Computing, Telecommunication and Control, 2025, Vol. 18, No. 4, Pp. 7–19.
Информатика, телекоммуникации и управление. 2025. Том 18, № 4. С. 7–19.

Intelligent Systems and Technologies, Artificial Intelligence

Интеллектуальные системы и технологии, искусственный интеллект

Research article
DOI: https://doi.org/10.18721/JCSTCS.18401
UDC 004.85

CONCEPT-BASED LEARNING IN HETEROGENEOUS TREATMENT EFFECT

L.V. Utkin ✉ , *A.V. Konstantinov* , *N.M. Verbova*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ lev.utkin@gmail.com

Abstract. Estimating Heterogeneous Treatment Effects (HTE) is crucial for personalized decision-making in medicine, economics and engineering. While machine learning models for Conditional Average Treatment Effect (CATE) estimation have become increasingly accurate, they often remain black boxes, providing little insight into why treatments affect individuals differently. This paper introduces CATE-Concept Bottleneck Model (CATE-CBM), a novel framework that integrates concept-based learning with CATE estimation to bridge this interpretability gap. Our approach enforces a concept bottleneck that forces the model to express treatment effects through understandable concepts, enabling transparent reasoning about which concepts drive heterogeneous effects. Through experiments on a modified MNIST dataset, we demonstrate that CATE-CBM maintains competitive accuracy while providing meaningful concept-based explanations of treatment effect heterogeneity. The model successfully identifies how both the presence and absence of specific concepts influence treatment outcomes, offering clinicians and engineers both accurate effect estimates and interpretable rationales for personalized interventions. This work represents the first unification of concept bottleneck models with causal effect estimation, advancing the frontier of explainable artificial intelligence in causal inference.

Keywords: machine learning, concept-based learning, conditional average treatment effect, interpretation, neural network

Acknowledgements: The research was financially supported by the Russian Science Foundation, project "Machine learning models for assessing treatment effect with heterogeneous diagnostic information using expert rules" (Agreement No. 25-11-00021; available online: https://rscf.ru/project/25-11-00021/).

Citation: Utkin L.V., Konstantinov A.V., Verbova N.M. Concept-Based Learning in Heterogeneous Treatment Effect. Computing, Telecommunications and Control, 2025, Vol. 18, No. 4, Pp. 7–19. DOI: 10.18721/JCSTCS.18401

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18401
УДК 004.85

ОБУЧЕНИЕ НА ОСНОВЕ КОНЦЕПТОВ ДЛЯ ОЦЕНКИ УСЛОВНОГО ЭФФЕКТА ЛЕЧЕНИЯ

Л.В. Уткин ✉, *А.В. Константинов*, *Н.М. Вербова*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ lev.utkin@gmail.com

Аннотация. Оценка условного эффекта лечения имеет решающее значение для персонализированного принятия решений в медицине, экономике и технике. Хотя модели машинного обучения для оценки условного среднего эффекта лечения (CATE) становятся все более точными, они часто остаются «черными ящиками», не давая понимания того, почему лечение по-разному влияет на различных людей. Данная работа представляет новую модель CATE-CBM, которая интегрирует обучение на основе концептов с оценкой CATE, чтобы преодолеть этот разрыв в интерпретируемости. Предлагаемый подход использует обучение на концептах, заставляя модель выражать эффекты лечения через понятные для человека концепты, что позволяет прозрачно объяснять, какие именно концепты обуславливают эффекты. В экспериментах на модифицированном наборе данных MNIST демонстрируется, что CATE-CBM сохраняет конкурентоспособную точность, одновременно предоставляя содержательные объяснения предсказания эффекта лечения на основе концептов. Модель успешно идентифицирует, как присутствие или отсутствие конкретных концептов влияет на результаты лечения, предлагая клиницистам и политикам как точные оценки эффекта, так и интерпретируемые обоснования для персонализированных вмешательств. Данная работа представляет собой первую унификацию моделей с обучением на концептах и оценкой причинно-следственных связей, продвигая границы объяснимого искусственного интеллекта.

Ключевые слова: машинное обучение, обучение на концептах, условный средний эффект лечения, интерпретируемость, нейронная сеть

Финансирование: Исследование выполнено при финансовой поддержке Российского научного фонда в рамках реализации проекта «Модели машинного обучения для оценки эффекта лечения при разнородной диагностической информации с экспертными правилами» (Соглашение № 25-11-00021; https://rscf.ru/project/25-11-00021/).

Для цитирования: Utkin L.V., Konstantinov A.V., Verbova N.M. Concept-Based Learning in Heterogeneous Treatment Effect // Computing, Telecommunications and Control. 2025. Т. 18, № 4. С. 7–19. DOI: 10.18721/JCSTCS.18401

Introduction

The pursuit of explainability and the integration of human-centric reasoning into Machine Learning (ML) has catalyzed the development of Concept-Based Learning (CBL). Unlike conventional models that operate directly on raw, low-level features, CBL utilizes high-level, human-intelligible concepts as intermediate representations for making predictions [1]. This paradigm aims to bridge the gap between data-driven patterns and expert knowledge, leading to models that are not only more interpretable but also more data-efficient and robust [2, 3]. A prominent instantiation of this approach is the Concept Bottleneck Model (CBM), which enforces a compressed, concept-based representation of the input, forcing the final classifier to rely solely on these concepts for prediction [4]. This architectural constraint ensures that the model's decision-making process is intrinsically tied to a vocabulary of meaningful concepts, significantly enhancing its explainability [3].

