IMO: Greedy Layer-Wise Sparse Representation Learning for Out-of-Distribution Text Classification with Pre-trained Models

Tao Feng, Lizhen Qu [*], Zhuang Li, Haolan Zhan, Yuncheng Hua, Gholamreza Haffari
Monash University, Australia
{firstname.lastname}@monash.edu

Abstract

Machine learning models have made incredible progress, but they still struggle when applied to examples from unseen domains. This study focuses on a specific problem of domain generalization, where a model is trained on one source domain and tested on multiple target domains that are unseen during training. We propose IMO: **I**nvariant features **M**asks for **O**ut-of-Distribution text classification, to achieve OOD generalization by learning domain-invariant features. During training, IMO employs a greedy algorithm to learn sparse representations for each layer in a top-down manner. It performs better than the opposite direction and learning of sparse representations for all layers simultaneously. Our comprehensive experiments show that IMO substantially outperforms strong baselines such as prompt-based methods and large language models, in terms of various evaluation metrics and settings. [1]

1 Introduction

When deploying natural language processing (NLP) models trained on labeled data in the wild, it is well known that their predictive performance declines significantly on samples drawn from distributions that differ from their training data (Wang et al., 2021b). Although various domain adaptation (DA) methods have been proposed (Liu et al., 2022; Saunders, 2022), they assume the availability of labeled or unlabeled data from target domains, along with information about target domains. However, for many real-world applications, especially for early-stage businesses, users may apply their models to arbitrary data so the test data may well be Out-of-Distribution (OOD). Hence, target domain information may not be available for DA. In addition, some training datasets are expensive to acquire so they are available only in one domain. Therefore, this work focuses on single-source *domain generalization* (DG) for text classification, which aims to enable classifiers trained in *one* source domain to *robustly* work on the same classification tasks in any unseen OOD data without any model tuning.

Pre-trained large language models (LLMs) have drawn a lot of attentions due to their strong predictive performance across a variety of tasks. Although generative models or classifiers built on top of pre-trained LLMs outperform prior models in multiple domains, their performance is still not *robust* on tasks when the testing distribution differs substantially from the training distribution (Bang et al., 2023). Recent works (Wang et al., 2021a; Feng et al., 2023; Veitch et al., 2021) show that one of the key reasons is *spurious correlations*, which refer to the correlations between features and model outputs that are not based on causal relationships.

To take a step towards "train it once, apply it anywhere", we propose a novel greedy layer-wise **I**nvariant **M**asking technique for **O**OD text classification, coined IMO, which selects domain-invariant features and key token representations from appropriate layers of a pre-trained deep transformer encoder to mitigate spurious correlations. The resulting hidden representations are sparse from the top layer to a specific layer of the pre-trained model. We demonstrate the effectiveness of this technique through theoretical justifications and extensive experiments. Similar to (Zhang et al., 2021) on computer vision tasks, we shed light on how to apply sparsity as an effective inductive bias to deep pre-trained models for OOD text classification. Our contributions are:

- We propose IMO, a novel top-down greedy layer-wise sparse representation learning method for pre-trained text encoders for robust OOD classification by sharply reducing task-specific spurious correlations. In com-

[*] Corresponding Author.

[1] Codes are available at https://github.com/WilliamsToTo/IMO.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2625–2639
August 11-16, 2024 ©2024 Association for Computational Linguistics

parison with bottom-up layer-wise and simultaneous search across all layers, we discover that the top-down greedy search is decisive for performance improvement.

- We develop a theoretical framework that elucidates the relationship between domain-invariant features and causal features. Additionally, we provide an explanation of how our method learns invariant features.

- Our comprehensive experimental results show that: (i) using IMO with BART (Lewis et al., 2020) significantly outperforms competitive baselines, including CHATGPT, on topic classification and sentiment polarity prediction in most of the target domains. Notably, CHATGPT has 10 times more parameters than BART; (ii) using IMO with CHATYUAN (Clue-AI, 2023) achieves superior performance in Chinese social factor classification compared to strong competitors like CHATGPT; (iii) IMO achieves robust OOD performance w.r.t. varying training data size. The accuracy difference between using 1k and 3.5 million training instances using IMO is less than 6%.

2 Related Work

Domain Generalization. Numerous DG methods have been proposed in the past decade, and most of them are designed for multi-source DG (Chattopadhyay et al., 2020; Zhao et al., 2020; Ding et al., 2022; Zhang et al., 2022; Lv et al., 2022). Existing DG methods can be roughly classified into two categories: invariant representation learning and data augmentation. The key idea of the former is to reduce the discrepancy between representations of source domains (Muandet et al., 2013; Li et al., 2018a,b; Shao et al., 2019; Arjovsky et al., 2020). The key idea of data augmentation is to generate out-of-distribution samples, which are used to train the neural network with original source samples to improve the generalization ability (Xie et al., 2020; Wei and Zou, 2019; Volpi and Murino, 2019).

