Multimodal Temporal Fusion Transformers Are Good Product Demand Forecasters

Maarten Sukel, Stevan Rudinac, and Marcel Worring, *University of Amsterdam, 1098 XH, Amsterdam, The Netherlands*

Multimodal demand forecasting aims at predicting product demand utilizing visual, textual, and contextual information. This article proposes a method for such forecasting using an integrated architecture composed of convolutional, graph-based, and transformer-based networks. Since traditional forecasting methods depend on historical demand and factors like manually generated categorical information, they face challenges such as the cold start problem and handling of category dynamics. To address these challenges, our architecture allows for incorporating multimodal information, such as geographical information, product images, and textual descriptions. Experiments with the multimodal approach are performed on a real-world dataset of more than 50 million data points of article demand. The pipeline presented in this work enhances the reliability of the predictions, demonstrating the potential of leveraging multimodal information in product demand forecasting.

Demand forecasting is a crucial task that has garnered extensive research in time series analysis and regression analysis. Typical examples are predicting energy consumption, sales trends, and various tasks in health care. The potential for forecasting in the economic sphere, particularly in the retail sector, is promising. The retail industry operates on thin margins, and goods are perishable, making an accurate forecast of product demand essential to optimize inventory management, minimize waste, and stay competitive. Due to the large volumes in the retail industry, even small improvements in prediction accuracy have a high impact.

The task of product demand forecasting based on the various sources of information available is challenging due to often having to predict multiple horizons into the future for a wide range of granularities. The time horizons used in this work range from days to weeks. The granularities we consider are *warehouses*, *products*, and different *ordering moments* during the day. Traditional demand forecasting methods often rely on tabular information, like historical sales data, weather forecasts, events, and seasonalities, only. Even though these features are important, they do not fully represent information about the type of product the demand is being predicted for or the context in which the product can be purchased.

In online retail, purchase decisions often rely on product images and descriptions.[18] Thus, visual and textual product information are decisive factors in potential product demand. There is abundant textual product information available for customers to decide what to buy, such as the name and description of the product, the nutrition label, and the list of ingredients. The visual aspect is important, too, for the customer since customers prefer appealing and properly depicted products.[12]

Geographic information is another important indicator of customer preference for specific types of products.[13] It is particularly interesting for product demand forecasting because the different regions may be associated with different customer behavior and, consequently, demand patterns and also have regional distribution centers. Differences arise, for example, from demographics, socioeconomic factors, and location-dependent data, such as type of housing. Weather, events, and seasonality have also been proven invaluable for understanding customer preference.[17]

To further improve product demand forecasting, tabular information should be combined with multimodal product information and geographical context about the delivery area. In this article, we propose a novel multimodal approach to product demand forecasting that achieves just that.

1070-986X © 2024 IEEE
Digital Object Identifier 10.1109/MMUL.2024.3373827
Date of publication 7 March 2024; date of current version 8 July 2024.

Another aspect that makes product demand forecasting in a real-world setting challenging is the fast-changing assortment and high volumes of inference required to support supply chain operations. There are multiple reasons for the dynamic nature of categorical product information. For example, AB testing and hyperpersonalization are commonplace in such settings. Another common source of category dynamics is that the information about products is constantly optimized to improve the search, browsing, and filtering experience. The third reason is the frequent use of custom categories during special events, such as the holiday season. Due to the abrupt category changes next to more gradual evolution, the data often contain noise. We conjecture that, by relying more on product information and utilizing multimodal feature extraction methods, the effects of category changes causing noise in real-world demand forecasting can be reduced.

In this work, we create a multimodal method to extract features from products that are effective for product demand forecasting and fuse them with geographical and tabular information. Our approach is based on the use of a temporal fusion transformer (TFT).[10] Transformers have recently achieved state-of-the-art results in a wide range of natural language processing and computer vision tasks while also having state-of-the-art results in the domain of demand forecasting[10,11] for tasks like predicting electricity consumption and traffic occupancy. In this article, we make a significant step forward in multimodal demand forecasting by developing a multimodal TFT (MTFT), showing that, by incorporating both product text and images as input modalities, a TFT is able to learn a more comprehensive representation of the product and its potential demand while also capturing the preferences of customers in different geographic areas.

For the textual product information, we propose a transformer-based component optimized for demand forecasting to extract the features from the product text, such as the name, description, and ingredients. For the extraction of visual features from product images, we deploy a convolutional neural network (CNN)-based component, optimized for generating input for the MTFT. The textual and visual approaches do not need large quantities of demand-forecasting-specific training data because their weights are pretrained on large Internet collections and fine-tuned afterward. These multimodal features are then combined with the tabular features using static covariate encoders (for continuous numerical values, often referred to as ”reals”) and long short-term memory (LSTM) encoders (for static and dynamic tabular data) to pass the information through the temporal fusion decoder, as illustrated in Figure 1. For these continuous numerical values, separate static covariate encoders are used to feed information into the temporal fusion decoders. This allows for leveraging both static multimodal contextual information and temporal sequence information. The information extraction modules and the fusion layer form the core of our multimodal pipeline for demand forecasting.

To study the role of different modalities, a large-scale, real-world dataset of the online grocery store Picnic is used for evaluation. Picnic operates in The Netherlands, Germany, and France. The company works by delivering groceries to people's houses the next day and operates out of several fulfillment centers that cover different geographical regions. Product demand forecasting is at the core of Picnic's operation and is used to accurately assess the quantity of all products customers will order. Due to the company being available online only, there is a large quantity of structured product information available in several different modalities. In an ablation study, we investigate what multimodal information is most beneficial for product demand forecasting. In addition, different design choices for the architectures used for multimodal product demand forecasting are compared. This study shows that an MTFT outperforms a baseline using tabular features only.

The main contributions of this article are as follows:

- A state-of-the-art transformer-based approach for product demand forecasting that incorporates multimodal product descriptions composed of textual and visual product information as well as geographical demand information.
- An innovative method that automatically processes and categorizes multimodal product data, eliminating the need for manual product category generation, not only reducing human intervention but also enhancing adaptability to dynamic real-world scenarios.
- An extensive evaluation on a large real-world dataset demonstrating the effectiveness and applicability of the proposed approach, paving the way for using multimodal approaches in product demand forecasting.

RELATED WORK

In this section, we first discuss traditional demand forecasting approaches and then the methods for integration of multiple modalities, and, finally, we discuss multimodal product representation learning.

