\mid \{u \in G \mid (u \leftrightarrow a) \wedge s\} \mid}{\mid \{u \in G \mid s\} \mid} \tag{4c}$$

where s is the year's season that day t belongs to, $I(\oplus, G, p, t)$ and $I(\ominus, G, p, t)$ denote the number of positive and negative interactions between G and p at day t respectively, u denotes user, $\mid \cdot \mid$ denotes set cardinality and $u \leftrightarrow a$ denotes positive interaction of user u with a product having attribute a. Hence, *likability* is the probability that demographic group G likes product p at time t, while *reachability* is the probability that demographic group G interacts with the set of attributes A at season s. The reason for incorporating reachability to calculate popularity is that positive interaction of small fractions of demographic groups with unusual attributes results in unexpected high likability. For instance, we found dresses, jumpsuits and heels in the top categories for men 18-25 based only on likability. The incorporation of reachability not only mitigated this issue but gave reasonable seasonal patterns to all garment categories as well. To create the time series used in QAR, we compute the weekly mean popularity for each of the 22 categories and 109 attributes which exhibit 1.13% and 8.49% sparsity respectively. We apply linear interpolation to fill the missing values. Time series of certain fashion classes can be seen in Figure 3 where we can observe clear seasonal patterns.

4.2.4 Amazon Reviews: Home and Kitchen

Fashion is the central focus of our work but we also deem important to examine the generalisation of MuQAR to other domains. To this end, we utilize the Amazon Reviews dataset[6] and specifically the "Home and Kitchen" subset [20]. Our selection criteria (cr.) for the dataset are: (1) relate to a primarily visually-driven domain, (2) provide a popularity metric, (3) provide images or extracted visual features, (4) provide categories that produce dense time series.

We select the "Home and Kitchen" subset since it mostly contains products relating to furniture, decorative items, artwork posters and kitchen appliances. We consider that visual appearances play a very important role in driving customer choices in this domain similarly to fashion (cr. 1). We define the 'star ratings' as the target variable (cr. 2). The dataset provides the visual features extracted from the products' images by a convolutional neural network pretrained on ImageNet (cr. 3).

The dataset spans years 1999 to 2014 and comes with 965 unique categories. However categories contain duplicates (e.g "sheet" and "sheets") and rare items (e.g "charcoal drawings" and "crayon drawings" only make a single appearance in the dataset). Aggregating the weekly time series for the 965 categories results in 64.34% sparsity. Even after filtering out the products before 2009 (a period with higher sparsity rates), results in 33.34% sparsity. In order to mitigate this issue, we use the K-Means algorithm to cluster categories based on their TF-IDF textual representation into K=300 clusters. The weekly time series sparsity is reduced to 36.91% for the whole dataset and to only 2.13% for the filtered subset. (cr. 4). Filtering out the products before 2009 reduced the total size of the dataset from 3,261,846 to 3,002,786 which we consider a sensible trade-off between data loss and reduced time series sparsity. The dataset is split into 2,840,178 samples for *established products* (training set) and 149,414 samples for *new products* (validation and testing sets).

4.3 Implementation Details

We perform a grid search for hyper-parameter optimization for FusionMLP and QAR modules. For FusionMLP we define $n_{mlp=}$ 3, 4, 5 or 6 fully connected layers with progressively narrower units u_{mlp} = (2048, 1024, 512, 256, 128) (2048 units layer used twice in the case of 6 layers only) and an embedding space of $d_c = d_t = d_g = 8$, 16 or 32 dimensions. For the QAR networks we define a hyper-parameter grid search as shown in Table 1. For each QAR network we define the number

[6] https://jmcauley.ucsd.edu/data/amazon/

Table 1: Grid search for hyper-parameter optimization of the QAR networks in terms of the number of layers $n_{\{Layer\}}$ and their units $u_{\{Layer\}}$.