This paper focuses on single-source DG, where the model is trained on a single source domain, then evaluated on multiple unseen domains. Wang et al. (2021c) proposes a style-complement module to synthesize images with unseen styles, which are out of original distributions. Qiao et al. (2020) proposes adversarial domain augmentation to encourage semantic consistency between the augmented and source images in the latent space. Ouyang et al. (2023) uses a causality-inspired data augmentation approach to encourage network learning domain-invariant features. In terms of text classification, Ben-David et al. (2022); Jia and Zhang (2022) apply prompt-based learning methods to generate a prompt for each sample, then use large language models to predict labels.

Causal Representation Learning (CRL). CRL addresses OOD generalization by exploring causal features that lead to labels. It is based on the assumption that causal features are stable across different environments or data selections. Since CRL is very ambitious and even infeasible in real application, a more practical method is invariant representation learning. Peters et al. (2016) investigated that invariant features, to some extent, infer the causal structure. Arjovsky et al. (2020) also assumes that prediction conditioned on invariant features is stable under different environments. Following such assumption, a strand of methods tries to learn invariant features by mitigating spurious correlated features, which vary across environments (Muandet et al., 2013; Chattopadhyay et al., 2020; Asgari et al., 2022; Izmailov et al., 2022; Hu et al., 2022b). This paper also follows this thread of methods, where we treat features that don't affect prediction as spurious correlated features.

3 Learning Sparse Domain-Invariant Representations

LLMs are pre-trained on large-scale corpora so that they can capture rich correlations between tokens across various domains. To enable trained models incorporating LLMs to work across domains, our key idea originates from the *Invariance Assumption* that the conditional distributions of labels conditioned on invariant features do not change across domains (Peters et al., 2016). Zhang et al. (2021) show that there is a subnetwork inside a full network that can achieve better OOD performance than the full network, if this assumption holds. This hypothesis is also referred to as the functional lottery ticket (Liang et al., 2021). For a specific classification task, such as sentiment polarity analysis, the assumption indicates that there are certain sparse representations that are potential *causes* of labels (Wang and Jordan, 2022) across domains. Our method IMO realizes this idea by constructing sparse domain-invariant representations from the hidden representations of the selected layers of pre-trained transformer-based encoders.

Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the label space, a *domain* is characterized by a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$. In the context of a single source DG, we have access to the data of one source domain $\mathcal{S} = \{(x^s, y^s)\}$ drawn from its joint distribution $P^{\mathcal{S}}_{XY}$. The goal is to learn a predictive model $f : \mathcal{X} \to \mathcal{Y}$ using only the data sampled from $P^{\mathcal{S}}_{XY}$ to minimize the prediction error on K unseen target domains, each of which is associated with a joint distribution P^k_{XY}. Due to domain drifts, $P^{\mathcal{S}}_{XY} \neq P^k_{XY}, \forall k \in 1, ..., K$.

Following (Quinzan et al., 2023), we make the same assumptions that (i) $Y = f(\text{Pa}(Y)) + \epsilon$, where $\text{Pa}(Y)$ denote the features that directly cause Y, (ii) ϵ is exogenous noise, independent of any features, and (iii) Y has no direct causal effect on any features because classification labels are assigned after observing the corresponding texts. Although $P^{\mathcal{S}}_{XY} \neq P^{(k)}_{XY}, \forall k \in 1, ..., K$, we show in §3.3 that under all above assumptions, there is a sparse representation $\mathbf{H}_i$ such that the function $Y = f(\mathbf{H}_i) + \epsilon$ exists in both source and target domains. We empirically study the presence of invariant representations and influence of spurious correlations in §4.3.

As illustrated in Figure 1, our method constructs sparse domain-invariant representations at both feature and token levels in a top-down manner. At the feature level, given embeddings produced by the transformer block of the top layer, a parametric mask layer identifies invariant features from the embeddings. Then, the mask layer is frozen and the algorithm learns the mask layer for the lower layer. The process is repeated until a pre-specified layer is reached. At the token level, a soft attention mechanism incorporates the selected features from the top layer to identify the tokens strongly correlated with Y and use attention weights to create aggregated sparse representations based on the selected features for binary classification. For multi-class classification tasks, a sparse representation is created for each class so that each of them can focus on class-specific information. The model is regularized during training to increase the divergences of the representations between classes.

