Learning Cold-start Time Series and Product-Image Joint Embeddings

Stijn Verdenius, Andrea Zerio, and Roy L.M. Wang

WAIR, Amsterdam, NL, https://wair.ai

Abstract. This work discusses the challenges of an open research agenda on contrastive embedding learning between product images and sales time series for downstream cold start forecasting and feature aggregation.

Keywords: Multi-modal · Representation Learning · Cold-start · Computer Vision · Time-series · Machine Learning for Logistics · Fashion

What if an image holds the secret to a product's market success? Cold start forecasting tackles the challenge of making accurate predictions in the absence of any historical data, posing a challenge for traditional forecasting methods [6]. This problem is particularly relevant for industries like fashion retail [4, 7]. We explore a novel approach centred on learning a robust representation of product images, which captures the relationship between visual features and sales patterns at product launches. Taking inspiration from existing literature on multi-modal embedders, contrastive learning and using images for cold-start forecasting [2, 5, 7, 8, 10–12], we propose minimizing dot product between product image and cold-start time series embedding-pairs. In this work, we will explore some open questions of how to create a joint time-series-image embedding space and whether time series data can serve as an effective soft target, similar to the role of language in vision-language pre-training models [5].

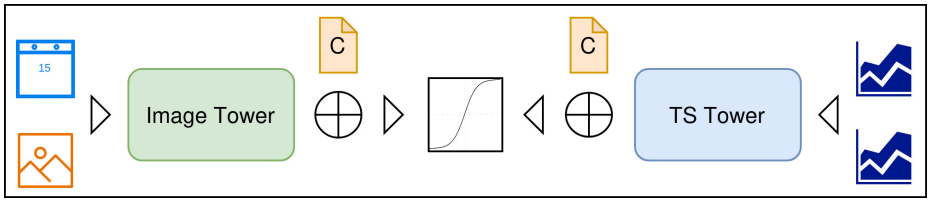

Fig. 1: *Dual-tower methodology: left tower processes images with additive metadata embeddings, the right post-launch multivariate time series. Sigmoid loss from SigLIP [11] aids contrastive learning. Category-specific bias (c) enhances distinctiveness per segment. Towers are finetuned ViT [1] and Lat-PFN-encoder [9]*

Figure 1 illustrates the proposed method. The left side presents product images and metadata available prior to launch – e.g. the proposed release season. The right displays multivariate data on online and offline sales during the first 10 weeks post-launch. Further details will be provided in subsequent documents.

This approach offers several significant advantages over forecasting directly. First, embeddings provide flexibility and can support various tasks like classification and image generation. Second, it is common for large catalogues to have incomplete data. The embedding space allows the system to make informed predictions even when, for example, pictures are absent by using fall-backs or averages of similar products or aggregates. Similarly, it enables the aggregation of visual features – which is impossible with raw pixels – allowing usage of these features on aggregate levels, like a product category. Finally, embedding spaces allow for vector arithmetic which enables manipulations in latent space, such as interpolations or vector composition – similar to Word2VEC [3].

In our ongoing work, we observe a reasonable fit, especially when including the category and metadata embeddings; however, further refinements are necessary to achieve the desired level of accuracy on downstream tasks. Some important unresolved discussions remain:

1. It is unclear if a time series can be an informative enough soft target for an image, and vice versa. Whilst images are common inputs to dual-tower embedders [5, 11], the selection of time series is far less trivial. Currently, we only select covariates aligned with a typical target for downstream forecasting. However, consider a more flexible approach and temporarily set aside the constraints of the downstream task. By doing so, we could include a broader range of covariates that describe the situation post-launch, such as weather statistics, analogous to how descriptive language in captions functions as soft targets in prior image-language contrastive learning works [5, 11]. This raises the question: could incorporating more diverse covariates lead to richer and more holistic embeddings? Or might this approach introduce distractions or even create opportunities for overfitting or leakage? Further experimentation is necessary to explore these possibilities and determine the optimal approach.
2. How much structural information about the data should be provided to the model during training? To strengthen the relationship between modalities in each data point we included a bias term (the product category), which we added to the outputs of both sets of embedders to help the model distinguish existing clusters of data points within its embedding space, which improved the alignment of the data modalities in the joint space. The trade-off, however, is that the more extra information is added, the more each data point becomes uniquely identifiable, reducing the model's generalization capabilities.
3. There is a risk of model collapse where images are bypassed by only matching non-image features with time series data. However, omitting these features is not desirable since pictures don't exist in a vacuum. For example, the textual soft-label "a boy in a *fashionable* shirt" should yield different image matches depending on the date of the prediction. The same is likely true with the time-series modality. Additionally, events such as COVID-19 are impactful.

In this work, we discussed challenges in creating a robust multi-modal image embedder for cold start applications like forecasting, aiming to prompt further academic discussion and guide future research.

Acknowledgments. This study was funded fully by WAIR (AI4R B.V.), Luchtvaarstraat 4, Amsterdam, Netherlands and executed by its research team.

Disclosure of Interests. All authors are employed by funder WAIR, which has an interest in producing, advertising and selling a strategy for cold start forecasting.

Bibliography

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[2] Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23390–23400 (2023)

[3] Mikolov, T.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[4] Pereira, A.M., Moura, J.A.B., Costa, E.D.B., Vieira, T., Landim, A.R., Bazaki, E., Wanick, V.: Customer models for artificial intelligence-based decision support in fashion online retail supply chains. Decision Support Systems **158**, 113795 (2022)

[5] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[6] Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M.: Methods and metrics for cold-start recommendations. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 253–260 (2002)

[7] Skenderi, G., Joppi, C., Denitto, M., Scarpa, B., Cristani, M.: The multimodal universe of fast-fashion: the visuelle 2.0 benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2241–2246 (2022)

[8] Sun, M., Li, F., Zhang, J.: A multi-modality deep network for cold-start recommendation. Big Data and Cognitive Computing **2**(1), 7 (2018)

[9] Verdenius, S., Zerio, A., Wang, R.L.: Lat-pfn: A joint embedding predictive architecture for in-context time-series forecasting. arXiv preprint arXiv:2405.10093 (2024)

[10] Wei, Y., Wang, X., Li, Q., Nie, L., Li, Y., Li, X., Chua, T.S.: Contrastive learning for cold-start recommendation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 5382–5390 (2021)

[11] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)

[12] Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18123–18133 (2022)