Model	Layer	$n_{\{Layer\}}$	$u_{\{Layer\}}$
LSTM	lstm	1, 2, 3	(512, 256, 128)
CNN	cnn	1, 2, 3	(512, 256, 128)
DA-RNN	lstm	2	64 or 128
ConvLSTM	cnn lstm	1, 2, 3 1, 2, 3	(512, 256, 128) (512, 256, 128)
F-LSTM	lstm mlp	1, 2, 3 0,1	(512, 256, 128) 256 or 512
Transformer	block head	1, 2, 3 2 or 4	128, 256 or 512

of its component layers and their units. Values in parenthesis indicate experiments with progressively narrower units. After each LSTM, CNN or MLP layer we add a dropout layer of 10% probability in order to reduce overfitting with the exception of Transformers which we define at 20% probability.

For training we consider the Adam optimiser, cyclical learning rate [26] (initial learning rate: 1e-4, max learning rate: 1e-2, step size: 2, gamma: 0.1), the mean squared error (MSE) loss function for regression tasks (VISUELLE, SHIFT15m, Mallzee-P) and the categorical cross entropy for classification tasks (Amazon Reviews: Home and Kitchen). We use a batch size of 1024 for Mallzee-P and Amazon datasets, 8192 for SHIFT15m since it is a larger dataset and we wanted to exploit parallelization and 16 for VISUELLE since it is a very small dataset.

5 Results

5.1 Ablation Analysis

Table 2 presents the ablation analysis of MuQAR on Mallzee-P, SHIFT15m and Amazon Reviews datasets.

QAR: By performing the hyper parameter optimisation grid search on the Mallzee-P dataset we identify the following best performing combinations, LSTM: $u_{lstm} = (512)$ units, CNN: $u_{cnn} = (512, 256, 128)$, DA-RNN: $u_{lstm} = (128)$, ConvLSTM: $u_{cnn} = (512, 256, 128)$, $u_{lstm} = (512, 256)$ and $u_{mlp} = (128)$, F-LSTM: $u_{lstm} = (256)$ and $u_{mlp} = (512)$, Transformer: $n_{block} = 1$, $n_{head} = 2$ with 256 units, ConvLSTM+X: we use the same hyper parameters as ConvLSTM. We apply the same hyper-parameters on the rest of the datasets. The CNN, ConvLSTM and Transformer QAR networks yielded the best performance for Mallzee-P, SHIFT15m and Amazon Reviews datasets respectively when only using [A], the target attribute time series with 12 previous weeks as input and 1 as output. We integrate these specific QAR models for each of the three datasets in the MuQAR experiments. We can observe that there is not a single QAR network that consistently performs better. Experimentation is required in order to identify the most appropriate QAR network for specific tasks and datasets.

When QAR utilizes both [A + X] its performance decreases on Mallzee-P and SHIFT15m; compared to the best QAR with [A]. However, we observe an impressive improvement on the Amazon dataset with +18% and +10% increases in terms of Accuracy and AUC respectively. This may be attributed to the fact that Amazon initially had 965 categories which we clustered into 300 with the use of K-Means - a lot more than in Mallzee-P and SHIFT15m - some of which may be quite noisy. Thus, feeding the exogenous time series [X] may help QAR discern the informative from the noisy time series. Overall, we observe that ConvLSTM+X tends to outperform DA-RNN in 6 out of 8 cases. We therefore integrate ConvLSTM+X in the final MuQAR experiments.