3.1 Extraction of Invariant Features

Given a text input $X = [x_i]^T_{i=0}$, where x_i is a token in X, a transformer-based pre-trained language model is employed to convert x_i to a continuous token representation. We use hidden states produced by each transformer layer l as token representations, denoted as $\boldsymbol{H}^l = [\boldsymbol{h}^l_i]^T_{i=0}$. $\boldsymbol{h}^l_i$ embeds both invariant features (useful for prediction in different domains) and spuriously correlated features (irrelevant for prediction) produced by layer l. Based on the Invariance Assumption, the invariant features $\boldsymbol{h}^*$ ensure $p^k(Y|\boldsymbol{h}^*)$ to be the same for each domain k. In a transformer layer l, the spuriously correlated features are filtered out by performing element-wise multiplication between token representation $\boldsymbol{h}^l_i$ and a learnable mask $\boldsymbol{m}^l$.

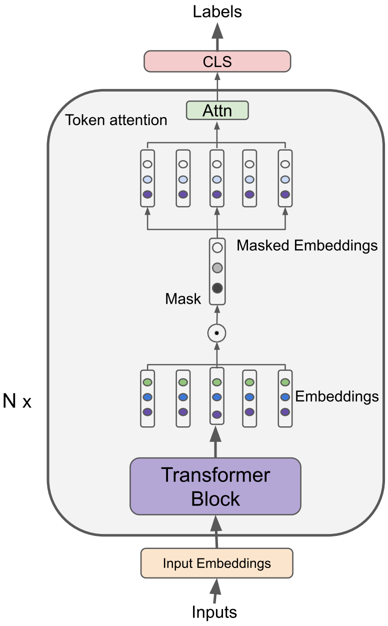


Figure 1: The overall architecture of our method IMO.

A parametric filtering vector $\mathbf{m} = \mathbf{r} \odot \mathbf{q}$ contains zero and non-zero elements, where we define a trainable weight vector $\mathbf{r} \in \mathbb{R}^d$ and a trainable pruning threshold vector $\mathbf{s} \in \mathbb{R}^d$. A unit step function $g(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases}$ is applied to get a binary mask $\mathbf{q} = g(|\mathbf{r}| - \mathbf{s})$. By applying element-wise multiplication $\mathbf{e}^l_i = \mathbf{h}^l_i \odot \mathbf{m}^l$, the zero elements of $\mathbf{m}$ remove corresponding features in token embeddings $\mathbf{h}^l$, while non-zero elements characterize the importance of corresponding features (Liu et al., 2020).

As the unit step function g is not differentiable, we approximate its derivative by using the derivative estimator proposed in (Xu and Cheung, 2019) such that all parameters of a mask layer are trainable by using back-propagation and the family of stochastic gradient descent algorithms,

$$\frac{d}{dt}g(t) = \begin{cases} 2-4|t|, & -0.4 \leq t \leq 0.4 \\ 0.4, & 0.4 \leq |t| \leq 1 \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Following (Xu and Cheung, 2019; Liu et al., 2020), we add a sparse regularization term L_{sparse} to the training loss to encourage the sparsity of mask layers:

$$\mathcal{L}_{sparse} = \sum_{i=1}^{N} \exp(-\boldsymbol{s}_i), \boldsymbol{s} \in \mathbb{R}^d \quad (2)$$

where $\exp(-\boldsymbol{s}_i)$ encourages high (but not extremely large) thresholds. A higher threshold leads to removal of more features. During *inference*, we retain the mask layers to retain invariant features while discarding irrelevant ones.

3.2 Identification of Invariant Tokens

Given a long token sequence, not all information is useful for target tasks. For example, function words, such as 'the', or 'that', provide little information for predicting sentiment polarity. Thus, we employ a token-level attention mechanism to focus on important tokens. Instead of using all features of a token representation, we compute attention scores by using only the invariant features. The proposed attention mechanism differs slightly between binary and multi-class classification.

Binary Classification. For binary classification, we treat the filtering vector $\mathbf{m}^L$ from the last layer L as the query vector and compute the attention weight by performing the matrix product between $\mathbf{m}^L$ and each token embedding from the last layer $\mathbf{e}_i^L$: $a_i = \mathbf{m}^L\mathbf{e}_i^L$. Here, the filtering vector and token embeddings are interpreted as matrices, with $\mathbf{m}^L \in \mathbb{R}^{1\times d}$ and $\mathbf{e}_i^L \in \mathbb{R}^{d\times 1}$. For an input token sequence, we aggregate the masked token embeddings to obtain a sequence representation $\mathbf{v} = \sum_i^T a_i\mathbf{e}_i^L$, where $\mathbf{v} \in \mathbb{R}^{1\times d}$. Finally, the sequence representation is fed into a fully-connected layer, followed by generating a distribution over the label space as follows: $\hat{\mathbf{y}} = \text{softmax}(\mathbf{vP})$.