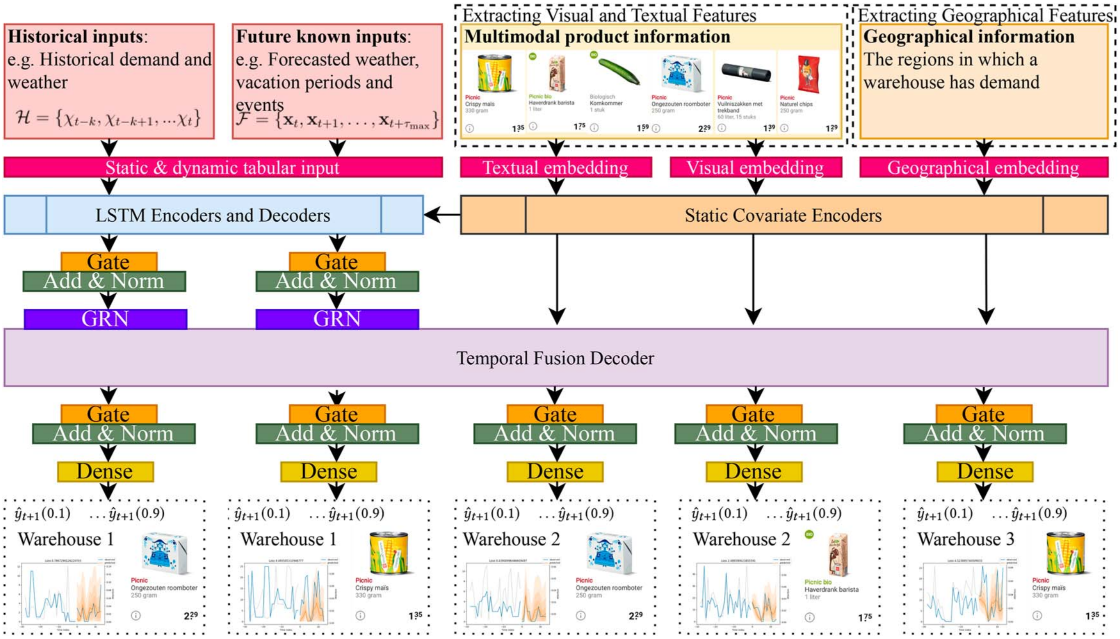

Historical inputs:
e.g. Historical demand and weather
$\mathcal{H} = \{\chi_{t-k}, \chi_{t-k+1}, \ldots \chi_t\}$
Future known inputs:
e.g. Forecasted weather, vacation periods and events
$\mathcal{F} = \{\mathbf{x}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+\tau_{\max}}\}$
Extracting Visual and Textual Features
Multimodal product information
Extracting Geographical Features
Geographical information
The regions in which a warehouse has demand
Static & dynamic tabular input
Textual embedding
Visual embedding
Geographical embedding
LSTM Encoders and Decoders
Static Covariate Encoders
Gate
Add & Norm
GRN
Temporal Fusion Decoder
Dense
$\hat{y}_{t+1}(0.1)$ ... $\hat{y}_{t+1}(0.9)$
Warehouse 1
Warehouse 2
Warehouse 3

FIGURE 1. Tabular data sources, such as historical inputs (in particular, historical demand and weather) are combined with future known inputs. These traditional data sources are fused with embeddings of multimodal product information and geographical delivery information to make multimodal product demand predictions. The approach uses an encoder-decoder-based temporal fusion transformer to handle static and dynamic information. Predictions are made for a range of quantiles per warehouse, delivery period, and product. GRN: gated residual network; LSTM: long short-term memory.

Demand Forecasting

Traditional approaches that extrapolate historical demand and more advanced methods using tree-boosting techniques have been effective in demand forecasting,[19] but they are underperforming when the demand goes outside of the range found in the training set. Advancements in deep learning have paved the way for more sophisticated techniques, including TFTs[10] and time-series dense encoder (TiDe).[1] They have been proven to be even more effective in accurately predicting demand in various fields, including retail, transportation, and finance. TFTs are more complex architectures and yield the best results on short-term forecasting. TiDe has shown that an architecture based on multilayer perceptrons can achieve even better results than TFTs while having lower training costs and higher inference speed.

The mentioned state-of-the-art demand forecasting approaches are not capable of out-of-the-box handling of raw multimodal input. Therefore, in this work, we develop approaches that can be used to feed multimodal features into a TFT model likely capable of handling such complexity and incorporating multimodal information. We evaluate the approach in a real-world application. To the best of our knowledge, this is the first study of multimodal product demand forecasting, thus paving the way for using multimodal approaches in product demand forecasting, which is a novel problem for the community.

Integration of Multiple Modalities

Multimodal approaches have proven useful for a variety of tasks. For single modalities, like text in natural language processing, there are transformer-based architectures, such as DistilBERT,[15] that achieve state-of-the-art results in a variety of tasks, such as question answering and text classification. In the visual domain, where tasks such as image classification and semantic segmentation were often performed using CNN-based approaches,[7] transformers have also been achieving state-of-the-art results. More recently, multimodal approaches improved a variety of tasks by being able to make effective representations that capture both visual and textual information. These multimodal approaches often unlock new possibilities. For example, combining visual and textual tasks allows for unified vision-language understanding and generation with approaches such as BLIP.[9] In addition to unlocking new tasks, multimodal approaches can also improve performance on existing tasks. Examples are using geographical and user click data to represent hotels or the forecasting of traffic flows. We also believe that in the case of product demand forecasting, incorporating multimodal features can result in improved performance as well as generalizability, and multimodal transformers are a natural way to bring those information sources together.

Multimodal Product Representation Learning

Multimodal product analysis has been a popular research topic in multimedia. Food images and texts are used for extracting recipes and ingredient assessment tasks.[4,8,20] Product embeddings can be used for various tasks, such as search, recommendation, and personalization. Approaches for creating product embeddings based on shopping behavior, like One Embedding to Do Them All,[16] are often used to improve recommender systems while also showing potential in other tasks. Product images are often used in information retrieval tasks, such as the use of clothing images to help customers find relevant fashion,[3] or novel applications like screenless shopping.[5]

Due to the variety of products that occur in the fashion retail industry, retrieving the correct product for a customer is not a trivial task. Recent advancements in multimodal conversational search have allowed the user to more intuitively find the right products. Multimodal deep neural networks are also used for attribute prediction and in e-commerce catalog enhancement.[14] In work by Dheenadayalan et al.,[2] multimodal information from news articles is used for demand forecasting. This resulted in an increase in performance, demonstrating the potential of using multiple modalities to extract new and richer information for improving demand forecasting. Based on these positive experiences with using multimodal product analysis in such applications, we conjecture that it is likely that a multimodal approach for demand forecasting will yield good results.

