The Role of Expert-driven Prompt Engineering for Fine-grained Zero-shot Classification in Fashion

Dhanashree Balaram* , Matthew Nokleby[1], Thiyagarajan Ramanathan[2]

Ajitesh Gupta[3], Ravi Shankar[4]

1 Purpose

E-commerce fashion retailers face the increasingly challenging task of tracking fast-changing trends in delivering content-specific recommendations. A key differentiator among competitors in this space is the ability to identify fine-grained fashion attributes that distinguish these trends. Such fine-grained attributes go deeper than, e.g., the colors, fabrics, and sizes of clothing that are easily captured in retailers' catalogs, and involve nuanced attributes, such as the specific sleeve style of a garment, its intended occasion, or whether it belongs to a trending style. In order to classify these attributes via standard methods, we require large amounts of fine-grained, accurately labeled data which is expensive and labor intensive.

2 Problem

We use CLIP [1] to carry out zero-shot classification for fine-grained fashion attributes, where text prompts represent the classes. We demonstrate that for fine-grained fashion attributes, expert-driven prompts deliver higher accuracy than coarse, naive prompts. Because CLIP depends on joint text-image similarity, we hypothesize that adding detailed fashion descriptions like the captions included in the CLIP training set, lead to more well-defined embedding spaces and higher classification accuracy.

3 Method

We evaluate the performance of zero-shot classification on fine-grained style attributes, using "standard" prompts generated automatically vs. prompts selected by an in-house team of fashion experts. We measure the accuracy of the prompt-based classifiers in terms of per-class F1 scores.

4 Result

We show the accuracy results of a "Sleeve Style" classification problem in Table 1. For Bishop sleeve, we see that expert driven prompts elevate the performance when compared to naive prompts. This is because a naive coarse-grained prompt like "Bishop Sleeve" is not descriptive enough. In contrast, "Bell Sleeve" quite obviously describes a bell shaped sleeve. To aid our investigation, we generate activation maps in Figure 1 using [2] to help visualize how different prompts affect the propagation of gradients (ResNet backbone). We see a correlation between the activation map and the F1 scores.

F1 Score		
Sleeve style	Coarse/Naive	Expert
Bell	0.33	0.32
Bishop	0.00	0.28
Cap	0.60	0.56
None	0.00	0.37

Table 1: sleeve style vs. prompt granularity

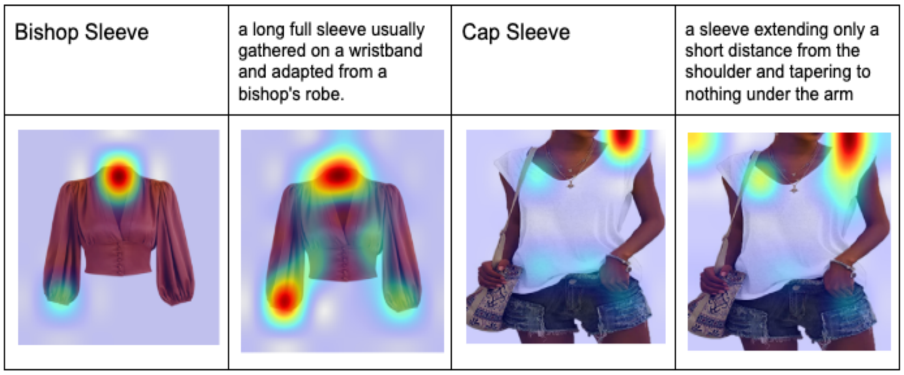
Bishop Sleeve
a long full sleeve usually gathered on a wristband and adapted from a bishop's robe.
Cap Sleeve
a sleeve extending only a short distance from the shoulder and tapering to nothing under the arm

Figure 1: Gradient maps comparison between naive prompts and expert driven prompts

5 Conclusion

We conclude that expert-driven prompts work better than naive prompts, specifically when class names vary in meaning based on their context, like “Bishop” in chess is different from fashion. Though using too many similar words in prompts can cause over-fitting (to one prompt), the key is to bring out elements that differentiate the attributes, like shapes, relative location, etc. Our analysis brings out the intricacies of expert-driven prompt engineering in fine-grained fashion classification.

References

[1] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18-24 Jul 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[2] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.