Multi-class Classification. For the multi-class classification task, we propose using multiple mask layers $\boldsymbol{m}_y^L$ in the last layer L to capture corresponding features and tokens for labels $\boldsymbol{y}$. The number of mask layers equals the number of labels. Each label has its own attention weights $\boldsymbol{a}_y^L = \boldsymbol{m}_y^L\boldsymbol{e}$, and its own representation $\boldsymbol{v}_y^L = \sum_i^T a_{yi}^L\boldsymbol{e}_i$. Instead of using a fully-connected layer, we use a learnable weight vector per class to project $\boldsymbol{v}^L$ to a scalar: $c^L = \boldsymbol{v}^L\boldsymbol{p}^L$, where $\boldsymbol{v}^L \in \mathbb{R}^{1\times d}$ and $\boldsymbol{p}^L \in \mathbb{R}^{d\times 1}$. The rationale behind this is that each class should have its own weight vector and hidden representations for encoding class-specific information. Then, we concatenate these scalars to a vector $\boldsymbol{c} = [c^L]$, and compute the predictive distribution by $\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{c})$.

To encourage mask layers to extract label-specific features, we propose the following regularization term to penalize pairwise cosine similarities between the corresponding mask layers (where N is the number of label-specific mask layers):

$$\mathcal{L}_{dist} = \frac{1}{N(N-1)} \sum_{i \neq j} \cos(\boldsymbol{m}^i, \boldsymbol{m}^j). \quad (3)$$

Training Procedure. Rather than training all mask layers simultaneously, we adopt a layer-wise training procedure to train them sequentially from the top layer to the bottom layer. As illustrated in Figure 1, for each layer, a new filtering layer, $\mathbf{m}^{L-i}$, is introduced on the top of the $(L-i)$-th transformer layer, with $i \in \{0, 1, 2, ...L-1\}$. Crucially, during this phase, the previously trained mask layers remain frozen to preserve their learned parameters. Upon each layer's training completion, the model is stored as $\theta_{L:L-i}$. This iterative procedure continues until the training of the most bottom filtering vector, $\mathbf{m}^1$, is completed. Consequently, a suite of models, ranging from θ_L to $\theta_{L:1}$, is collected. We empirically determine the model's efficacy by evaluating its performance on the validation set from the source domain. The best-performing model is chosen as the model to test on the target domains.

Objective Function. During training, the overall objective for binary classification is to (1) have good predictive performance on classification tasks and (2) maximize sparsity in mask layers to only keep invariant features. When training mask at layer l, the loss function is:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha\mathcal{L}_{sparsity}^l \quad (4)$$

where $\mathcal{L}_{ce}$ denotes the cross entropy loss and f denotes the predictive model. α, where $\alpha > 0$, is a hyperparameter that controls the balance between predictive performance and sparsity in mask layers. $\mathcal{L}_{sparsity}^l$ is the sparse regularization term for mask at layer l.

For multi-class classification, we add a distance regularization term:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha\mathcal{L}_{sparsity}^l + \beta\mathcal{L}_{dist} \quad (5)$$

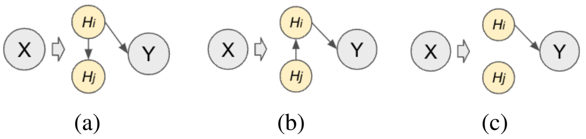


Figure 2: Illustration of potential causal graphs between the variables H_i, H_j of two features (encoded from an input X) and a target variable Y.

The hyperparameter β serves to calibrate the equilibrium between features specific to individual labels and those shared across multiple labels.

3.3 Theoretical Analysis

Based on our assumptions, $Y = f(\mathbf{H}_i) + \epsilon$ exists, when $\mathbf{H}_i$ are the parent nodes of Y in the underlying causal graph. Because $\mathbf{H}_i$ are a subset among all possible hidden representations correlated with Y, there should be a subset of hidden representations serving as parents of Y, otherwise the invariance assumption does not hold. Due to the widely used faithfulness assumption stating that statistical independences imply the corresponding causal structures (Neal, 2020), we aim to find out $\mathbf{H}_i \not\perp\!\!\!\perp Y|\mathbf{H}_j$, where $\mathbf{H}_j$ is any feature set non-overlapped with $\mathbf{H}_i$.

We start our theoretical analysis by introducing a sparsity regularization term $\Omega(Y, H_i, ..., H_j)$, which counts the number of edges between Y and the random variables of features in an underlying causal graph, where Y is the variable for labels and H_k denotes the random variable of the feature h_k. Then we introduce a loss function $\mathcal{L}_\Omega(Y, H_i, ..., H_j) = \mathcal{L}_{ce} + \alpha\Omega(Y, H_i, ..., H_j)$ with $\alpha > 0$, analogous to Eq. (4).

Considering the simplest case that there is only a causal feature h_i and a non-causal feature h_j, the corresponding random variables are denoted by H_i and H_j. From any causal graphs in Fig. 2, we conclude that $p(Y|H_i, H_j) = p(Y|H_i)$ so that the cross entropy term in $\mathcal{L}_\Omega$ remains the same when using the term $p(Y|H_i)$, but the loss decreases after removing the non-causal feature from the loss due to the regularization term $\Omega(Y, H_i, H_j)$.