In conclusion, the use of multimodal product features for product demand forecasting is a relatively unexplored and exciting new field of research.

METHODOLOGY

MTFT

In this section, we describe our proposed method for multimodal product demand forecasting.

Our MTFT is a transformer-based model that leverages self-attention to capture the complex temporal dynamics of multiple time sequences and modalities. This makes it an effective tool for product demand forecasting, accommodating both multimodal static and dynamic temporal input. It is important to understand the distinction between dynamic and static features: dynamic features, such as historical demand and weather patterns, shift and evolve over time, directly influencing product demand. Static input, on the other hand, encompasses elements like product images $\mathcal{V}$ and textual input $\mathcal{T}$. These remain unchanged over time but offer essential context about the product. These two types of inputs undergo distinct processing pathways, as illustrated in Figure 1.

Dynamic features are processed in real time to accommodate the changing landscape, while static inputs are embedded into the model, offering a stable product context. To ensure no data leakage between historical and future dynamic features, these representations are combined with historic input data. They are only used alongside known future inputs in a temporally separated manner.

For the integration of the static multimodal input, static covariate encoders are utilized. At their core, static covariate encoders use gated residual networks, proposed by Lim et al.,[10] to transform static inputs into context vectors. These context vectors encapsulate the essence of static data and are integrated at various junctures in the temporal fusion decoder. By doing so, the MTFT ensures that multimodal static information, which remains unchanged over time, is still used to condition or influence the temporal dynamics of the forecast. For example, consider a product's static textual input consisting of the product's name, description, and ingredients. While this input does not change over time in our setup, it can still provide valuable insight into how the product's demand might fluctuate in different scenarios. The static covariate encoder will capture this invariant information and relay it as a context vector to other components of the MTFT, ensuring that even static features play a part in shaping the forecasting model's decisions.

Being a forecasting task, our method distinguishes historical $\mathcal{H}$ and future inputs $\mathcal{F}$.

$$\mathcal{H} = \{\chi_{t-k}, \chi_{t-k+1}, \dots \chi_t\} \tag{1}$$

$$\mathcal{F} = \{\mathbf{x}_t, \mathbf{x}_{t+1}, \dots, \mathbf{x}_{t+\tau_{\max}}\}. \tag{2}$$

The MTFT employs a simple yet effective sequence-to-sequence LSTM model to naturally accommodate these two types of inputs in time series data, especially when differentiating between observed inputs due to varying numbers of historic and future entries. To further incorporate the influence of static product information into the local processing of dynamic features, the initial cell state and hidden state of the first LSTM layer are initialized using the context vectors derived from the static covariate encoders. Approaches that are more complex than an LSTM layer are likely possible; however, to keep the architecture similar to the TFT base, for comparison reasons, we did not adjust this element of the network.

The output of the MTFT is in quantiles $\hat{y}$, which are formulated through the simultaneous prediction of multiple percentiles for each forecasted timestep. For instance, the model might provide estimates for the 10th percentile (indicating a lower bound), the 50th percentile (the median), and the 90th percentile (an upper bound). These percentiles give a holistic view of the probable outcomes, allowing for the anticipation of underforecasting and overforecasting scenarios. The computation behind the quantile forecasts is a linear transformation of the output emanating from the temporal fusion decoder.

There are a variety of modalities that serve as static input features. Each product $p_l \in \mathcal{P}$ is associated with visual content $v_l \in \mathcal{V}$. Textual content $t_l \in \mathcal{T}$ of the product consists of names N, descriptions D, and ingredients I. In geographical content $g_l \in \mathcal{G}$, we have the regions and warehouses. There is also traditional time-dependent static and dynamic tabular information. While historical inputs $\mathcal{H}$ represent past data, "known future inputs" $\mathcal{F}$ are unique, as they pertain to future events but are already known at the time of inference. A detailed description of the variables is provided in Table 1, and in Figure 2 some further examples are given. Therefore, we can write the prediction problem as

$$\hat{y} = \mathrm{MTFT}(\mathcal{V}, \mathcal{T}, \mathcal{G}, \mathcal{H}, \mathcal{F}). \quad (3)$$

Feature Extraction Methods

In the following sections, we describe how the multimodal input is constructed for the static covariate encoders of the MTFT. Our approaches for feature extraction consist of four components. First, we present methods for creating visual and textual product embeddings by fine-tuning networks and taking the learned representation from the network's last layer as an embedding. Then, we discuss the approach for the creation of geographical embeddings for the same task using a graph-based approach.

Extracting Visual Features

For visual feature extraction, we consider two approaches, one based on CNNs and one based on transformers.

We adapt ResNet152,[7] pretrained on ImageNet, for demand forecasting using visual features. Product images are preprocessed by padding, resizing to 224 × 224, and undergoing random horizontal flips and rotations before normalization to ImageNet's pretrained weights. We configure the ResNet's final layers for linear output, reducing their size for use in the TFT as a static embedding. Trained with mean square error loss on average product demand, visual features from the penultimate layer are aimed at forecasting product demand. Alternatively, we employ the transformer-based BLIP[9] multimodal network for feature extraction. Known for handling noisy data across domains like e-commerce, BLIP is trained on tasks such as visual question answering and image captioning.

Due to high data volumes in real-world applications, the visual embedding's dimensionality in both extraction methods is condensed. An overview is available in Figure 3. Through our experiments, principal component analysis (PCA) was identified as the optimal feature reduction method.

Extracting Textual Features

Textual product features, like descriptions, names, and ingredients, convey crucial product insights. For instance, a mention of "cheerful" could have a relationship with higher sales during a festive season, as illustrated in Figure 2. Similarly, sustainability certifications for salmon can sway customer preferences in certain regions.

For product demand forecasting, we adopt a pretrained multilingual DistilBERT[15] by appending the final layer with reduced dimensionality suitable for embedding into the TFT. DistilBERT is favored for its efficiency in extracting multilingual textual information, matching larger models in performance with reduced size. It is based on a multilingual uncased tokenizer and model and is trained on product average demand, considering names, ingredients, and descriptions. An overview of this pipeline is provided in Figure 3. As an alternative, we also use the BLIP architecture for textual embeddings, with similar dimensionality reduction.