The application of CBL can extend beyond standard supervised learning into more complex domains, such as estimating Heterogeneous Treatment Effects (HTE). HTE is the recognition that the effect of a treatment (e.g., a drug, a policy) varies across different individuals [5–7]. While HTE describes this general property, the Conditional Average Treatment Effect (CATE) quantifies it by measuring the average treatment effect for a specific subpopulation with given characteristics. To estimate the treatment effect, patients are typically divided into treatment and control groups, and their average outcomes are compared. This comparison provides an estimate of the causal effect of the treatment. Various approaches to estimating this effect are considered in several surveys [5, 8–12].

The primary aim of combining HTE and CBL is to provide explanations for why a certain treatment is predicted to be more effective for one individual than for another. For instance, in a medical context, a model might predict a stronger positive treatment effect for patients characterized by the concepts “high genetic marker expression” and “early disease stage”, providing clinicians with a clear, conceptual rationale for personalized treatment plans. Incorporating CBL into CATE estimation moves us beyond simply knowing that a treatment effect is heterogeneous; it provides the crucial why, explaining this heterogeneity through the lens of understandable concepts. This can lead to more informed and personalized decisions in healthcare, economics and public policy. Moreover, to the best of our knowledge, no existing method combines CBL with HTE or CATE.

Motivated by the above reasoning, we propose a model called CATE-CBM, which integrates concept learning and CATE estimation into a single framework. The model consists of two main components:

1. The first produces concept probabilities, which serve as a type of embedding.
2. The second part solves the CATE estimation problem using these predicted concept probabilities for patients in the treatment and control groups.

An important characteristic of CBL is the interpretation of predictions in terms of human-intelligible concepts. Therefore, this paper demonstrates how to locally interpret the CATE predictions made by our model.

Numerical experiments conducted on a modified MNIST dataset demonstrate how the integration of CATE estimation and CBL can improve accuracy and provide explanations for the CATE-CBM predictions in terms of concepts.

Related work

Concept-based learning

The growing interest in CBL has led to a proliferation of models aimed at improving the interpretability and explainability of ML predictions [1, 2]. These models leverage understandable concepts to make model reasoning more transparent and to align machine decisions with user intuition. The CBM [4] as a special case of CBL serves as a foundational architecture for many CBL approaches. Its efficient two-stage design in predicting concepts from inputs, then targets from concepts, has inspired numerous extensions. These include models that learn continuous concept embeddings [13], probabilistic variants to handle uncertainty [14] and investigations into concept independence and intervention [15, 16]. Further adaptations have integrated powerful pre-trained models like CLIP [17, 18] and addressed performance disparities between different CBM formulations [19].

Survey papers [20–22] comprehensively discuss aspects of CBL, CBM and their applications.

Estimating CATE

Accurately estimating CATE is fundamental to various applications. Early statistical methods were ranged from LASSO-based estimators [23] to causal forests [24]. Subsequent research extended these ideas, developing methods for censored data [25] and anomaly detection [26].

A key development was the formalization of meta-learners, flexible estimation strategies like T-learners, S-learners and X-learners [27]. More recently, neural networks have emerged as a powerful framework for CATE estimation, leading to numerous specialized architectures [28–30].

Recent work has extended CATE estimation to transformer-based architectures, leveraging attention mechanisms to model complex dependencies [31–33]. While Nadaraya-Watson kernel regression provides a theoretically grounded approach to CATE estimation [34, 35], its practical application is often limited by data sparsity, particularly in the treatment group.

Background

Concept-based learning

The paradigm of CBL formalizes a ML problem where a model must reason using a set of understandable, high-level concepts in addition to, or instead of, raw input features [36]. Formally, this framework assumes the availability of a training dataset $\mathcal{D}=\left\{\left(\mathbf{x}_i,\mathbf{c}_i,y_i\right)\right\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a d-dimensional input feature vector; $y_i \in \mathcal{Y} \subset \mathbb{R}$ is a continuous-valued regression target; $\mathbf{c}_i = \left(c_i^{(1)},\ldots,c_i^{(m)}\right) \in \mathcal{X} \subset \mathbb{R}^d$, in particular $\mathbf{c}_i \in \{0, 1\}^m$ is a vector of m binary concept annotations associated with $\mathbf{x}_i$. Here, $c_i^{(j)}=1$ indicates the presence of the j-th concept in the sample.

The core objective of CBL is twofold. The primary goal is to learn a hypothesis $h:\mathcal{X}\to\left(\mathcal{C},\mathcal{Y}\right)$ that can accurately predict both the target variable and the underlying concepts for a new input. The secondary, and equally critical, goal is to achieve a high degree of model explainability. By leveraging concepts as intermediate representations, CBL provides a transparent interface through which a user can understand *which concepts* present in an input were most influential in arriving at a final prediction, and to what degree they influence the predicted value.

A seminal architecture that instantiates this paradigm is the CBM [4]. The CBM explicitly decomposes the function h into two distinct stages: a concept encoder $g:\mathcal{X}\to\mathcal{C}$ that maps the raw input $\mathbf{x}$ to a vector of predicted concepts $\hat{\mathbf{c}}$; a *label predictor* $f:\mathcal{C}\to\mathcal{Y}$ that maps the predicted concepts to a final, continuous target $\hat{y}$.

