iFetch: Multimodal Conversational Agents for the Online Fashion Marketplace

Ricardo Gamelas Sousa
Pedro Miguel Ferreira
Pedro Moreira Costa
Pedro Azevedo
ricardo.sousa@farfetch.com
FARFETCH

Joao Paulo Costeira
Carlos Santiago
jpcosteira@e-tecnico.ulisboa.pt
Institute for Systems and Robotics
(ISR/IST), Universidade de Lisboa

Joao Magalhaes
David Semedo
Rafael Ferreira
jmag@fct.unl.pt
Universidade Nova de Lisboa

Alexander I. Rudnicky
Carnegie Mellon University

Alexander Georg Hauptmann
Carnegie Mellon University

ABSTRACT

Most of the interaction between large organizations and their users will be mediated by AI agents in the near future. This perception is becoming undisputed as online shopping dominates entire market segments, and the new "digitally-native" generations become consumers. iFetch is a new generation of task-oriented conversational agents that interact with users seamlessly using verbal and visual information. Through the conversation, iFetch provides targeted advice and a "physical store-like" experience while maintaining user engagement. This context entails the following vital components: 1) highly complex memory models that keep track of the conversation, 2) extraction of key semantic features from language and images that reveal user intent, 3) generation of multimodal responses that will keep users engaged in the conversation and 4) an interrelated knowledge base of products from which to extract relevant product lists.

CCS CONCEPTS

• **Information systems** → Information retrieval; • **Applied computing**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**;

KEYWORDS

Conversational Commerce, Natural Language Processing, Computer Vision

ACM Reference Format:
Ricardo Gamelas Sousa, Pedro Miguel Ferreira, Pedro Moreira Costa, Pedro Azevedo, Joao Paulo Costeira, Carlos Santiago, Joao Magalhaes, David Semedo, Rafael Ferreira, Alexander I. Rudnicky, and Alexander Georg Hauptmann. 2021. iFetch: Multimodal Conversational Agents for the Online Fashion Marketplace. In *Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI (MuCAI '21), Oct. 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3475959.3485395

MuCAI '21, October 24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8674-6/21/10.
https://doi.org/10.1145/3475959.3485395

1 INTRODUCTION

High-fashion marketplaces require top-class customer interaction. Users of such online platforms demand reliable, precise and timely service and expect the best in class customer experience throughout the customer journey. Therefore, a frictionless experience with a high-touch feeling is key to clientele satisfaction. As our high-end fashion business continues to grow, our services are more than ever under increasing pressure. More recently, with the impact of COVID, it created an additional demand that was not anticipated.

So it is without surprise that scaling up FARFETCH business to our ambition while maintaining our customer's loyalty is a challenge. We are currently witnessing an enormous paradigm shift away from the traditional search and click thanks to conversational assistants in online shopping and other social media-related tasks, especially for younger generations. WeChat, the largest social network in China, and WhatsApp, now a core service in Facebook, are the most unambiguous signals of this paradigm shift by closing the gap between consumers and brands.

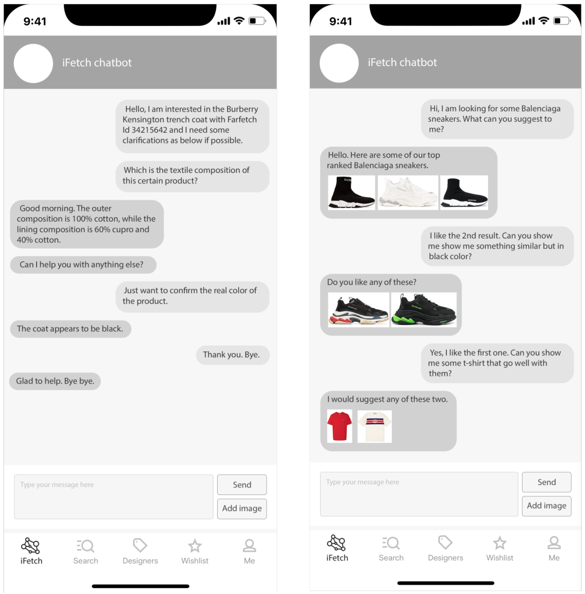
The challenge addressed by iFetch is to mimic a fashion specialist that understands the customer needs and provides fashion advice leveraging the vast textual and visual data together with knowledge accumulated by past experiences with a massive number of users. Our vision is to make a step-change in the online high-fashion marketplace by advancing conversational AI technology with multimodal capabilities (see Fig. 1).

For an e-commerce marketplace, task-oriented multimodal conversational agents (MCA) have the potential to make a ground-breaking shift in the way users do their online shopping.

2 METHODS

Soon, most interactions between large organizations and their users will be mediated by Artificial Intelligence ("AI") agents. This perception is becoming indisputable, as online purchases conquer entire market segments and the new "digitally native" generations become consumers.

iFetch is integrating a new generation of chat agents that interact with users, seemingly using textual and visual information. Through conversation, iFetch is providing its customer base guidance and a "physical store"-like experience while maintaining user engagement. Such implies contributions to the following components:



Hello, I am interested in the Burberry Kensington trench coat with Farfetch Id 34215642 and I need some clarifications as below if possible.