FusionMLP: an embedding space of $d_c = d_t = d_g = 32$ dimensions, $n_{mlp} = 4$ with $u_{mlp} = (2048, 1024, 512, 256)$ and a dropout rate of 10% performed best on the three datasets. FusionMLP performed significantly better than the QAR networks on the Mallzee-P dataset but not as good on SHIFT15m and Amazon Reviews. The visual features [I] in the Mallzee-P data are extracted by a fine-tuned network on fashion imagery, while the other two datasets utilized networks pre-trained on ImageNet and are therefore less specialised. This fact may have affected the results since, presumably, the quality and specialisation of the extracted visual features plays a crucial role in the task. When image captions [C] are added in FusionMLP there is only a negligible improvement in terms of MAE but a small decrease in terms of PCC and Accuracy. This issue may be due to

Table 2: Popularity forecasting models trained on 'established' and evaluated on 'new' garments for three datasets: Mallzee-P (MLZ-P), SHIFT15m and Amazon Reviews: Home and Kitchen. Features used: Images [I], target attributes time series [A], exogenous attributes time series [X] and image captions [C]. [A] and [X] have 12 weeks-long time series as input. The models forecast the next week.
Bold denotes the best overall performance per metric and dataset. Underline denotes the best performing QAR network per dataset; which are used in the final MuQAR models.
**[C] are only available on MLZ; they are ignored on Amazon and SHIFT15m, [I+A+X] are used instead.*

Input	Model	MAE(↓)		PCC(↑)		Accuracy(↑)			AUC(↑)
		MLZ-P	SHIFT15m	MLZ-P	SHIFT15m	MLZ-P	SHIFT15m	Amazon	Amazon
[A]	LR	0.1878	0.1162	0.2439	0.3177	63.10	59.58	48.52	65.41
	CNN	0.1611	0.1148	0.5379	0.3406	70.99	61.51	47.18	69.34
	LSTM	0.1656	0.1150	0.5109	0.3371	69.67	61.42	45.58	67.54
	F-LSTM	0.1809	0.1149	0.3395	0.3376	64.62	61.43	44.95	67.89
	Transformer	0.1842	0.1149	0.3071	0.3398	63.67	61.28	51.10	71.29
	ConvLSTM	0.1641	0.1147	0.5225	0.3411	69.98	61.58	46.58	68.59
[A+X]	ConvLSTM+X	0.1686	0.1185	0.4913	0.2191	68.86	59.50	60.61	78.95
	DA-RNN	0.1863	0.1187	0.2652	0.2050	64.39	59.55	60.61	78.90
[I]	LR	0.1599	0.1186	0.5314	0.1940	71.86	57.93	41.46	68.18
	FusionMLP	0.1074	0.1148	0.7893	0.2811	81.52	60.89	46.96	71.69
[I+C]	FusionMLP	0.1073	-	0.7879	-	81.30	-	-	-
[I+A]	MuQAR	0.0949	**0.1100**	0.8362	**0.3934**	83.41	**63.57**	51.51	74.24
[I+A+C]	MuQAR	0.0928	-	0.8474	-	83.69	-	-	-
[I+A+X+C]	MuQAR	**0.0911**	0.1118*	**0.8484**	0.3448*	**84.26**	62.30*	**60.63***	**80.40***

the hyper-parameter tuning being done on the [I] features and not fine-tuned for [I+C].

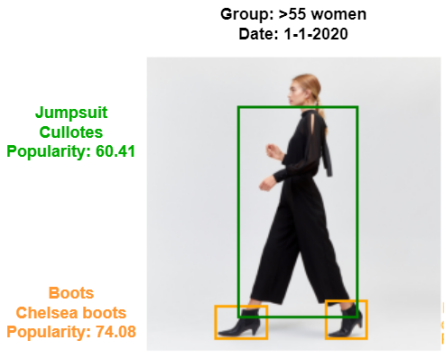
MuQAR: we can observe that MuQAR using [I+A] is capable of consistently surpassing all QAR networks and FusionMLP on the three datasets. By combining the visual features [I] with the target time series [A], MuQAR is able to improve on the task of forecasting the popularity of new products while retaining high performance even when less specialised visual features were used. Furthermore, MuQAR with [I+A+C] performs better than FusionMLP with [I+C] and MuQAR with [I+A] on the Mallzee-P dataset. Adding [A] to [I+C] results in a 13.5% percentage decrease in terms of MAE while adding [C] to [I+A] results in a smaller 2.2% percentage decrease; showing that [A] has a more important contributions to MuQAR's performance on the Mallzee-P dataset. Finally, adding exogenous time series [X] further improves the performance of MuQAR for Mallzee-P and Amazon datasets but not on SHIFT15m. The latter most likely occurs due to SHIFT15m's time series being nearly identical; meaning that all categories exhibit the same popularity fluctuation and seasonal patterns. When SHIFT15m's exogenous time series are utilized, the neural network receives near identical data points which increases the network's complexity without offering any useful information and causes the model's decrease in performance. We can conclude that the quality of the popularity time series is of high importance when training QAR models as is the quality of the visual features when training FusionMLP. While [A+X] did not consistently improve the performance of QAR, neither did [C] significantly improve FusionMLP, when combined and integrated within the MuQAR architecture they can show significant improvements.