The final prediction for a new input $\mathbf{x}$ is thus computed as $\hat{y}=f\left(g\left(\mathbf{x}\right)\right)$. This architectural design imposes a "concept bottleneck": all information from the input must flow through the intermediate concept representation before a final prediction is made. This ensures that the model's output is intrinsically and interpretably linked to the human-defined concepts, allowing a user to trace the predicted value $\hat{y}$ back to the specific concepts $\hat{\mathbf{c}}$ that caused it, thereby fulfilling the central promise of CBL to provide explainable predictions.

Treatment effect estimation

Let the available data be partitioned into two groups: a control group and a treatment group. The control group consists of c patients and is denoted as $\mathcal{C}=\left\{\left(\mathbf{x}_1,y_1\right),\ldots,\left(\mathbf{x}_c,y_c\right)\right\}$, where each patient i is characterized by a M-dimensional feature vector $\mathbf{x}_i=\left(x_{i1},\ldots,x_{id}\right)\in\mathbb{R}^M$ and a continuous outcome $y_i \in \mathbb{R}$ (e.g., survival time, blood pressure). Similarly, the treatment group contains t patients and is denoted as $\mathcal{T}=\left\{\left(\mathbf{z}_1,h_1\right),\ldots,\left(\mathbf{z}_t,h_t\right)\right\}$, with feature vectors $\mathbf{z}_i \in \mathbb{R}^M$ and outcomes $h_i \in \mathbb{R}$. For notational consistency across all $n = c + t$ patients, we define the treatment assignment indicator $T_i \in \{0, 1\}$, where $T_i = 0$ indicates assignment to the control group and $T_i = 1$ to the treatment group.

The central goal of causal inference is to estimate the effect of a treatment on an outcome. For a given patient, we define two *potential outcomes*: Y (the outcome if the patient does not receive the treatment, $T = 0$) and H (the outcome if the patient does receive the treatment, $T = 1$). A fundamental problem in causal inference is that for any single patient, we can only observe one of these potential outcomes, either Y or H, but never both. To overcome this, we estimate the *CATE*, which is the expected treatment effect for a subpopulation defined by a specific feature vector $\mathbf{x}$ [37]:

$$\tau\left(\mathbf{x}\right)=\mathbb{E}\left[H-Y\middle|\mathbf{X}=\mathbf{x}\right].$$

One of the important concepts if the treatment effect is the *propensity score* $e(\mathbf{x})$ which is the probability that a specific patient will receive the treatment given its observed characteristics (covariates), i.e., there holds $e(\mathbf{x}) = \Pr(T = 1 | \mathbf{X} = \mathbf{x})$. It is used to adjust differences between treatment and control groups and to isolate the true effect of a treatment from the effects of pre-existing differences.

Under some assumptions [37], CATE can be identified from the observed data as the difference between two conditional expectations:

$$\tau(\mathbf{x}) = \mathbb{E}\left[H | T = 1,\ \mathbf{X} = \mathbf{x} \right] - \mathbb{E}\left[Y | T = 0,\ \mathbf{X} = \mathbf{x} \right].$$

Let the outcome for a control patient be governed by g_0: $\mathbb{R}^d \to \mathbb{R}$ and for a treated patient by g_1: $\mathbb{R}^d \to \mathbb{R}$. Then we can write

$$y = g_0(\mathbf{x}) + \varepsilon,\ \ \mathbf{x} \in \mathcal{C},\ \ h = g_1(\mathbf{z}) + \varepsilon,\ \ \mathbf{z} \in \mathcal{T},$$

where ε is a random noise variable with $\mathbb{E}[\varepsilon] = 0$.

Hence, the CATE is simply the difference between these two response surfaces:

$$\tau(\mathbf{x}) = g_1(\mathbf{x}) - g_0(\mathbf{x}).$$

Proposed model

The proposed model, CATE-CBM, can be regarded as a combination of a CATE estimation model and a CBM. Its architecture is inspired by the Dragonnet model [38], which was introduced for CATE estimation. The architecture of CATE-CBM is depicted in Fig. 1.

It can be seen from the figure that the convolutional neural network (CNN) extracts a feature vector $\mathbf{v}$ which is fed to fully-connected neural networks (FCN-0 and FCN-1) for predicting the concept probability distributions $\mathbf{p} = (p_1, \ldots, p_m)$ for controls and $\mathbf{q} = (q_1, \ldots, q_m)$ for treatments, respectively. The use of CNNs is important for reducing the dimensionality of images. The whole network has three heads: two heads predict targets y and h from the corresponding concept probabilities; the third head can be regarded as the propensity score regularization. It forces the model to learn the structure of the confounding [38]. We propose to implement the propensity score regularization by means of the attention mechanism. In this case, the propensity score $e(\mathbf{x})$ or $e(\mathbf{z})$ is computed through the attention weights $a(\mathbf{q}_i, \mathbf{q}_j, \theta)$ with trainable parameters θ as follows:

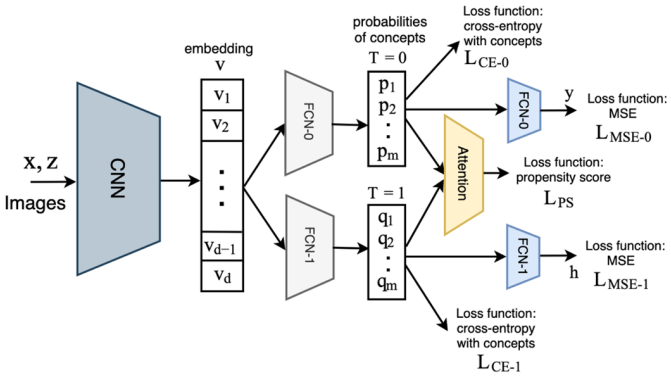

Fig. 1. Architecture of the proposed model CATE-CBM

$$e(\mathbf{x}_i) = \sum_{j=1, j\neq i}^{n} a(\mathbf{q}_i, \mathbf{q}_j, \theta) \cdot T_j,$$

where the attention weights are defined as:

$$a(\mathbf{q}_i, \mathbf{q}_j, \theta) = \frac{K_\theta(\mathbf{q}_i, \mathbf{q}_j)}{\sum_{k=1, j\neq i}^{n} K_\theta(\mathbf{q}_i, \mathbf{q}_k)}.$$