The two feature case can be easily extended to the case having more than two features. It is trivial that excluding a non-causal feature from the loss $\mathcal{L}_\Omega$ leads to the decrease of $\mathcal{L}_\Omega$ due to the Markov property of causal graphs (Peters et al., 2017).

Corollary 1. *If there is no edge between Y and H_k in a causal graph, then $\mathcal{L}_\Omega(Y, H_i, ..., H_j) < \mathcal{L}_\Omega(Y, H_i, ..., H_j, H_k)$.*

During training, we start with a loss $\mathcal{L}_\Omega(Y, H_1, ..., H_N)$ with a complete set of features. If a non-causal feature H_k is removed, $\mathcal{L}_\Omega(Y, H_i, ..., H_j)$ decreases according to Corollary 1. In contrast, if a causal feature H_k is removed, the cross entropy term increases because the mutual information $I(Y; H_k|H_i, ..., H_j) > 0$. Namely, H_k adds additional information for predicting Y. However, in that case, $\mathcal{L}_\Omega(Y, H_i, ..., H_j)$ may still decrease if the increase of $\mathcal{L}_{ce}$ is smaller than the decrease of the regularization term $\alpha\mathcal{L}_\Omega(Y, H_i, ..., H_j)$, where $\alpha > 0$. The exceptional case can be mitigated if α is sufficiently small. As a result, the loss $\mathcal{L}_\Omega$ provides an effective way to guide the search for the features serving as the causes of the labels, although we cannot recover the underlying true causal graphs. Herein, the loss (4) is a surrogate of $\mathcal{L}_\Omega(Y, H_i, ..., H_j)$ by using a deep neural network.

4 Experiments

We show that our approach significantly outperforms the competitive baselines in almost all settings, followed by empirically verifying that domain-invariant sparse representations indeed exist and spurious features deteriorate model performance in Sec. 4.3, as well as justifying the effectiveness of top-down greedy search strategy and individual modules in the ablation study.

4.1 Experimental Setup

Tasks and Datasets We evaluate our method on binary and multi-class classification tasks. Herein, we adopt *accuracy* as the metric for binary sentiment polarity classification and *macro-F1* for multi-class classification tasks. All models are trained with five different random seeds to assess the statistical significance.

The datasets for binary sentiment analysis include Amazon Review Polarity (Zhang et al., 2015a), Yelp Review Polarity (Zhang et al., 2015a), IMDB (Maas et al., 2011), TweetEval Sentiment (Barbieri et al., 2020) [2] and Yahoo! Answers Sentiment (Li et al., 2019). For multi-class classification, we consider topic classification task in AG News dataset (Gulli, 2005; Del Corso et al., 2005; Zhang et al., 2015b) and social factor prediction task in SocialDial (Zhan et al., 2023, 2024). More details about datasets can be found in Appendix A.2.

[2] We remove all neutral instances to turn it into a binary classification task.

	IMDB→			Amazon→			Yelp→			TweetEval→			
Models	**Amazon**	**Yelp**	**TweetEval**	**IMDB**	**Yelp**	**TweetEval**	**IMDB**	**Amazon**	**TweetEval**	**IMDB**	**Yelp**	**Amazon**	**Avg.**
BERT	89.77*	87.12*	78.52*	88.09*	92.18*	83.75*	86.98*	92.10*	87.55*	82.59*	84.87*	86.80*	86.69*
BART	89.91*	88.01*	68.47*	87.93*	91.01*	82.98*	86.44*	91.97*	88.21*	78.21*	89.51*	87.01*	85.80*
BERT-EDA	87.73*	87.47*	72.10*	88.89*	92.43*	86.40*	88.11*	92.98*	87.92*	81.64*	85.82*	87.77*	86.61*
BERT-UDA	87.76*	87.02*	70.23*	89.87*	93.78*	86.37*	86.89*	92.81*	84.91*	82.83*	85.95*	87.29*	86.31*
BERT-PGB	88.40*	83.61*	70.51*	89.70*	93.66*	86.19*	86.09*	92.72*	87.95*	81.88*	85.13*	87.54*	86.11*
PADA	85.73*	89.84*	88.40	84.47*	93.96	85.92*	87.71*	91.42*	90.33	80.30*	84.69*	90.61	87.78*
PDA	89.35*	90.59*	87.71*	88.16*	94.20	85.61*	88.17*	93.59	89.88*	82.05	86.37	86.41	88.51*
CHATGPT	91.08	92.06	81.01	90.50	92.06	81.01	**90.50**	91.08	81.01	**90.50**	92.06	91.08	88.66
ALPACA-7B	90.14	92.30	88.66	83.01	92.30	88.66	83.01	90.14	88.66	83.01	92.30	90.14	88.52
ALPACA-7B-LoRA	89.80	82.80	87.77	81.00	82.80	87.77	81.00	89.80	87.77	81.00	82.80	89.80	85.34
IMO-BART (Our)	**93.97**	**94.63**	**89.58**	**90.86**	**95.14**	**91.08**	90.08	**94.87**	**91.62**	85.39	**92.84**	91.66	**91.81**
IMO-BART B2T	75.86*	75.37*	71.90*	73.27*	73.74*	72.58*	72.90*	73.47*	72.06*	69.74*	73.29*	75.81*	73.33*
IMO-BART w/o **sq**	74.88*	76.41*	67.97*	70.47*	72.33*	71.98*	71.59*	72.30*	71.73*	71.25*	71.62*	70.63*	71.93*
IMO-BART last	91.71*	92.82*	89.01	89.41	93.01*	89.85*	89.67	93.51	90.10*	84.69*	91.22*	90.95*	90.49*