We hypothesize that merging visual and textual features will enhance the MTFT's demand forecasting capability. We thus experiment with a BLIP-based combined input, analogous to its unimodal versions.

Extracting Geographical Features

The geographical location of customers influences their purchasing preferences.[13] To enable the TFT model to assimilate this, we devised a geographical feature generation pipeline. As an example, areas celebrating Halloween might show increased demand for pumpkin-related products (compare Figure 2).

We constructed a graph highlighting regions around specific warehouses and their associated delivery routes, as depicted in Figure 4. The graph's nodes represent regions $\mathcal{R}$ and warehouses $\mathcal{W}$, with edges illustrating delivery pathways. Using Node2vec,[6] we transformed this graph into representations suitable for demand forecasting, assessing the potential of geometric deep

TABLE 1. Overview of the traditional features used, the multimodal features, and the actual.*

Feature	Description	Example
Time index	Integer with time index from prediction	–28 to 14
Prediction date	The date of predicting	17 December 2022
Prediction timestamp	The timestamp of predicting	17 December 2022 12:01:00
Delivery date	The date of delivery	20 December 2022
Delivery weekday	The weekday of delivery	Monday
Delivery month	The month of delivery	December
Warehouse	The warehouse of the forecast	UID
Part of day	Morning or evening delivery	Evening
Product	UID for product	UID
Products ordered at prediction time	Confirmed consumer units	53
Confirmed number of deliveries at prediction time	The number of deliveries confirmed during prediction	3550
Rate of confirmed deliveries	The rate at which deliveries are confirmed	0.3
Minutes until slot cutoff	The number of minutes until slot cutoff	42
Maximum temperature	Maximum temperature of the day, °C	3.1
Vacation period	If there is a vacation period in the delivery region	0
Christmas	Proximity to Christmas	0
Text and visual (BLIP)	Multimodal embedding created by using BLIP and PCA	Dimension of 10
Text (BLIP): Product description, name, and ingredients	Textual embedding created by using BLIP and PCA	Dimension of 10
Visual (BLIP): Product image	Visual embedding created by using BLIP and PCA	Dimension of 10
Visual (ResNet): Product image	ResNet152 fine-tuned for product demand forecasting	Dimension of 10
Text (DistilBERT): Product description, name, and ingredients	DistilBERT fine-tuned for product demand forecasting	Dimension of 10
Geo (Node2Vec): Warehouses and regions	Geographical embedding created with node2vec	Dimension of 10
Amount of product demand	The total number of consumer units, the target of the forecast	203

*PCA: principal component analysis; UID: unique identifier.

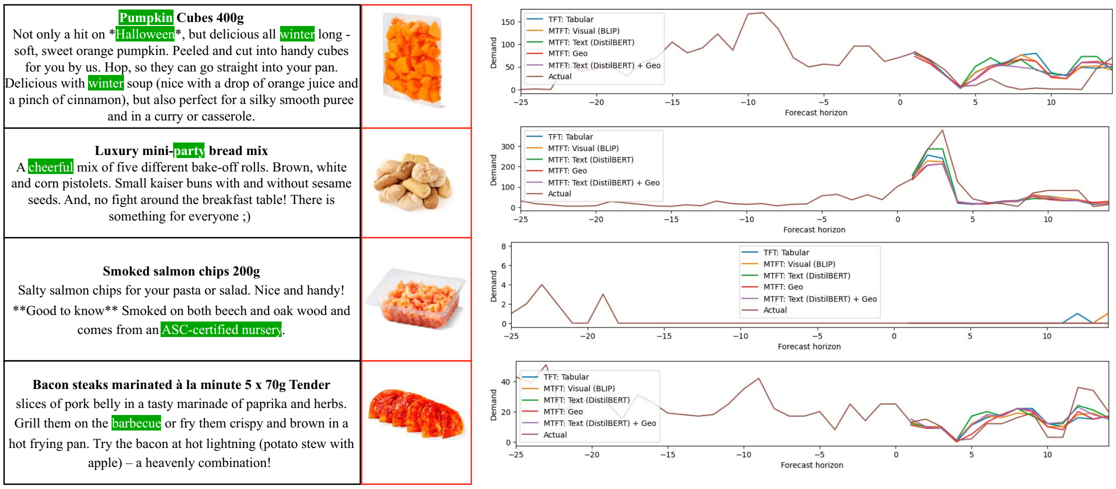
learning. An overview is provided in Figure 3. The generated embeddings encapsulate the nuances of localized delivery areas, facilitating MTFT's understanding of regional demand nuances.

EXPERIMENTAL SETUP

In this section, we first discuss the data that have been used for the experiments and then the evaluation criteria used for the task of multimodal demand forecasting.

Data

The dataset used for the experiments contains texts and images for a wide range of products. The product texts include product descriptions, names, and ingredients, while the images depict the products themselves. In addition, we have several years of product demand to train and evaluate the model on. The dataset has the granularity of warehouse, product, delivery date, and delivery moment expressed as a part of the day.



Pumpkin Cubes 400g
Not only a hit on *Halloween*, but delicious all winter long - soft, sweet orange pumpkin. Peeled and cut into handy cubes for you by us. Hop, so they can go straight into your pan. Delicious with winter soup (nice with a drop of orange juice and a pinch of cinnamon), but also perfect for a silky smooth puree and in a curry or casserole.

Luxury mini-party bread mix
A cheerful mix of five different bake-off rolls. Brown, white and corn pistolets. Small kaiser buns with and without sesame seeds. And, no fight around the breakfast table! There is something for everyone ;)

Smoked salmon chips 200g
Salty salmon chips for your pasta or salad. Nice and handy! **Good to know** Smoked on both beech and oak wood and comes from an ASC-certified nursery.

Bacon steaks marinated à la minute 5 x 70g Tender
slices of pork belly in a tasty marinade of paprika and herbs. Grill them on the barbecue or fry them crispy and brown in a hot frying pan. Try the bacon at hot lightning (potato stew with apple) – a heavenly combination!

FIGURE 2. Several examples of products where the textual and visual data can improve the product demand forecast due to, e.g., the availability of seasonal context. A random selection of the forecasted product demand is shown for several architectures in combination with the actuals at the granularity of a specific warehouse and part of day (right). For illustrative purposes, the Dutch texts utilized in the experiments are translated to English. Geo: geographical; MTFT: multimodal temporal fusion transformer; TFT: temporal fusion transformer.