Here $K_\theta(\mathbf{q}_i, \mathbf{q}_j)$ is a kernel. In particular, if the kernel is Gaussian, then the attention weight can be expressed through the softmax function as:

$$a(\mathbf{q}_i, \mathbf{q}_j, \theta) = softmax(-\mathbf{q}_i - \mathbf{q}_j^2/\theta).$$

It should be noted that instead of the Gaussian kernel, we can use a neural network to learn a complex, data-driven similarity metric. However, this replacement significantly complicates the propensity score regularization.

In the context of the attention mechanism [39], the vector $\mathbf{q}_i$ is referred to as the *query*, while vectors $\mathbf{q}_j$ and indicators T_j are called the *keys* and *values*, respectively.

Predictions y and h for every $\mathbf{x}$ and $\mathbf{z}$, respectively, are obtained as outputs of the corresponding FCNs.

The loss function for training the whole model consists of the following five components:

• $\mathcal{L}_{MSE-0}$ and $\mathcal{L}_{MSE-1}$ are the Mean Squared Error (MSE) loss functions for the outputs y and h, respectively. The loss functions are of the form:

$$\mathcal{L}_{MSE-0} = \frac{1}{c}\sum_{i=1}^{c}(y_i - \hat{y}_i), \quad \mathcal{L}_{MSE-1} = \frac{1}{t}\sum_{i=1}^{t}(h_i - \hat{h}_i),$$

where $\hat{y}_i$ and $\hat{h}_i$ are predicted values of y_i and h_i, respectively.

• $\mathcal{L}_{CE-0}$ and $\mathcal{L}_{CE-1}$ are the cross-entropy functions controlling probabilities of concepts $\mathbf{p}$ and $\mathbf{q}$, respectively. The loss functions correspond to solving the concept classification task. For a single example with true concept values $(c_1, \ldots, c_m)$, the loss $\mathcal{L}_{CE-0}$ is defined as:

$$\mathcal{L}_{CE-0} = -\frac{1}{c}\sum_{i=1}^{c}\log(\hat{p}_i).$$

Here $\hat{p}_i$ is the predicted value of p_i. The loss $\mathcal{L}_{CE-1}$ is defined in the same way replacing $\hat{p}_i$ with $\hat{q}_i$.

• $\mathcal{L}_{PS}$ is the binary cross-entropy loss for the propensity score:

$$\mathcal{L}_{PS} = -\sum_{i=1}^{n}\left(T_i \cdot \log(\hat{e}(\mathbf{z}_i)) + (1 - T_i) \cdot \log(\hat{e}(\mathbf{x}_i))\right),$$

where $\hat{e}$ is the predicted propensity score value.

In sum, the whole loss function is defined as

$$\mathcal{L} = \gamma_1\mathcal{L}_{MSE-0} + \gamma_2\mathcal{L}_{MSE-1} + \gamma_3\mathcal{L}_{CE-0} + \gamma_4\mathcal{L}_{CE-1} + \gamma_5\mathcal{L}_{PS},$$

where γ_i, $i = 1,\ldots,5$, are hyperparameters weighting the loss components.

The predicted value of the treatment effect is computed as $\hat{\tau}(\mathbf{x}) = \hat{h}(\mathbf{x}) - \hat{y}(\mathbf{x})$. The model is trained in the end-to-end manner.

During inference, an image $\mathbf{x}$ is fed into a CNN, the output of which is an embedding $\mathbf{v}$. This embedding is then passed to two FCNs to obtain the concept probability vectors $\mathbf{p}$ and $\mathbf{q}$, such that one vector corresponds to the control group and the other to the treatment group. These vectors are fed into neural networks that generate the predictions of y and h.

It is important to note that an explicit embedding step is not strictly necessary. The concept probability vectors $\mathbf{p}$ and $\mathbf{q}$ can be obtained directly as the output of the CNN, bypassing the intermediate embedding representation. However, the intermediate embedding may have advantage in comparison with the direct implementation of the concept probabilities as the output of the CNN. The embedding layer can act as a form of regularization, preventing the concept predictors from overfitting to the training data by forcing information compression. Moreover, the CNN's feature maps are often low-level (edges, textures). An embedding layer can learn to combine these low-level features into a more sophisticated, high-level representation that is better suited for predicting complex concepts.

Local interpretation of the CATE predictions

An important question in CATE estimation is its interpretation that lies in answering the question of which concept change had the strongest impact on the estimated CATE value. The proposed model allows us to answer this question in the following way.

First, it should be noted that we consider the local interpretation which allows us to explain an individual prediction at a point of interest. Methods of the local interpretation are based on a linear approximation of the predictive model in a neighborhood around the explainable point [40, 41]. A well-known local explanation method is the Local Interpretable Model-agnostic Explanations (LIME) [42] interpreting the black-box model predictions by approximating the model at a point by a linear model whose coefficients can be viewed as a quantitative representation of the feature impacts on the prediction [43]. The approach to interprete predictions by means of the black-box model approximation at a point by the linear model can be applied to many classification and regression tasks. Therefore, we consider its use for interpreting the CATE predictions.