In Figure 4 we present an inference sample predicted by MuQAR trained on the Mallzee-P dataset. We use the same image, depicting a jumpsuit and a pair of chelsea boots and MuQAR performs predictions for different demographic groups. We can observe that the popularity of all garments increases in June for women over 55 compared with January (Fig. 4a and 4b). Moreover, the outfit is less popular with younger women (Fig. 4c) and very unpopular with men (Fig. 4d). In Figures 4e and 4f we observe that a sweater

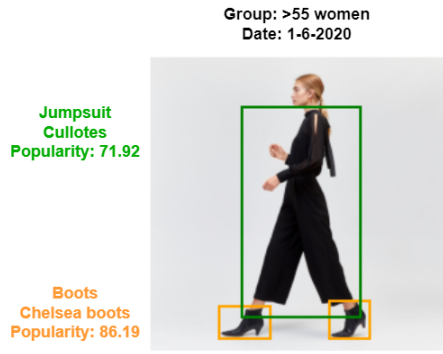
Table 3: Comparative analysis between MuQAR and its modules against state of the art networks on the VISUELLE dataset using 52 week-long time series as input from Google Trends and forecasting the next 6. Features used: [T]ext, [I]mage, target [A]ttribute time series from Google trends, e[X]ogenous time series from google trends and image [C]aptions. Underline denotes the best performing network per input type. **Bold** denotes the best overall performance.

Method	**Input**	**IN:52, OUT:6**	
		WAPE(↓)	MAE(↓)
GTM-Transformer [25]	[T]	62.6	34.2
Attribute KNN [8]		59.8	32.7
FusionMLP		55.15	30.12
Image KNN [8]	[I]	62.2	34.0
GTM-Transformer [25]		56.4	30.8
FusionMLP		54.59	29.82
Transformer	[A]	62.5	34.1
LSTM		58.7	32.0
ConvLSTM		58.6	32.0
GTM-Transformer [25]		58.2	31.8
F-LSTM		58.0	31.7
CNN		57.4	31.4
ConvLSTM+X	[A + X]	55.73	30.44
DA-RNN		58.05	31.71
Attribute + Image KNN [8]	[T + I]	61.3	33.5
Cross-Attention RNN [8]		59.5	32.3
GTM-Transformer [25]		56.7	30.9
FusionMLP		54.11	29.56
FusionMLP	[T + I + C]	53.50	29.22
GTM-Transformer AR [25]	[T + I + A]	59.6	32.5
Cross-Attention RNN+A [8]		59.0	32.1
GTM-Transformer [25]		55.2	30.2
MuQAR w/ Transformer		54.87	29.97
MuQAR w/ F-LSTM		54.37	29.7
MuQAR w/LSTM		54.3	29.66
MuQAR w/CNN		53.9	29.44
MuQAR w/ ConvLSTM		53.61	29.28
MuQAR w/ DA-RNN	[T + I + A + X + C]	54.43	29.73
MuQAR w/ ConvLSTM+X		**52.63**	**28.75**