In the context of demand forecasting for online grocery stores, one of the most important input features is the confirmed demand. Given that orders are placed in advance for delivery within a specified timeframe, it is possible to obtain knowledge about the number of specific products that have already been ordered for any given delivery moment. However, more traditional tabular data are also used, an overview of which can be found in Table 1.

The dataset for training encompasses 38,206,962 entries, covering the timeframe from December 2021 to December 2022. For validation purposes, a subset containing 5,643,705 entries from the same period is employed. Evaluation is executed on two fronts: a test set of 10,867,373 entries from the concluding two weeks of 2022 and a secondary test set involving 16,856,549 entries for delivery dates from 1 January 2023, to 1 April 2023.

The strategic selection of the final two weeks of 2022, notably the Christmas season, is deliberate. It presents a unique challenge in forecasting due to the distinct product mixes and heightened and varying

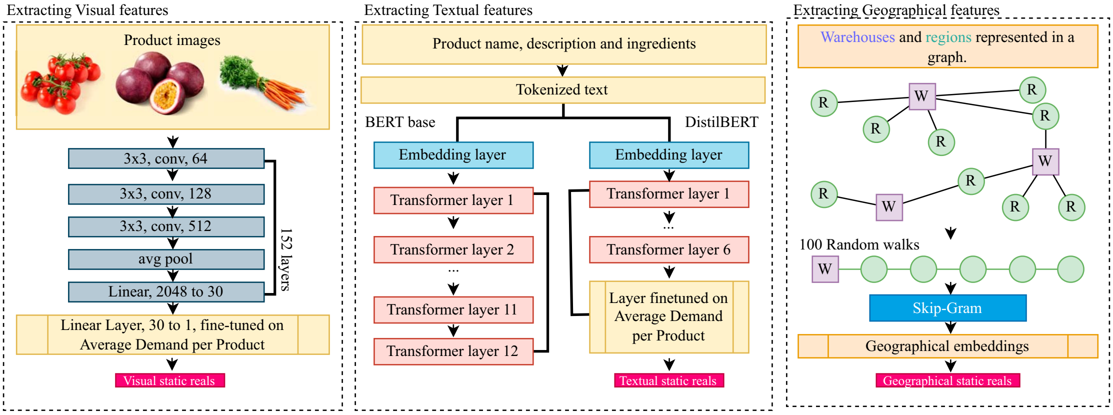


FIGURE 3. Feature extraction methods used in different parts of the proposed pipeline. We extract visual features using ResNet152[7] and, for the textual feature extraction, we utilize DistilBERT,[15] both modified for demand forecasting. To create the geographical embeddings, we make use of Node2Vec[6] on a graph of warehouses and regions. R: region; W: warehouse; conv.: convolutional.

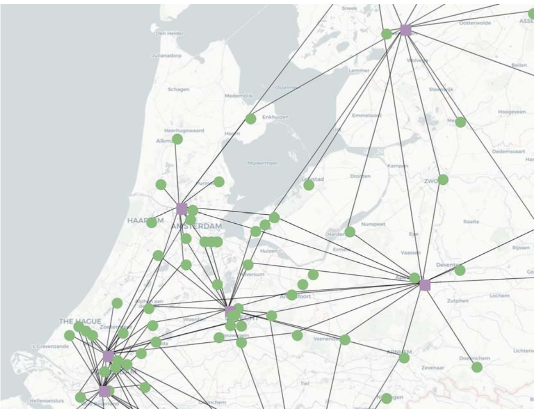


FIGURE 4. Geographical overview of how the warehouses ■ are connected to specific regions ○. For illustrative purposes, the number of regions has been reduced.

purchasing activities characteristic of this period. The evaluation during these weeks aims to validate the proposed method's robustness under such fluctuating conditions. Concurrently, the additional test set captured during a longer period serves to demonstrate the model's sustained accuracy over time without retraining, reflecting more consistent consumer behavior patterns.

Evaluation Criteria

To assess the effectiveness of the different approaches, we train a baseline using only traditional input data and compare it to different combinations of multimodal input. All experiments have been conducted on a single NVIDIA Tesla K80, where training took approximately four hours, getting optimal validation loss after 75–125 epochs.

We utilize a range of metrics to evaluate the various dimensions of demand forecasting accuracy and reliability. The weighted absolute percent error [WAPE (4)] plays a pivotal role in assessing the impact of significant forecasting errors, which carry direct real-world implications. The mean absolute error (MAE) provides an average measure of the model's error, without accounting for outliers or the direction of forecast errors (either under- or overforecasting). The mean signed deviation [MSD (6)] reveals any bias toward

TABLE 2. Evaluation of approaches on 2662 products as well as performance differences between multimodal approaches and the baseline.*

Name	Hidden size	Test data	RMSE	MSD	MAE	WAPE
MTFT: Text (DistilBERT) + geo	240	Christmas 2022	**7.74**	–1.06	**3.35**	**39.20**
MTFT: Geo	240	Christmas 2022	7.85	–0.92	3.36	39.33
MTFT: Text (DistilBERT)	240	Christmas 2022	7.76	**–0.89**	3.42	39.97
MTFT: Visual (BLIP)	240	Christmas 2022	7.98	–1.35	3.43	40.10
MTFT: Visual (ResNet)	240	Christmas 2022	8.29	–1.22	3.47	40.57
MTFT: Visual (BLIP) + text (BLIP)	240	Christmas 2022	7.97	–0.96	3.49	40.87
MTFT: Text (BLIP)	240	Christmas 2022	8.32	–1.18	3.55	41.57
TFT: Tabular (baseline)	240	Christmas 2022	8.40	–1.19	3.63	42.39
MTFT: Text (DistilBERT)	240	Q1 2023	**10.79**	**–0.88**	**3.99**	**38.83**
MTFT: Text (DistilBERT) + geo	240	Q1 2023	11.04	–1.37	4.01	38.99
MTFT: Geo	240	Q1 2023	11.03	–0.94	4.01	39.05
MTFT: Visual (ResNet)	240	Q1 2023	11.42	–0.99	4.10	39.88
TFT: Tabular	240	Q1 2023	11.09	–0.90	4.16	40.48
TFT: Tabular (baseline)	24	Christmas 2022	8.40	**–1.19**	**3.63**	**42.39**
MTFT: Image (BLIP)	24	Christmas 2022	8.56	–1.38	3.64	42.58
MTFT: Text (BLIP)	24	Christmas 2022	8.35	–1.26	3.64	42.61
MTFT: Multimodal (BLIP)	24	Christmas 2022	**8.01**	–1.20	3.66	42.78

*The MTFT model outperforms the baseline across multiple metrics and a variety of test data given a hidden size of 240. geo: geographical; MAE: mean absolute error; MSD: mean signed deviation; Q1: quarter 1; RMSE: root-mean-square error; WAPE: weighted absolute percent error.