The output FCN-0 and FCN-1 are linear, which makes it possible to interpret each predicted function $\hat{y}(\mathbf{x})$ and $\hat{h}(\mathbf{x})$ in terms of concepts. If we identify the concepts with the highest probability, then the largest weights of the neural networks precisely show which of the identified concepts are significant for each function. It should be noted that the absence of a concept can also be significant. The change in their values, if present for a given concept, is exactly what answers the question of CATE interpretation.

Suppose that functions $f_0(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{x} + a_0$ and $f_1(\mathbf{x}) = \mathbf{b}^{\mathrm{T}}\mathbf{x} + b_0$ are linear approximations of functions $\hat{y}(\mathbf{x})$ and $\hat{h}(\mathbf{x})$, respectively, at a point $\mathbf{x}$, where $\mathbf{a} = (a_1,...,a_d)^{\mathrm{T}}$ and $\mathbf{b} = (b_1,...,b_d)^{\mathrm{T}}$. Then we can write $\hat{y}(\mathbf{x}) = f_0(\mathbf{x}) + e_0$ and $\hat{h}(\mathbf{x}) = f_1(\mathbf{x}) + e_1$. Here e_0 and e_1 are the approximation errors. The CATE in this case is defined as

$$\hat{\tau}(\mathbf{x}) = \hat{h}(\mathbf{x}) - \hat{y}(\mathbf{x}) = f_1(\mathbf{x}) - f_0(\mathbf{x}) + e_1 - e_2.$$

Note that there holds

$$|e_1 - e_0| \le |e_1| + |e_0|.$$

Hence, the linear difference $f_1(\mathbf{x}) - f_0(\mathbf{x})$ can be regarded as a linear approximation of $\hat{\tau}(\mathbf{x})$ at point $\mathbf{x}$ if the approximation errors e_1 and e_0 are small. Finally, if we know weights of FCN-0 and FCN-1, $\mathbf{a}$ and $\mathbf{b}$, the the values of the concept importance are defined as the difference $\mathbf{b} - \mathbf{a}$.

Numerical experiments

A difficulty of comparing the proposed model with other models is that CATE-CBM is the first model combining CATE and CBM. Therefore, we will show some properties of CATE-CBM by means of numerical experiments.

To study the proposed model, a synthetic dataset is constructed from the well-known MNIST dataset [44] which represents 28×28 pixel handwritten digit images. The original MNIST dataset has a training set of 60000 instances and a test set of 10000 instances[1].

Each instance in the synthetic dataset consists of four different digits randomly taken from MNIST such that the instance has two digits in the first row and two digits in the second row as it is shown in Fig. 2.

Each instance has the size 56×56. A similar dataset is used in [14] and in [45]. Concepts $c^{(1)},...,c^{(10)}$ are binary and defined by the presence of the corresponding number 1,...,9,0 in the instance. For example, the first instance has concepts (1,1,1,0,0,0,0,0,0,1), the second instance has concepts (0,0,0,0,1,1,1,0,0,1).

We analyze the proposed model by its training on numbers of instances (from 1000 till 5000). The number of testing images is 20000. The cross-validation in all experiments is performed with 50 repetitions.

For experiments, we apply functions similar to those used in [27]. They are expressed through the indicator function I taking value 1 if its argument is true. The function for controls is represented as

$$g_0(\mathbf{c}) = \mathbf{b}^{\mathrm{T}}\mathbf{c} + 5I(c_4 = 1),$$

where $\mathbf{b}^{\mathrm{T}} = (1,2,3,4,5,4,3,2,1,0.01)$.

The function for treatments is represented as

$$g_1(\mathbf{c}) = \mathbf{b}^{\mathrm{T}}\mathbf{c} + 5I(c_3 = 1) + 7I(c_8 = 1),$$

where $\mathbf{b}^{\mathrm{T}} = (0.01,1,2,3,4,5,4,3,2,1)$.

Values of y and h are generated by adding the normally distributed random numbers ε with the zero expectation and the standard deviations $\sigma_0 = 1.5$ for controls and $\sigma_1 = 2.0$ for treatments.

Suppose that we have additional information about peculiarities of digits in controls and treatments, namaly controls do not have digits “1”, treatments do not have digits “7”. In accordance with this information, we generate examples for control and treatment groups. It can be seen from Fig. 2 that the first example belongs to treatments, the second belongs to controls, the third and fouth can belong to treatments as well as to controls, therefore, they are randomly referred to the controls or treatments. If an example contains “1” and “7” simultaneously, then it is removed from the generated dataset.

For the modified MNIST datasets, we employ CNN (Fig. 1) consisting of four convolutional layers with progressively decreasing kernel sizes, starting from (8×8) to (4×4). LeakyReLU activation functions are used throughout, and the final layer is linear. FCN-0 and FCN-1 consist of two layers with sigmoid activation functions. The output FCNs contain one layer which is linear to implement interpretability of the CATE. Parameter θ of the Gaussian kernel in the propensity score regularization is trainable.

The MSE measure is used as an accuracy measure of CATE in experiments. We compare the proposed model CATE-CBM with the same implementation of the CATE estimator but without using concepts. MSE of CATE with and without using concepts is shown in Fig. 3, where MSE as functions

[1] The dataset is available at http://yann.lecun.com/exdb/mnist/.

