*Concept*ualizing *Concept* Drift

Isaac Roberts[*], Fabian Hinder, Valerie Vaquet,
Alexander Schulz and Barbara Hammer [†‡]

Bielefeld University – Faculty of Technology
Inspiration 1, 33619 Bielefeld – Germany

Abstract. Concept drift refers to the phenomenon that the underlying data distribution changes over time. While detection methods or model adjustment methods exist, a proper explanation of drift in high-dimensional settings is still widely unsolved. This problem is crucial since it enables an understanding of the most prominent drift characteristics. In this work, we propose to explain concept drift of high-dimensional data objects by means of concept activation vectors which give rise to local, phase, and a novel, global explanation called the *Concept*^{2} Drift Distribution.

1 Introduction

Most machine learning contributions assume the availability of a batch setup where a dataset is drawn i.i.d. from the data-generating distribution. However, in many real-world scenarios data is collected over time and possibly subject to changes in the underlying distribution. These changes can be induced by environmental, measurement, or societal factors. This phenomenon, called *concept drift* [1], is a challenge for many machine learning systems.

Effectively addressing *concept drift* requires three key components: accurate detection, comprehensive explanation, and an appropriate response strategy. Detection methods identify when drift occurs, while explanations characterize the nature of drift, enabling operators to respond effectively. To achieve this, explanations must convey the characteristics of a drift, by considering feature attributions or directly using histograms [2]. However, in high-dimensional scenarios, a feature-based drift explanation remains an open challenge.

One promising approach [3] addresses this by proposing a model-based explanation framework that combines drift localization [4] with well-established xAI techniques. While this approach has shown effectiveness in certain contexts, it currently focuses on local explanations for high-dimensional data, requiring operators to examine each item in the data stream individually to understand the full extent of the drift. This further underscores the need for a more comprehensive set of drift explanations in high-dimensional contexts.

To address this gap, we propose a novel drift explanation pipeline based on Concept Activation Vectors (CAVs) [5], which offers human-interpretable explanations targeted at high-dimensional settings by combining a well-established

[*]Corresponding author: iroberts@techfak.uni-bielefeld.de

[†]This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101073307 and the BMBF project KI Akademie OWL under grant agreement No 01IS24057A.

[‡]Experiment code: `https://github.com/robertsi20/ConceptualizingConceptDrift`.

concept extraction technique with the model-based explanation framework. Our proposed pipeline produces drift explanations at multiple levels—local, phase, and global—enabling a nuanced perspective. Central to our method is the novel *Concept*^{2} Drift Distribution which provides a holistic, global view of the drift. We demonstrate its utility and give quantitative and qualitative evidence that it offers a comprehensive drift characterization. In this work, we focus on image streams but are not restricted to this domain.

This paper is organized as follows: Section 2 provides background information needed to understand our proposed pipeline. In Section 3, we provide the details of the methodology and explanation scheme. Section 4 presents a quantitative evaluation of the pipeline's effectiveness and includes a case study to qualitatively evaluate the generated explanations.

2 Background

Given a data stream $S = (x_t)_{t=0}^{T}$, where at time t a sample x_t is generated by the probability measure p_t. *Concept drift* occurs if $\exists t_0 \neq t_1 : p_{t_0} \neq p_{t_1}$ [2] which describes various types of drift including abrupt, gradual, and reoccurring. We focus on a single change point, prominent in abrupt drift and inducible with drift detectors [4, 2]. *Drift localization*, i.e. identifying which samples are drifting, can be framed as a binary classification problem that predicts whether a given sample x_t appears before or after the drift [1, 4]. This formulation has two key advantages: (1) drifting samples that occur exclusively in either *phase (before or after the drift)* can be identified with high certainty, while non-drifting samples, appearing in both, are classified with low certainty; and (2) since the classification model captures information about the drift, state-of-the-art xAI techniques can be applied to explain both the model and the drift [3]. However, global explanations for high-dimensional data are not widely explored [3].