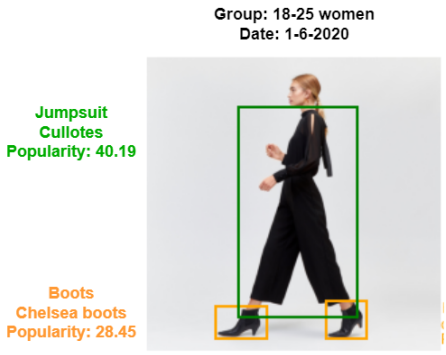
with a "playful" graphic design is more popular with younger (mostly teenagers) than older women. Finally, a minimal white "boxy t-shirt" shows an increase in popularity from January (Fig. 4g) to June (Fig. 4g) presumably due to the seasonal change and by extension the warmer weather - while still remaining relatively unpopular for young teenage boys. MuQAR seems to have learned both seasonal trends and the average preferences of different demographic groups.



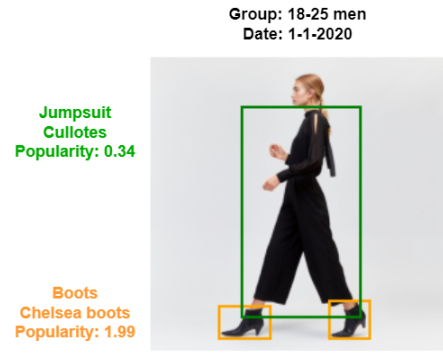
(a)



(b)



(c)



(d)

(e)

(f)

(g)

(h)

Fig. 4: Inference samples by the MuQAR on Mallzee-P data for different demographic groups and dates.

5.2 Comparative Analysis

Table 3 presents the comparison between MuQAR and its modules against state of the art networks proposed by [8] and [25] on the VISUELLE dataset. We do not fine tune the hyperparameters of MuQAR and its modules on the VISUELLE dataset. Rather, we use the same values as described in 5.1 with the exception of increasing the dropout rate to 0.3 to avoid overfitting since VISUELLE is a significantly smaller dataset.

Regarding QAR using [A], CNN and F-LSTM were able to surpass both GTM-Transformer and our Transformer. This indicates that it is advisable to also experiment with "simpler" neural network architectures and not immediately leaping to more complicated networks. Moreover, ConvLSTM+X utilizing both target and exogenous time series [A+X] surpasses all QAR models using [A]. FusionMLP utilizing [T+I], or solely [T] or [I] can not only outperform its similar-input competitors but also the GTM-Transformer using [T+I+A]. On top of that, adding image captions [C] improves the performance of FusionMLP. In the ablation study we noted that adding image captions had only marginal improvement on FusionMLP in terms of MAE for the Mallzee-P dataset. In contrast, combining [T+I+C] results in a notable improvement both in terms of WAPE and MAE on the VISUELLE dataset. This can be explained by Mallzee-P having fine-tuned visual features while VISUELLE having visual features extracted from an ImageNet pre-trained model that do not specialize on fashion. The former captures deeper and finer-grained aspects of fashion imagery while the latter represents more general visual aspects. Therefore, a model relying on pre-trained visual features may have relatively more to gain from image captions since they can provide additional information about the garment's visual characteristics and attributes.

We also observe that MuQAR with any of the QAR networks is capable of surpassing GTM-Transformer (non-AR), Cross Attention RNN+A and GTM-Transformer AR (an autoregressive variant) when using [T+I+A]. Finally, the proposed MuQAR w/ ConvLSTM+X employing all features [T+I+A+X+C] is capable of surpassing all other models, setting a new state-of-the-art on VISUELLE with 4.65% and 4.8% improvements in terms of WAPE and MAE respectively. These results further prove the validity of our two main proposals, the use of image captioning and the inclusion of exogenous attributes time series within the MuQAR architecture when forecasting the visual popularity of new garment designs.