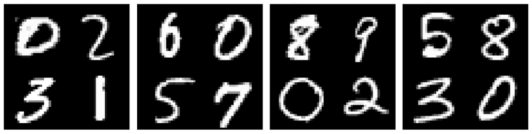

Fig. 2. Examples of the modified MNIST dataset

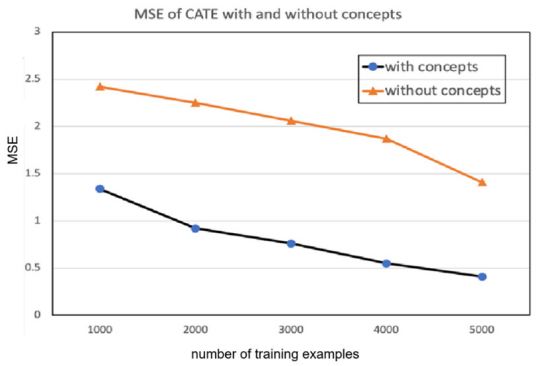

Fig. 3. MSE of CATE with and without using concepts

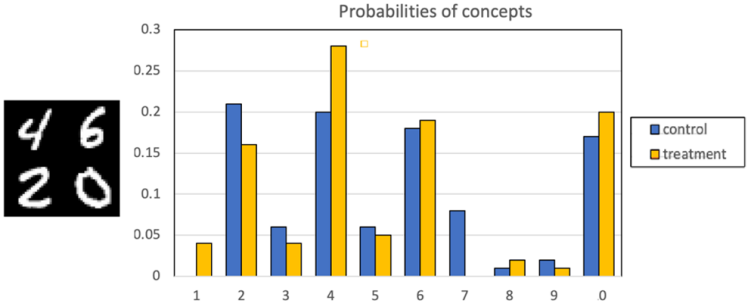

Fig. 4. Probabilities of concepts for the testing example

of the training set size of the modified MNIST is demonstrated. It can be seen from the graphs that information about concepts significantly improves the model performance.

Another important consideration is the interpretation of the results. Following the aforementioned interpretation method, we compute the probabilities of concepts under the assumptions that an example belongs to the control and treatment groups. Fig. 4 illustrates these concept probabilities for an example containing the digits 4, 6, 2 and 0. The probabilities for these corresponding concepts are the largest, indicating that the CATE-CBM model correctly recognizes them.

Fig. 5 shows the normalized weights from the output layers FCN-0 and FCN-1. The importance of each concept for the CATE prediction can be derived from the difference between the weights of FCN-1 and FCN-0. The results, depicted in Fig. 6, reveal that the most important concept is “5”, which is not present in the example. This interesting finding demonstrates that the absence of a concept can also be significant. Furthermore, the importance value for this concept is negative, implying that concept 5 acts to reduce the treatment effect. In contrast, concepts 3 and 7 have positive importance.

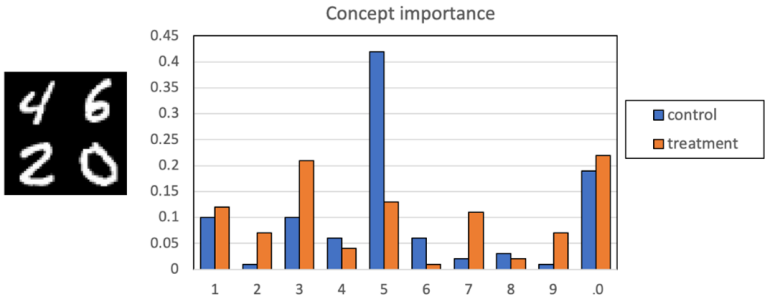

Fig. 5. Importance of concepts for the testing example

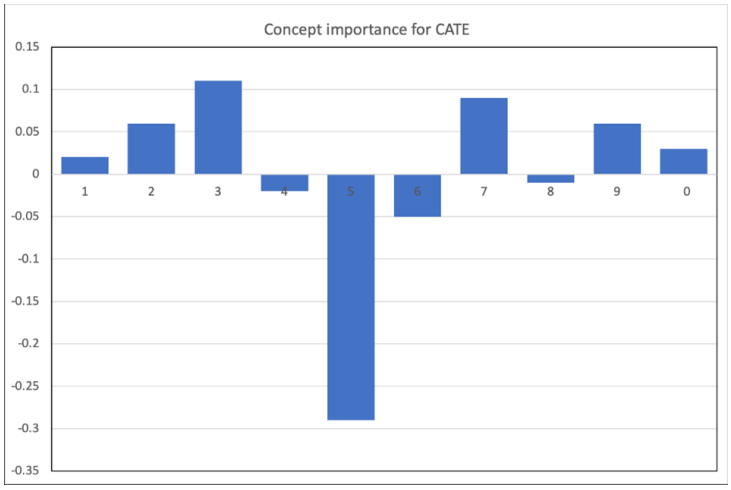

Fig. 6. Values of the concept importance for the CATE interpretation of the considered example

It should be noted that the above interpretation results pertain only to the specific example with digits 4, 6, 2 and 0.

Conclusion

This paper introduced CATE-CBM, the first model to integrate CBL with CATE estimation. By combining interpretable concept bottlenecks with CATE estimation, the model provides both accurate treatment effect predictions and understandable explanations through concept importance analysis.

The numerical experiments conducted on the modified MNIST dataset demonstrate several important properties and advantages of the proposed CATE-CBM model.

1. First, the incorporation of concept information significantly enhances model accuracy, as evidenced by the consistently lower MSE of CATE estimation compared to the same model without concept utilization.