TABLE 3. Improvement in WAPE split out over the forecast from one to seven days ahead and eight to 14 days ahead for the Christmas period 2022.

Name	1–7 days	8–14 days
TFT: Tabular (baseline)	36.7	46.2
MTFT: Visual (BLIP)	34.8	43.7
MTFT: Text (DistilBERT)	35.3	43.2
MTFT: Geo	34.4	42.7
MTFT: Text (DistilBERT) + geo	**34.2**	**42.6**

under- or overforecasting, indicating a systematic deviation in predictions. Furthermore, the root-mean-square error [RMSE (7)] is instrumental in identifying outliers, offering a gauge for the extent of forecasting errors.

These metrics collectively provide a detailed understanding of the multifaceted performance of demand forecasting. Despite its application in related research, we have chosen not to incorporate the symmetric mean absolute percent error (SMAPE) into our analysis. This decision stems from SMAPE's limitations when dealing with datasets that include a considerable number of zero values, which can undermine its effectiveness and accuracy.

In the equations displayed, $\hat{y}_i$ is the predicted value, y_i is the true value, and n is the number of samples.

$$\mathrm{WAPE} = 100 \times \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} y_i} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{5}$$

$$\mathrm{MSD} = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i) \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{7}$$

EXPERIMENTAL RESULTS

In this section, several experiments and their results are discussed to answer the following research questions:

- Are MTFTs effective in multimodal product demand forecasting?
- What are the benefits of visual, textual, and geographical modalities as well as the different pipelines for multimodal product demand forecasting?

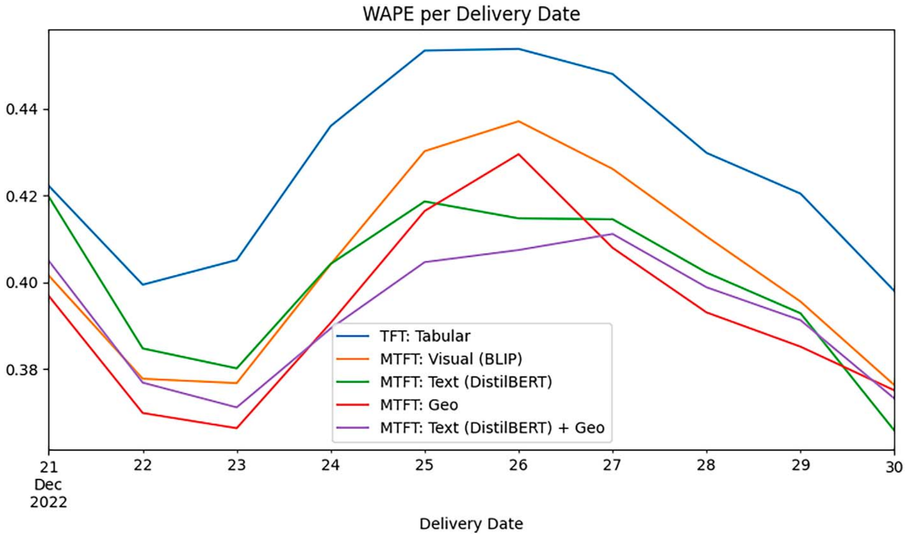


FIGURE 5. WAPE per delivery date of the prediction displays both the TFT baseline and the various MTFT variants. Every delivery date exhibits a marked enhancement using diverse multimodal strategies. It is worth noting that on 25 and 26 December, given the significant volume of Christmas-specific deliveries, the improvement attributed to the textual component is even more pronounced. WAPE: weighted absolute percent error.

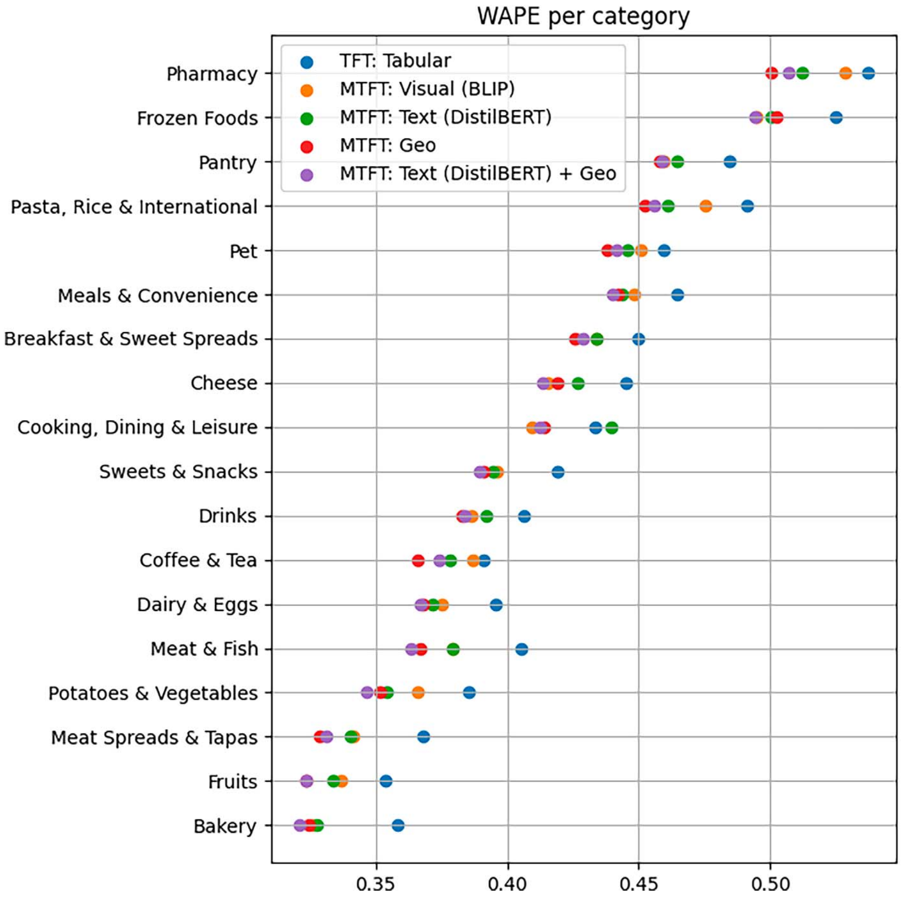


FIGURE 6. WAPE for a selection of product categories shows an improvement on almost all categories of products. The figure is based on the evaluation performed on the test set of Christmas 2022.