Table 1: Single-source domain generalization on sentiment analysis datasets. “B2T”: bottom-up layer-wise search. “**w/o sq**”: simultaneous search. “last”: applying the mask on only the last layer. The metric is accuracy. Asterisk * shows a significant difference compared to IMG-BART using a t-test with a $p \leq 0.05$.

	AG News		
Models	**Title → Desc**	**Desc → Title**	**Avg-F1**
BERT	81.11*	67.95*	74.68*
BART	80.12*	71.22*	75.96*
BERT-EDA	80.52*	72.10*	76.58*
BERT-UDA	80.41*	71.81*	75.82*
BERT-PGB	78.53*	73.51*	76,02*
PADA	82.39*	75.52*	78.96*
PDA	83.61*	75.96*	79.79*
CHATGPT	85.13	79.28	82.21
ALPACA-7B	70.61	70.44	71.49
ALPACA-7B-LoRA	56.17	49.44	52.81
IMO-BART (Our)	**89.40**	**81.97**	**85.68**
IMO-BART B2T	70.31*	64.59*	67.45*
IMO-BART w/o **sq**	62.59*	57.27*	59.93*
IMO-BART last	88.22	80.05*	84.13*

Table 2: Results for multi-class classification datasets. ‘Desc’ represents description. The metric is macro F1.

	SocialDial			
Models	**Loc (Synthetic) → Loc (Human)**	**SD (Synthetic) → SD(Human)**	**SR (Synthetic) → SR(Human)**	**Avg- F1**
BERT-zh	18.11*	35.05*	32.39*	28.51*
CHATYUAN	18.23*	34.94*	33.92*	29.03*
BERT-EDA	13.98*	35.71*	26.38*	25.36*
BERT-UDA	15.20*	33.59*	27.03*	25.27*
CHATGPT	21.44	38.46	35.12	31.67
CHATGLM-6B	20.57	20.53	11.55	17.55
IMO-CY (Our)	**23.22**	**46.04**	**42.71**	**37.32**
IMO-CY B2T	14.31*	30.29*	32.45*	25.68*
IMO-CY w/o **sq**	13.37*	29.81*	29.05*	24.07*
IMO-CY last	21.47*	44.73	39.89*	35.36*

Table 3: Evaluation results on SocialDial dataset. CY represents the pre-trained language model CHATYUAN. Loc represents Location; SD represents Social Distance; SR represents Social Relation. The metric is macro F1.

Baseline Models. As Gulrajani and Lopez-Paz (2021) showed, simple empirical risk minimization (ERM) outperforms many SOTA domain generalization algorithms. So we finetune **BERT** (Devlin et al., 2019) and encoder of **BART** (Lewis et al., 2020) using cross-entropy loss as baselines. For Chinese text classification, we use **BERT-zh** (Devlin et al., 2019), **BART-zh** (Shao et al., 2021) and **CHATYUAN** (Clue-AI, 2023).

Domain Generalization Models. **PADA** (Ben-David et al., 2022) is an example-based autoregressive prompt learning algorithm for domain generalization based on the T5 language model (Raffel et al., 2020). **PDA** (Jia and Zhang, 2022) is a prompt-based learning algorithm for domain generalization.

Large Language Models. As CHATGPT shows promising zero-shot ability on various NLP tasks (OpenAI, 2023), we treat **CHATGPT** (gpt-3.5-turbo) as a baseline. **ALPACA-7B** (Taori et al., 2023) is another baseline, which is finetuned from LLaMA 7B (Touvron et al., 2023) on 52K instruction-following data generated by self-instruct (Wang et al., 2022). **ALPACA-7B-LoRA** is a finetuned ALPACA-7B model using low-rank adaptation (Wang, 2023; Hu et al., 2022a). **CHATGLM-6B** (THUDM, 2023) is an open large language model based on General Language Model (Du et al., 2022), optimized for Chinese question-answering and dialogue. All LLMs use few-shot in-context learning. The specific query templates used for the LLMs can be found in Appendix A.3.