2. Second, the model successfully identifies and extracts relevant concepts from complex image data, as shown by the high probability scores assigned to correct digit concepts in test examples.

3. Third, CATE-CBM provides transparent insights into treatment effect mechanisms through concept importance analysis.

4. Fourth, the CNN-Concept architecture proves effective for handling complex visual data while maintaining interpretability, successfully balancing predictive performance with explanatory capabilities.

These findings establish CATE-CBM as a promising approach for CATE estimation in settings where both accuracy and interpretability are crucial, particularly when dealing with high-dimensional data requiring meaningful feature extraction.

Several promising directions emerge from this work.

1. First, extending CATE-CBM to handle continuous-valued concepts and temporal treatment effects would broaden its applicability.

2. Second, developing methods for automatic concept discovery rather than relying on pre-defined concepts could enhance model flexibility.

3. Third, incorporating uncertainty quantification for both concept predictions and treatment effects would provide crucial reliability measures for decision-making.

Applications in real-world clinical trials and policy evaluation settings would further validate the approach's practical utility.

REFERENCES

1. **Kim B., Wattenberg M., Gilmer J, Cai C., Wexler J., Viegas F., Sayres R.** Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning*, 2018, Vol. 80, Pp. 2668–2677. DOI: 10.48550/arXiv.1711.11279

2. **Yeh C.-K., Kim B., Arik S.Ö., Li C.-L., Pfister T., Ravikumar P.** On completeness-aware concept-based explanations in deep neural networks. *Proceedings of the 34th International Conference on Neural Information Processing Systems* (*NeurIPS 2020*), 2020, Pp. 20554–20565.

3. **Wang B., Li L., Nakashima Y., Nagahara H.** Learning bottleneck concepts in image classification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2023, Pp. 10962–10971. DOI: 10.1109/CVPR52729.2023.01055

4. **Koh P.W., Nguyen T., Tang Y.S., Mussmann S., Pierson E., Kim B., Liang P.** Concept bottleneck models. *Proceedings of the 37th International Conference on Machine Learning*, 2020, Vol. 119, Pp. 5338–5348.

5. **Caron A., Baio G., Manolopoulou I.** Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 2022, Vol. 185, No. 3, Pp. 1115–1149. DOI: 10.1111/rssa.12824

6. **Künzel S.R., Stadie B.C., Vemuri N., Ramakrishnan V., Sekhon J.S., Abbeel P.** Transfer learning for estimating causal effects using neural networks. *arXiv:1808.07804*, 2018. DOI: 10.48550/arXiv.1808.07804

7. **Zhou X., Xie Y.** Heterogeneous treatment effects in the presence of self-selection: A propensity score perspective. *Sociological Methodology*, 2020, Vol. 50, No. 1, Pp. 350–385. DOI: 10.1177/0081175019862593

8. **Chu Z., Li S.** Continual treatment effect estimation: Challenges and opportunities. *Proceedings of Machine Learning Research*, 2023, Vol. 208, Pp. 11–17.

9. **Curth A., Peck R.W., McKinney E., Weatherall J., van der Schaar M.** Using machine learning to individualize treatment effect estimation: Challenges and opportunities. *Clinical Pharmacology & Therapeutics*, 2024, Vol. 115, No. 4, Pp. 710–719. DOI: 10.1002/cpt.3159

10. **Wang Y., Li H., Zhu M., Wu A., Xiong R., Wu F., Kuang K.** Causal inference with complex treatments: A survey. *arXiv:2407.14022*, 2024. DOI: 10.48550/arXiv.2407.14022

11. **Yao L., Chu Z., Li S., Li Y., Gao J., Zhang A.** A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*), 2021, Vol. 15, No. 5, Art. no. 74. DOI: 10.1145/344494

12. **Zhang W., Li J., Liu L.** A unified survey of treatment effect heterogeneity modelling and uplift modelling. *ACM Computing Surveys* (*CSUR*), 2021, Vol. 54, No. 8, Art. no. 162. DOI: 10.1145/3466818

13. **Ismail A.A., Adebayo J., Bravo H.C., Ra S., Cho K.** Concept bottleneck generative models. *Proceedings of ICML 2023*. Workshop on Deployment Challenges for Generative AI, pages 1–10, 2023.

14. **Kim E., Jung D., Park S., Kim S., Yoon S.** Probabilistic concept bottleneck models. *The 12th International Conference on Learning Representations*, 2024.

15. **Raman N., Zarlenga M.E., Heo J., Jamnik M.** Do concept bottleneck models obey locality? *XAI in Action: Past, Present, and Future Applications*, 2023.

16. **Laguna S., Marcinkevičs R., Vandenhirtz M., Vogt J.E.** Beyond concept bottleneck models: How to make black boxes intervenable? *arXiv:2401.13544*, 2024. DOI: 10.48550/arXiv.2401.13544

17. **Kazmierczak R., Berthier E., Frehse G., Franchi G.** CLIP-QDA: An explainable concept bottleneck model. *Transactions on Machine Learning Research Journal*, 2024.

18. **Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I.** Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 2021, Vol. 139, Pp. 8748–8763.

19. **Havasi M., Parbhoo S., Doshi-Velez F.** Addressing leakage in concept bottleneck models. *36th Conference on Neural Information Processing Systems* (*NeurIPS*), 2022.