Our pipeline addresses the above limitation by building on a method, known as CRAFT [6], which provides automatic discovery of human-interpretable concepts learned by deep neural networks. The authors of CRAFT begin by embedding inputs into a non-negative activation space of a pre-trained model and utilize Non-negative Matrix Factorization (NMF) to decompose the embedded data matrix A into a product of non-negative matrices U and V, solved by reconstructing A, i.e. $(U, V) = \arg\min_{U \geq 0, V \geq 0} \|A - UV^\top\|_F^2$. The decomposition yields: V the dictionary of concepts (or concept bank) and U a reduced representation of A according to the basis V. Furthermore, by considering the variance fluctuations by perturbing U, they use Sobol Indices to attribute importance scores to the concepts for a predicted class and individual inputs.

3 Methodology

In Figure 1, we illustrate our proposed pipeline which aims to characterize and explain a drift using extracted concepts. Given a high-dimensional data stream, $x_1, x_2, \ldots, x_T$ and a drift detector that estimates a drift time t and thereby two

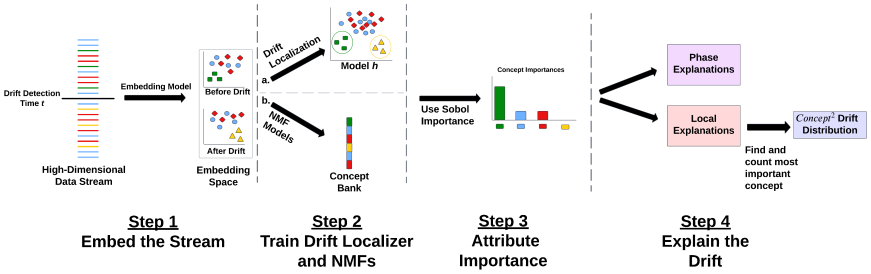

Fig. 1: Our proposed pipeline for generating concept drift explanations

sets before drift and after drift: BD $= \{i | i \leq t\}$, AD $= \{i | i > t\}$, we arrange our pipeline into 4 steps:

Step 1 - We represent the data in a meaningful form by embedding the stream using a foundation model g into an activation space with the condition that for each x_i, $g(x_i) \geq 0$ (e.g. after a ReLU layer).

Step 2a - Next, we train the model that we are going to explain: a Drift Localization model h trained to predict for a given $g(x_i)$ if $i \in$ BD or $i \in$ AD.

Step 2b - To generate the concepts, we train one NMF on patches from $\{g(x_i) | i \in \text{BD}\}$ and another NMF on patches from $\{g(x_i) | i \in \text{AD}\}$, producing two concept banks, V_{BD} and V_{AD}, each with dimension d. We then concatenate these to obtain $[V_{\text{BD}}, V_{\text{AD}}] = V_{\text{drift}}$ and can represent each $g(x_i)$ as a linear combination of the concepts in V_{drift}, with scaling factors U_i.

Step 3 - To estimate the importance of the concepts in V_{drift}, we follow CRAFT's approach and utilize the Sobol Indices [7], which measure the variance of the classifier h predictions resulting from masking concepts.

Step 4 - Repeating Step 3 for every data point, we obtain a *local* importance score $e_l(g(x_i)) \in \mathbb{R}^{2d}$ with $2d$ concepts. We augment the local score from above with consistent phase and global explanations. The *phase* importances can be computed as in literature by averaging over $e_l(g(x_i))$ for points predicted in BD and AD, respectively. We further propose a *global* explanation which is computed as follows: for each concept c, we compute the relative number of times c is the most important concept among BD points:

$$e_{\text{BD}}(c) = |\{x_i | \arg\max e_l(g(x_i)) = c, i \in \text{BD}\}| / |\{x_i | i \in \text{BD}\}|, \tag{1}$$

resulting in a probability distribution e_{BD} over the concepts. Doing the same for points from AD, we obtain e_{AD}. By assigning to each concept a probability of occurring before and/or after drift, e_{BD} and e_{AD} highlight the relevant changes and uniformities in the2 data stream. We display them together in a histogram and refer to them as the *Concept*^{2} Drift Distribution.