Data Augmentation. Wiles et al. (2022); Gokhale et al. (2022) find data augmentation benefit domain generalization tasks. **EDA** (Wei and Zou, 2019) uses four operations (*i.e.,* synonym replacement, random insertion, random swap, and random deletion) to augment text data. **UDA** (Xie et al., 2020) uses back-translation to generate diverse paraphrases while preserving the semantics of the original sentences. **PGB** (Shiri et al., 2023) generates syntactically and lexically diversified paraphrases using a fine-tuned BART.

4.2 Domain Generalization Results

Binary Classification. Table 1 reports the comparisons between our method and the baselines on sentiment polarity classification. Our method using

Models	Yelp→				Amazon→			
	Yelp (Source)	IMDB (Target)	Amazon (Target)	TweetEval (Target)	Amazon (Source)	IMDB (Target)	Yelp (Target)	TweetEval (Target)
IMO	95.94	-5.86	-1.07	-4.32	95.34	-4.48	-0.20	-4.26
IMO- SC	89.01*	-7.81*	-3.88*	-11.20*	90.12*	-7.64*	-3.47*	-12.59*

Table 4: Comparison between the proposed model and model using spurious features (SC). In target datasets, we report the reduced percentage of accuracy compared to the source domains.

BART as backbone (*i.e.,* IMO-BART) achieves superior performance over all baselines in 7 of 12 settings, and outperforms the best baseline ChatGPT by 2.63% on average. Interestingly, ChatGPT stands out as the best model in two out of 12 settings, though it remains unclear whether ChatGPT use those datasets for training. Moreover, it is noteworthy that data augmentation methods (*i.e.,* BERT-EDA, BERT-UDA, BERT-PGB) show slightly inferior performance in comparison to the simple fine-tuning of BERT in terms of average accuracy. This suggests that simply back-translating or paraphrasing instances within source domains does not enhance performance on target domains.

Multi-class Classification. As shown in Table 2 and Table 3, our method outperforms all baselines in terms of average macro-F1 by 3.22% and 5.16% on AG News and SocialDial respectively. Among baselines, ChatGPT exhibits the strongest performance on both datasets and surpasses Alpaca-7B, Alpaca-7B-LoRA, and ChatGLM by a large margin. This superior performance shows that current open-source large language models still have a substantial performance gap with ChatGPT when handling difficult tasks.

4.3 Analysis of Spurious Features

Presence of Invariant Representations. We inspect shared representations at both feature and token levels. Invariant features are expected to have non-zero values across domains. Taking the best performing model IMO-BART in the sentiment analysis as an example, we train the model in each domain respectively and visualize its masks of the top layer in each domains. As depicted in Fig. 3, there are indeed a set of features shared across domains selected by the masks. We further compute Cosine similarities between the filtering vectors m of the top layer trained on different source domains. As shown in Table 8 in Appendix, their similarities range from 0.68 and 0.85. At the token level, we inspect the shared attention weights visualized in Fig.4 (see Appendix A3), which indicate the keywords shared across domains in sentiment analysis, such as “great” and “slow”.

Impact of Spurious Correlations. To study whether our proposed masking mechanism indeed identifies robust features, we compare the performance of using the selected features with the non-selected ones. Specifically, we run additional experiments by replacing the learned binary masks $\mathbf{q}$ with $|1 - \mathbf{q}|$, followed by freezing all parameters except the classification head and training a model using those non-selected features. The results in Table 4 show that models using the non-selected features have an approximate 6% accuracy reduction in source domains and perform worse than using all features. In target domains, the corresponding performance drop using non-selected features is significantly higher than that using both our method as well as using all features. Hence, our masks indeed mitigate the use of spurious features.

4.4 Ablation Study

We compare top-down greedy search with alternative methods: bottom-up layer-wise search (B2T), simultaneous search (w/o **sq**), and only applying a mask on the last layer (last). From Table 1, 2 and 3, we can tell that top-down greedy performs significantly better than the alternative competitors. We conjecture that top-down layer-wise learning serves a regularization method that reduces the risk of loosing crucial features that are well correlated with Y and the corresponding optimization problem is easier to solve than learning all mask layers simultaneously. Representations from higher layers are shown to be more context-specific than lower layer representations (Ethayarajh, 2019). In contrast, the bottom up approach may drop key features in lower layers that significantly contribute to important higher layer features.