6 Conclusions

In this study we propose MuQAR, a Multimodal Quasi Auto-Regressive deep learning architecture, for forecasting the popularity of new products that lack historical data. The proposed architecture consists of two modules: (1) a multimodal multilayer perceptron (FusionMLP) representing visual, textual, categorical and temporal aspects of a product and (2) a quasi-autoregressive (QAR) neural network modelling the time series of the product's attributes along with all other attributes time series. For FusionMLP, we extract the visual and categorical features (fashion attributes) from a computer vision model representing the unique characteristics of new products along with rich textual descriptions extracted from an image captioning model. In QAR, the time series of its attributes provide a proxy of temporal patterns for the lack of historical data while the exogenous time series are used to identify relations among the target and all other attributes.

Our focus is centered around the fashion industry and new garment designs. We experiment with three fashion image datasets: Mallzee-P, SHIFT15m and VISUELLE. We also experiment with the Amazon Reviews: Home and Kitchen dataset to examine the generalisation of the proposed architecture. MuQAR proved capable of competing and surpassing the domain's current state of the art by 4.65% in terms of WAPE and 4.8% in terms of MAE on the VISUELLE benchmark dataset. Furthermore, our extensive ablation and comparative analysis validated the potential of the proposed methodologies, namely utilizing image captions, target and exogenous time series, to enhance the performance of multimodal popularity forecasting models. Target time series consistently achieved improvements in all four datasets. Similarly, exogenous time series offered further accuracy improvement; provided that the time series were of high quality and diverse enough. Last, image captions also contributed to MuQAR's performance, benefitting

models relying on ImageNet-pretrained visual features, instead of fine-tuned features, to a greater extent. In this study, we experiment with two QAR models that utilize the exogenous time series, namely DA-RNN and ConvLSTM+X. It would be interesting for future research to examine how other models perform within the QAR module such as Lavarnet [13], MTNet [5], LSTNet [14] or the Informer [31]. Moreover, we have only experimented with image datasets but the proposed architecture MuQAR could be adapted and applied to other domains with different types of data. For example audio feature combined with time series of musical genres could be used to forecast the popularity of new tracks or albums. In future work we plan on examining how visual-temporal features extracted from MuQAR can facilitate recommendation systems and especially in mitigating the cold start problem of new items.

7 Declarations

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.
Funding. This work is partially funded by the Horizon 2020 European project "eTryOn - virtual try-ons of garments enabling novel human fashion interactions" under grant agreement no. 951908.
Data availability. Amazon Reviews, VISUELLE and SHIFT15m are publicly available datasets and have been properly referenced in the text. The Mallzee-P dataset is proprietary and has been provided to us by our partner Mallzee purely for research purposes.

References

[1] Al-Halah Z, Grauman K (2020) From paris to berlin: Discovering fashion style influences around the world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10,136–10,145

[2] Al-Halah Z, Stiefelhagen R, Grauman K (2017) Fashion forward: Forecasting visual style in fashion. In: Proceedings of the IEEE international conference on computer vision, pp 388–397

[3] Billings SA (2013) Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains. John Wiley & Sons

[4] Chang AA, Ramadhan JF, Adnan ZKS, et al (2021) Fashion trend forecasting using machine learning techniques: A review. In: Proceedings of the Computational Methods in Systems and Software, Springer, pp 34–44, https://doi.org/10.1007/978-3-030-90321-3_5

[5] Chang YY, Sun FY, Wu YH, et al (2018) A memory-network based solution for multivariate time-series forecasting. arXiv preprint arXiv:180902105

[6] Cheng WH, Song S, Chen CY, et al (2021) Fashion meets computer vision: A survey. ACM Computing Surveys (CSUR) 54(4):1–41. https://doi.org/10.1145/3447239