Moreover, the *global* explanation enables a novel, simple model $\tilde{h}$, which approximates h, to be constructed. Model $\tilde{h}$ utilizes the e_l for a given $g(x_i)$ to find the concept c^* such that $c^* = \arg\max e_l(g(x_i))$, and then assigns a label based on the $\max(e_{\text{BD}}(c^*), e_{\text{AD}}(c^*))$, making it intrinsically interpretable.

		h	$\tilde{h}$	Recon c^*	Recon all
$D1$	Accuracy	0.815 ± 0.063	0.788 ± 0.055	0.764 ± 0.062	0.773 ± 0.071
	LP Accuracy	NA	0.829 ± 0.085	0.850 ± 0.073	0.889 ± 0.080
$D2$	Accuracy	0.790 ± 0.075	0.772 ± 0.067	0.751 ± 0.071	0.767 ± 0.077
	LP Accuracy	NA	0.848 ± 0.075	0.851 ± 0.072	0.886 ± 0.075

Table 1: We report for each data stream the average accuracy in relation to the ground truth labels ('Accuracy') and to h's predictions ('LP Accuracy') over 50 runs.

4 Experimental Evaluation

In this section, we empirically evaluate the correctness and robustness of our proposed *global* explanation for drift by comparing Model $\tilde{h}$'s performance to Model h. We additionally justify our selection of only the single, most important concept c^* by investigating the information loss by reconstructing h's input with only c^* and with all concepts. Following that, we present a case study of a possible drift scenario to illustrate the generated explanations.

4.1 Compressing Drift Localization to a Single Concept

To construct the drift, we construct data streams, D1 and D2, each with 500 images. D1 is derived from food-related items in six classes contained in No ImageNet Class Objects [8], while D2 involves various types of wolves and foxes from ImageNet [9]. For each stream, we randomly assign each class to a phase of drift (BD, AD, present in both), repeating 50 times. The non-drifting inputs are randomly distributed into BD and AD. We leverage a ResNet model pre-trained on ImageNet[9] to embed each image and train a random forest with 20 minimum leaf samples to localize the drift. Then we apply our proposed pipeline with 10 concepts for BD and AD each (default in [6]) and use a patch size of 100 ($\approx 25\%$), as a compromise between performance and interpretability.

To demonstrate that our *global* explanation characterizes the drift and closely approximates h, we compare h and $\tilde{h}$'s accuracy with the ground truth and how well $\tilde{h}$ matches h's predictions, denoted as 'LP Accuracy'. We report the same metrics by applying h on the reconstruction generated with only c^* ('Recon c^*') & all concepts ('Recon all').

The results in Table 1 show that $\tilde{h}$ closely matches the performance of h on the drifts, only dropping by 2-3 points (top rows) and by matching h's prediction nearly 83%-85% of the time (bottom rows). 'Recon c^*' demonstrates that by only using the most important concept we recover 75-76% accuracy and 85% of h's predictions, both of which only slightly increase when all concepts are used for reconstruction. These results show that $\tilde{h}$ indeed closely approximates the model h in performance and, therein, selecting the most important local concept c^* is justified, since representing the data with c^* leads to a similar performance as h and resembles its predictions to a high degree. We further exemplify this by including all concepts only observing slight performance boosts.

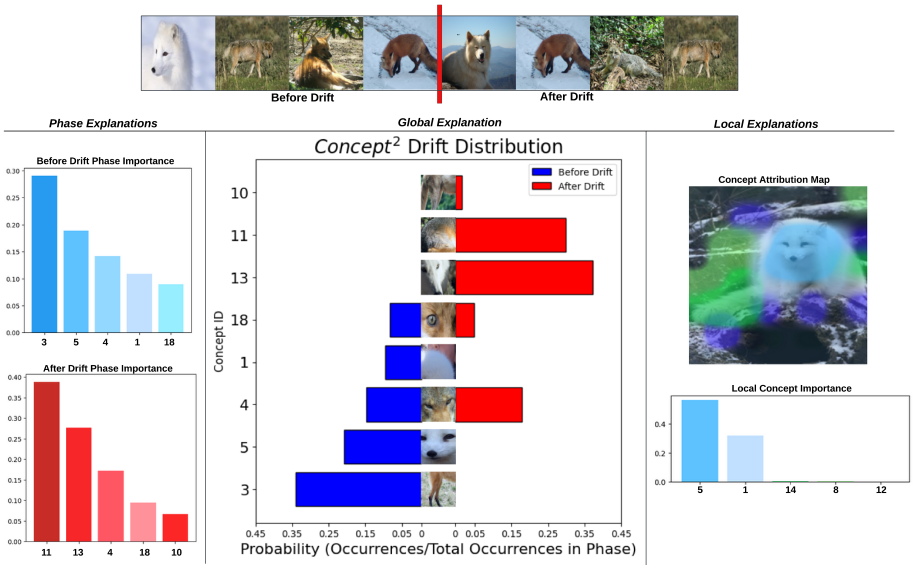

Fig. 2: Examples of the generated explanations from our proposed method in the given case study.