MTFT Effectiveness

Due to the intricacy of the multimodal features, we evaluate what hyperparameters for the TFT would work best. We expect a smaller network to not be effective at capturing the complexity of these features. To evaluate that hypothesis, a large and a small network size have been configured. The larger network had a hidden layer size of 240, a continuous size of 80, a dropout of 0.2, and a learning rate of 0.001. The smaller network had a hidden size of 24, a continuous size of 16, a dropout rate of 0.10, and a learning rate of 0.01. As can be seen in Table 2, when using a smaller network size, there are improvements in some metrics; however, these improvements were not consistent for the range of metrics that we want to improve on.

In Table 2, we observe that, with a larger network size, the MTFT improves performance compared to a traditional approach. Our methods for using the textual and geographical information improve performance on all metrics. When looking at forecasts that are made more than seven days ahead, the improvements of adding more detailed product information increase even further, as can be seen in Table 3.

It can be concluded that effectively capturing the complexity of multimodal features requires a sufficient number of hidden layers. Experiments show that a hidden layer size of 24 is inadequate for effectively utilizing additional input data. However, increasing the hidden layer size to 240 enables the effective use of the information.

Benefits of Visual, Textual, and Multimodal Features

To comprehend the impact of each modality on the model's overall performance, we conducted an ablation study. In this, the MTFT was trained using various combinations of unimodal and multimodal features and approaches.

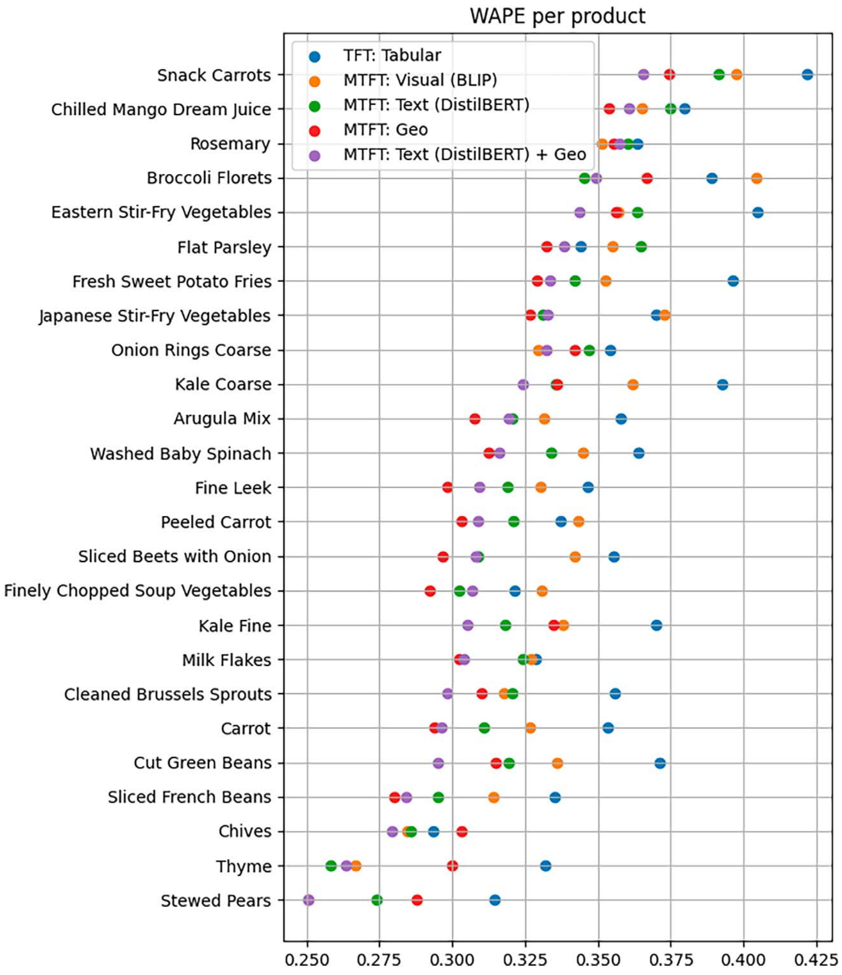


FIGURE 7. WAPE for a random selection of products, showing an improvement on almost all products. The type of features being most beneficial differ per product. The figure is based on the evaluation performed on the test set of Christmas 2022.

Results from our study illustrate the power of integrating multimodal information for demand forecasting. As shown in Table 2, a performance improvement was observed across all metrics: WAPE, RMSE, MSD, and MAE. Notably, the MSD metric indicates a consistent tendency across all models to underforecast. Figures 5–7 further spotlight the enhanced performance of the MTFT with the larger network size.

The MTFT consistently outperformed the baseline, evident across all test set delivery dates (Figure 5). A discernible weekly performance pattern, attributable to the absence of promotional data and unequal weekly demand distribution, further adds to our findings.

While the MTFT generally outpaces the baseline across categories, in Figures 6 and 7, we offer a granular perspective on WAPE scores across diverse categories and products. Key categories like "fruits," "drinks," and "meat spreads and tapas" exhibit significant WAPE improvements. This can be attributed to these product types possessing distinct characteristics that resonate with regional consumer preferences, and the textual information of these products offers insight into demand in different seasons. Figure 5 further underscores the advantage of integrating geographical and textual embeddings, especially evident during peak festive times such as 25 and 26 December. The Christmas period is often challenging to forecast with traditional methods due to the cold start problem and more frequently changing categories. The solid performance of our MTFT approach shows that it is effective in addressing these nontrivial problems.

When evaluating further ahead in the first quarter of 2023, model performance enhancements are predominantly seen in textual features, with geographical features still adding value, albeit to a lesser extent. This trend is likely a result of data drift, attributed to changes in the coverage areas of the warehouses. Addressing this issue through regular retraining, as customary in production environments, should resolve the problem.