20. **Gupta A., Narayanan P.J.** A survey on concept-based approaches for model improvement. *arXiv:2403.14566*, 2024. DOI: 10.48550/arXiv.2403.14566

21. **Lee J.H., Lanza S., Wermter S.** From neural activations to concepts: A survey on explaining concepts in neural networks. *arXiv:2310.11884*, 2023. DOI: 10.48550/arXiv.2310.11884

22. **Mahinpei A., Clark J., Lage I., Doshi-Velez F., Pan W.** Promises and pitfalls of black-box concept learning models. *arXiv:2106.13314*, 2021. DOI: 10.48550/arXiv.2106.13314

23. **Jeng X.J., Lu W., Peng H.** High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, 2018, Vol. 12, No. 1, Pp. 2074–2089. DOI: 10.1214/18-EJS1439

24. **Athey S., Tibshirani J., Wager S.** Generalized random forests. *arXiv:1610.0171*, 2016. DOI: 10.48550/arXiv.1610.01271

25. **Zhang W., Le T.D., Liu L., Zhou Z.-H., Li J.** Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 2017, Vol. 33, No. 15, Pp. 2372–2378. DOI: 10.1093/bioinformatics/btx174

26. **McFowland III E., Somanchi S., Neill D.B.** Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. *arXiv:1803.09159*, 2018. DOI: 10.48550/arXiv.1803.09159

27. **Künzel S.R., Sekhon J.S., Bickel P.J., Yu B.** Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences U.S.A.*, 2019, Vol. 116, No. 10, Pp. 4156–4165. DOI: 10.1073/pnas.1804597116

28. **Curth A., van der Schaar M.** Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2021, Vol. 130, Pp. 1–11.

29. **Du X., Fan Y., Lv J., Sun T., Vossler P.** Dimension-free average treatment effect inference with deep neural networks. *arXiv:2112.01574*, 2021. DOI: 10.48550/arXiv.2112.01574

30. **Qin T., Wang T.-Z., Zhou Z.-H.** Budgeted heterogeneous treatment effect estimation. *Proceedings of the 38th International Conference on Machine Learning*, 2021, Vol. 139, Pp. 8693–8702.

31. **Guo Z., Zheng S., Liu Z., Yan K., Zhu Z.** CETransformer: Casual effect estimation via transformer based representation learning. *Pattern Recognition and Computer Vision: 4th Chinese Conference*, 2021, Vol. IV, Pp. 524–535. DOI: 10.1007/978-3-030-88013-2_43

32. **Melnychuk V., Frauen D., Feuerriegel S.** Causal transformer for estimating counterfactual outcomes. *arXiv:2204.07258*, 2022. DOI: 10.48550/arXiv.2204.07258

33. **Zhang Y.-F., Zhang H., Lipton Z.C., Li L., Xing E.P.** Exploring transformer backbones for heterogeneous treatment effect estimation. *arXiv:2202.01336*, 2022. DOI: 10.48550/arXiv.2202.01336

34. **Imbens G.W.** Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 2004, Vol. 86, No. 1, Pp. 4–29.

35. **Park J., Shalit U., Scholkopf B., Muandet K.** Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. *Proceedings of the 38th International Conference on Machine Learning*, 2021, Vol. 139, Pp. 8401–8412.

36. **Xu K., Fukuchi K., Akimoto Y., Sakuma J.** Statistically significant concept-based explanation of image classifiers via model knockoffs. *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 2023, Pp. 519–526. DOI: 10.24963/ijcai.2023/58

37. **Rubin D.B.** Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, Vol. 100, No. 469, Pp. 322–331. DOI: 10.1198/016214504000001880

38. **Shi C., Blei D.M., Veitch V.** Adapting neural networks for the estimation of treatment effects. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, Pp. 2507–2517.

39. **Bahdanau D., Cho K., Bengio Y.** Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014. DOI: 10.48550/arXiv.1409.0473

40. **Bodria F., Giannotti F., Guidotti R., Naretto F., Pedreschi D., Rinzivillo S.** Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 2023, Vol. 37, Pp. 1719–1778. DOI: 10.1007/s10618-023-00933-9

41. **Burkart N., Huber M.F.** A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 2021, Vol. 70, Pp. 245–317. DOI: 10.1613/jair.1.12228

42. **Ribeiro M.T., Singh S., Guestrin C.** "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, Pp. 1135–1144. DOI: 10.1145/2939672.2939778

43. **Garreau D., von Luxburg U.** Explaining the explainer: A first theoretical analysis of LIME. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics* (*AISTATS*), 2020, Vol. 108, Pp. 1–9.

44. **LeCun Y., Bottou L., Bengio Y., Haffner P.** Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, Vol. 86, No. 11, Pp. 2278–2324. DOI: 10.1109/5.726791

45. **Kirpichenko S.R., Utkin L.V., Konstantinov A.V., Verbova N.M.** Survival concept-based learning models. *Journal of Intelligent Information Systems*, 2025, Vol. 63, Pp. 1687–1711. DOI: 10.1007/s10844-025-00958-0

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Lev V. Utkin
Уткин Лев Владимирович
E-mail: lev.utkin@gmail.com
ORCID: https://orcid.org/0000-0002-5637-1420

Andrei V. Konstantinov
Константинов Андрей Владимирович
E-mail: andrue.konst@gmail.com
ORCID: https://orcid.org/0000-0002-1542-6480

Natalia M. Verbova
Вербова Наталья Михайловна
E-mail: nag00@mail.ru
ORCID: https://orcid.org/0000-0001-8749-9470

Submitted: 18.11.2025; Approved: 13.12.2025; Accepted: 28.12.2025.
Поступила: 18.11.2025; Одобрена: 13.12.2025; Принята: 28.12.2025.