[7] Craparotta G, Thomassey S, Biolatti A (2019) A siamese neural network application for sales forecasting of new fashion products using heterogeneous data. International Journal of Computational Intelligence Systems 12(2):1537–1546. https://doi.org/10.2991/ijcis.d.191122.002

[8] Ekambaram V, Manglik K, Mukherjee S, et al (2020) Attention based multi-modal new product sales time-series forecasting. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 3110–3118, https://doi.org/10.1145/3394486.3403362

[9] Graves A (2013) Generating sequences with recurrent neural networks. arXiv preprint arXiv:13080850

[10] Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735–1780

[11] Hwang CL, Yoon K (1981) Methods for multiple attribute decision making. In: Multiple attribute decision making. Springer, pp 58–191

[12] Kimura M, Nakamura T, Saito Y (2021) Shift15m: Multiobjective large-scale fashion dataset with distributional shifts. arXiv preprint arXiv:210812992 https://doi.org/10.48550/arXiv.2108.12992

[13] Koutlis C, Papadopoulos S, Schinas M, et al (2020) Lavarnet: Neural network modeling of causal variable relationships for multivariate time series forecasting. Applied Soft Computing 96:106,685. https://doi.org/10.1016/j.asoc.2020.106685

[14] Lai G, Chang WC, Yang Y, et al (2018) Modeling long-and short-term temporal patterns with deep neural networks. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp 95–104, https://doi.org/10.1145/3209978.3210006

[15] Lim B, Zohren S (2021) Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A 379(2194):20200,209. https://doi.org/10.1098/rsta.2020.0209

[16] Lo L, Liu CL, Lin RA, et al (2019) Dressing for attention: Outfit based fashion popularity prediction. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3222–3226, https://doi.org/10.1109/ICIP.2019.8803461

[17] Loureiro AL, Miguéis VL, da Silva LF (2018) Exploring the use of deep neural networks for sales forecasting in fashion retail. Decision Support Systems 114:81–93. https://doi.org/10.1016/j.dss.2018.08.010

[18] Ma Y, Ding Y, Yang X, et al (2020) Knowledge enhanced neural fashion trend forecasting. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp 82–90, https://doi.org/10.1145/3372278.3390677

[19] Mall U, Matzen K, Hariharan B, et al (2019) Geostyle: Discovering fashion trends and events. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 411–420

[20] McAuley J, Targett C, Shi Q, et al (2015) Image-based recommendations on styles and substitutes. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp 43–52, https://doi.org/10.1145/2766462.2767755

[21] Niinimäki K, Peters G, Dahlbo H, et al (2020) The environmental price of fast fashion. Nature Reviews Earth & Environment 1(4):189–200

[22] Papadopoulos SI, Koutlis C, Sudheer M, et al (2022) Attentive hierarchical label sharing for enhanced garment and attribute classification of fashion imagery. In: Recommender Systems in Fashion and Retail. Springer, p 95–115, https://doi.org/10.1007/978-3-030-94016-4_7

[23] Qin Y, Song D, Chen H, et al (2017) A dual-stage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:170402971

[24] Singh PK, Gupta Y, Jha N, et al (2019) Fashion retail: Forecasting demand for new items. arXiv preprint arXiv:190701960

[25] Skenderi G, Joppi C, Denitto M, et al (2021) Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends. arXiv preprint arXiv:210909824 https://doi.org/10.48550/arXiv.2109.09824

[26] Smith LN (2017) Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 464–472, https://doi.org/10.1109/WACV.2017.58

[27] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008

[28] Wang P, Yang A, Men R, et al (2022) Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. arXiv preprint

arXiv:220203052

[29] Xingjian S, Chen Z, Wang H, et al (2015) Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, pp 802–810

[30] Zhao B, Lu H, Chen S, et al (2017) Convolutional neural networks for time series classification. Journal of Systems Engineering and Electronics 28(1):162–169. https://doi.org/10.21629/JSEE.2017.01.18

[31] Zhou H, Zhang S, Peng J, et al (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of AAAI