CONCLUSION

In this article, we introduced a novel approach to product demand forecasting, leveraging product texts, images, and geographical data. The use of transformer-based neural networks allowed for the effective integration of textual and geographical information, leading to improved performance compared to traditional approaches. Our experiments on a novel real-world dataset demonstrate the effectiveness of the proposed approaches in predicting demand for a wide range of products. Additionally, our approach can work with the often noisy categorical product information. It handles the cold start problem of a new product by using visual and textual modalities, which could allow for the quicker adoption of newly introduced products. In addition, our approach outperforms state-of-the-art baselines.

Due to the scale of an online retailer, even small forecasting improvements result in an enormous reduction in the waste of products while keeping sufficient products in stock. Our ablation study unveiled invaluable insights into the efficacy of multimodal strategies. It underscored the symbiotic relationship between different modalities in boosting the model performance. The BLIP-driven creation of visual features distinctly surpassed the baseline using a large enough network size. However, refining transformer- and graph-based models specifically for product demand forecasting achieved even better results. From the new multimodal input, textual features emerged as the most influential. After that, geographical features offered the most information, and even visual data also offered substantial forecasting enhancements for specific product categories.

In conclusion, the use of multimodal product information and geographical embeddings is effective for the tasks of product demand forecasting. Finally, since the retail sector is such a high-volume market with perishable goods, this work has a high potential for a positive impact on the environment and economic benefits to retailers while paving the way for research into multimodal product demand forecasting.

REFERENCES

1. A. Das, A. Leach, R. Sen, R. Yu, and W. Kong. "Long horizon forecasting with TiDE: Time-series dense encoder." Google Research. Accessed: Jun. 2023. [Online]. Available: https://openreview.net/forum?id=pCbC3aQB5W
2. K. Dheenadayalan, N. Kumar, S. Reddy, and S. Kulkarni, "Multimodal neural network for demand forecasting," in *Proc. 29th Int. Conf., Neural Inf. Process. (ICONIP)*, Cham, Switzerland: Springer-Verlag, 2023, pp. 409–421.
3. S. Gao, F. Zeng, L. Cheng, J. Fan, and M. Zhao, "Fashion image search via anchor-free detector," in *Proc. Int. Conf. Multimedia Retrieval*, 2022, pp. 416–425, doi: 10.1145/3512527.3531355.
4. F. Gelli, T. Uricchio, X. He, A. Del Bimbo, and T. S. Chua, "Learning subjective attributes of images from auxiliary sources," in *Proc. 27th ACM Int. Conf. Multimedia*, New York, NY, USA: ACM, 2019, pp. 2263–2271, doi: 10.1145/3343031.3350574.
5. Y. Gong, J. Yi, D. D. Chen, J. Zhang, J. Zhou, and Z. Zhou, "Inferring the importance of product appearance with semi-supervised multi-modal enhancement: A step towards the screenless retailing," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1120–1128.
6. A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 855–864, doi: 10.1145/2939672.2939754.
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
8. Y. Honbu and K. Yanai, "Unseen food segmentation," in *Proc. Int. Conf. Multimedia Retrieval*, 2022, pp. 19–23, doi: 10.1145/3512527.3531426.
9. J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2022, pp. 12,888–12,900.
10. B. Lim, S. Ö. Ar Ik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct./Dec. 2021, doi: 10.1016/j.ijforecast.2021.03.012.
11. B. Lim and S. Zohren, "Time-series forecasting with deep learning: A survey," *Philos. Trans. Roy. Soc. A*, vol. 379, no. 2194, 2021, Art. no. 20200209, doi: 10.1098/rsta.2020.0209.
12. M. Mazloom, R. Rietveld, S. Rudinac, M. Worring, and W. van Dolen, "Multimodal popularity prediction of brand-related social media posts," in *Proc. 24th ACM Int. Conf. Multimedia*, New York, NY, USA: ACM, 2016, pp. 197–201, doi: 10.1145/2964284.2967210.
13. N. Ramya and S. M. Ali, "Factors affecting consumer buying behavior," *Int. J. Appl. Res.*, vol. 2, no. 10, pp. 76–80, 2016.
14. L. F. Sales, A. Pereira, T. Vieira, and E. de Barros Costa, "Multimodal deep neural networks for attribute prediction and applications to e-commerce catalogs enhancement," *Multimedia Tools Appl.*,

vol. 80, no. 17, pp. 25,851–25,873, 2021, doi: 10.1007/s11042-021-10885-1.

15. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
16. L. Singh, S. Singh, S. Arora, and S. Borar, "One embedding to do them all," 2019, *arXiv:1906.12120*.
17. A. S. Vergori, "Patterns of seasonality and tourism demand forecasting," *Tourism Econ.*, vol. 23, no. 5, pp. 1011–1027, 2017, doi: 10.1177/1354816616656418.
18. E. S. Wang, "The influence of visual packaging design on perceived food product quality, value, and brand preference," *Int. J. Retail Distrib. Manage.*, vol. 41, no. 10, pp. 805–816, 2013.
19. J. Wolters and A. Huchzermeier, "Joint in-season and out-of-season promotion demand forecasting in a retail environment," *J. Retailing*, vol. 97, no. 4, pp. 726–745, 2021, doi: 10.1016/j.jretai.2021.01.003.
20. Y. Yamakata, A. Ishino, A. Sunto, S. Amano, and K. Aizawa, "Recipe-oriented food logging for nutritional management," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 6898–6904, doi: 10.1145/3503161.3549203.

MAARTEN SUKEL is a Ph.D. candidate on the topic of integrating of integrating geo, temporal, textual, and visual data through multimodal fusion for real-world applications at the University of Amsterdam, 1098 XH, Amsterdam, The Netherlands. His research interests include multimodal machine learning, real-world applications, and demand forecasting. Sukel received his Msc. in data science from the University of Amsterdam. Contact him at m.m.sukel@uva.nl.

STEVAN RUDINAC is an associate professor in the Amsterdam Business School and a guest researcher at the Informatics Institute, both at the University of Amsterdam, 1098 XH, Amsterdam, The Netherlands. His research interests include multimedia analytics, computer vision, information retrieval, and machine learning. Rudinac received a Ph.D. in video search and visual summarization from the Technical University of Delft. Contact him at s.rudinac@uva.nl.

MARCEL WORRING is a full professor in multimedia analytics in the Informatics Institute at the University of Amsterdam, 1098 XH, Amsterdam, The Netherlands. His research interests include multimedia analytics, artificial intelligence, and visual analytics. Worring received a Ph.D. in image analysis from the Free University Amsterdam. He is a Senior Member of IEEE. Contact him at m.worring@uva.nl.