4.2 Case study

We select a possible stream of 500 randomly sampled images from D2 to include an abrupt drift. As shown in the top of Figure 2, Red Wolves & Arctic Foxes appear exclusively before drift, White Wolves & Grey Foxes only after drift, and some appear both before and after. Our task is to generate explanations using the derived concepts which describe the occurred drift.

The results from the global explanations, e_{AD} and e_{BD}, are shown in the center of Figure 2. Due to space constraints, we plot the top five occurring concepts in e_{AD} & e_{BD}, and on the x-axis, display their probability of occurrence, denoting by color before drift (blue) or after drift (red). We also add the patch that most activates each derived concept on the central axis.

Using this *Concept*^{2} Drift Distribution, we can immediately spot which concepts occur *exclusively* in either phase. Concepts 3, 5, & 1 materialize only in the BD phase, and upon examination of the patch, they correspond to the body of a Red Wolf and the head & fur of an Arctic Fox, respectively. Likewise in the AD phase, concepts 13, 11, & 10 occur solely and correspond to the head of a White Wolf, the body of a Grey Fox, and the body of a White Wolf. Using our proposed explanation, we have found the core concepts that explain exactly the drifting components of the stream.

In addition to identifying the drift, our global explanation also characterizes the non-drifting components. Concepts 4 & 18 are detected in both phases and correspond to the face of a Timber Wolf and a Red Fox. Indeed, these were the

images not affected by drift in our generation.

The explanation adds the benefit of compressing the stream of images to concepts without losing much information, and thus, produces a feature-based representation of the drift, and subsequently, the interpretable model $\tilde{h}$.

For more narrow explanations of the drift, the framework additionally provides *phase* (left) and *local* (right) explanations. The phase explanations give insight into the concepts that contribute most to the predicted phase made by h, while the local explanations clarify what concepts are present (bottom right) and where the concepts are detected (top right).

5 Conclusion

We introduced a novel, comprehensive framework for explaining drift using automatically extracted concept activation vectors. Our proposed framework enables the *Concept*^{2} Drift Distribution which offers a global explanation by reducing the stream of images to single concepts and assigning probabilities to their occurrence in each phase of drift. This distribution also gives rise to the interpretable model $\tilde{h}$ which, as we showed, closely approximates the drift localizer h. Moreover, we also demonstrated that our proposed framework also supplies phase and local explanations, for more specific studies into the drift. In future work, examining the framework applied to text, a closer look at the assumptions for the deep embedding, and explaining other relevant problems such as rejection strategies and outlier detection using concepts seem to be promising directions.

References

[1] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE TKDE*, page 1–1, 2018.

[2] Fabian Hinder, Valerie Vaquet, and Barbara Hammer. One or two things we know about concept drift—a survey on monitoring in evolving environments. part b: locating and explaining concept drift. *Frontiers in Artificial Intelligence*, 7, 2024.

[3] Fabian Hinder, Valerie Vaquet, Johannes Brinkrolf, and Barbara Hammer. Model-based explanations of concept drift. *Neurocomputing*, 555:126640, 2023.

[4] Fabian Hinder, Valerie Vaquet, Johannes Brinkrolf, André Artelt, and Barbara Hammer. Localization of concept drift: Identifying the drifting datapoints. In *IJCNN*, pages 1–9, 2022.

[5] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, pages 2668–2677. PMLR, 2018.

[6] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *CVPR*, pages 2711–2721, 2023.

[7] Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *NeurIPS*, 34:26005–26014, 2021.

[8] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet out-of-distribution detection evaluation. In *ICML*, volume 202, pages 2471–2506, 2023.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.