Which is the textile composition of this certain product?

Good morning. The outer composition is 100% cotton, while the lining composition is 60% cupro and 40% cotton.

Can I help you with anything else?

Just want to confirm the real color of the product.

The coat appears to be black.

Thank you. Bye.

Glad to help. Bye bye.

Hi, I am looking for some Balenciaga sneakers. What can you suggest to me?

Hello. Here are some of our top ranked Balenciaga sneakers.

I like the 2nd result. Can you show me show me something similar but in black color?

Do you like any of these?

Yes, I like the first one. Can you show me some t-shirt that go well with them?

I would suggest any of these two.

Figure 1: iFetch Mockup design

(1) Complex memory models that keep track of the conversation;
(2) Extraction of critical semantic features from language and images that reveal user intent;
(3) Multimodal responses that will keep users engaged in the conversation;
(4) And, an interrelated knowledge base of products from which to extract relevant product lists.

Understanding utterances requires a natural language processing step aiming to extract several relevant aspects regarding the user intent. Given the current scientific advances, we decided to pursue algorithms that jointly do utterance encoding and state tracking.

Using the BERT-DST [1] baseline as a starting point, a slot-filling approach method discovers slots in the user utterances by classifying dialogue spans. We followed current state-of-the-art approaches from the state-of-the-art analysis SimpleTOD [5] and UBAR [8]. In both methods, a single decoder-only Transformer-based model (GPT-2) is used to develop an end-to-end task-oriented system with proven results on the MultiWOZ dataset. In specific, our interest is in generating slot-value pairs from a dialogue using a similar approach.

We know that fashion is a visually rich field. Their global visual characteristics (category, texture, colour) describe apparel and visual attributes especially relevant for clothing (neckline type/shape, sleeve length, zipped vs buttoned). However, these are often small details in the whole image that, as we know, require an overwhelming annotation effort for the daily insertion of new products. To mitigate this issue, we proposed a Self Explainable Noisy Label Flipping for Multi-Label (SELF-ML) [3]. This classification framework exploits the relation between visual attributes and appearance and the feature space's "low-rank" nature. It learns a sparse reconstruction of image features as a convex combination of very few images - a basis - that are correctly annotated.

Moving forward, the agent also needs to offer a set of products to the customer based on specific utterances of the dialogue. Therefore, the most fundamental aspect of faceted product searches is creating a product list that has a high diversity when users are uncertain about what they wish to shop. Search operations will support search constraints based on product properties, such as price.

Regarding the generation of responses, our agent utterances will fall into the categories: greet, catalogue, disambiguations, product lists, opinion and check-out. Such will be employed either by neural methods and template-based methods. Agent utterances generated by neural techniques will use multi-scale attention mechanisms over the dialogue state and the memorized facts to allow a more natural interaction, with fewer linguistic repetitions.

3 CONCLUSION

iFetch will spring on a service that will allow closer and faster interactions on the point of purchase and elevate the level of service on an increasingly on-demand market where the customer requires immediate answers to his needs. Doing so will significantly impact the CR of our marketplace visitors on a market that shows significant potential given the low online penetration rates. The customers are the new marketing channel of the brands, and the advantage and differentiating factor come from knowing them. Today, with the power of technology, it is already possible to re-create the personal relationship they are used to in the physical world through the investment in data analysis and technology-based tools.

ACKNOWLEDGMENTS

This work was funded by the iFetch project, Ref. 45920, co-financed by ERDF, COMPETE 2020, NORTE 2020 and FCT under CMU Portugal.

REFERENCES

[1] Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *arXiv preprint arXiv:1907.03040* (2019).

[2] Beatriz Quintino Ferreira, Luís Baía, João Faria, and Ricardo Gamelas Sousa. 2018. A unified model with structured output for fashion images classification. *arXiv preprint arXiv:1806.09445* (2018).

[3] Beatriz Quintino Ferreira, Joao P Costeira, and Joao P Gomes. 2021. Explainable Noisy Label Flipping for Multi-Label Fashion Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3916–3920.

[4] Rafael Ferreira, Mariana Leite, David Semedo, and Joao Magalhaes. 2021. Open-Domain Conversational Search Assistant with Transformers. *arXiv preprint arXiv:2101.08197* (2021).

[5] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796* (2020).

[6] Beatriz Quintino Ferreira, Joao P Costeira, Ricardo Gamelas Sousa, Liang-Yan Gui, and Joao P Gomes. 2019. Pose guided attention for multi-label fashion image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

[7] David Semedo and João Magalhães. 2020. Adaptive Temporal Triplet-loss for Cross-modal Embedding Learning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1152–1161.

[8] Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2020. UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. *arXiv preprint arXiv:2012.03539* (2020).