We compare variants of IMO by using varying backbone models and removing the corresponding components. For backbones, we compare BART with T5 and BERT, denoted them as **IMO-T5**, and **IMO-BERT**. To study the contribution of each component in our approach, we conduct experiments where we exclude the mask layers, attention mechanisms, or both. These models are denoted by **w/o** $\boldsymbol{m}$, **w/o** $\boldsymbol{a}$, and **w/o** $\boldsymbol{am}$, respectively. The corresponding results are reported in Table 5, Table 11 and Table 10 in Appendix. For comparison between backbones, we find that encoder-decoder neural architectures (*i.e.,* BART, T5) con-

	IMDB→			Amazon→			Yelp→			TweetEval→			
Models	**Amazon**	**Yelp**	**TweetEval**	**IMDB**	**Yelp**	**TweetEval**	**IMDB**	**Amazon**	**TweetEval**	**IMDB**	**Yelp**	**Amazon**	**Avg.**
BART w/o IMO	89.94*	89.13*	69.59*	88.19*	92.20*	82.69*	86.85*	90.64*	85.83*	78.98*	89.25*	87.58*	85.91*
IMO-BART (Our)	**93.97**	**94.63**	**89.58**	**90.86**	**95.14**	**91.08**	**90.08**	**94.87**	**91.62**	**85.39**	**92.84**	**91.66**	**91.81**
IMO-BART w/o m	92.15*	92.49*	85.61*	89.48*	92.97*	88.53*	88.28	92.75	87.44	80.10	89.57	88.09*	88.95*
IMO-BART w/o a	91.35*	91.04*	84.18*	88.51*	92.49*	84.97*	87.10*	91.87*	88.01*	83.31*	90.61*	88.87*	88.52*
IMO-BART STE	91.11*	91.71*	88.05*	88.29*	91.69*	87.09*	88.91*	91.39*	89.12*	82.48*	89.37*	88.50*	88.97*
IMO-BART STR	89.79*	88.97*	72.98*	86.26*	87.48*	79.48*	86.40*	88.31*	77.49*	81.43*	85.13*	82.49*	83.85*
IMO-BART Scalar	87.31*	89.92*	87.34*	87.73*	86.03*	83.41*	87.11*	86.43*	85.94*	81.44*	84.75*	85.41*	86.06*
T5 w/o IMO	87.53*	87.09*	66.37*	86.47*	89.38*	80.68*	84.94*	88.86*	83.78*	76.48*	86.89*	85.53*	83.67*
IMO-T5	93.45	93.88	84.92*	89.23*	93.38*	89.73*	88.27*	93.02*	91.01	81.39*	91.93	89.97*	90.01*
BERT w/o IMO	86.48*	86.19*	66.28*	86.12*	88.91*	81.45*	86.34*	88.34*	83.46*	77.25*	87.34*	84.82*	83.58*
IMO-BERT	92.09*	91.93*	85.34*	88.53*	92.19*	88.17*	87.46*	91.49*	89.55*	79.93*	89.23*	87.73*	88.64*

Table 5: Ablation study on sentiment analysis datasets.

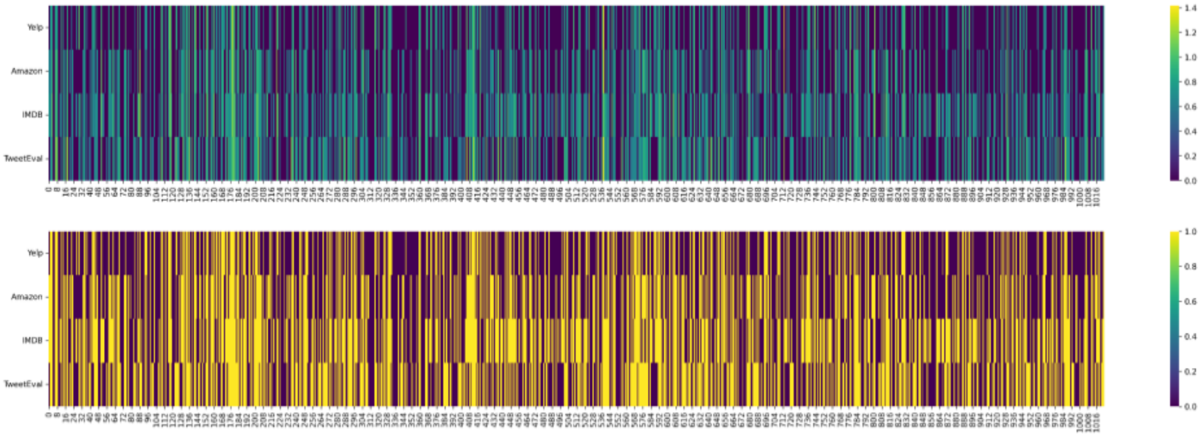


Figure 3: Visualization of filtering and mask vectors in IMO-BART. The top figure visualizes the filtering vectors $\boldsymbol{m}$, while the bottom one visualizes the mask vectors $\boldsymbol{q}$. The x-axis signifies the dimensionality of mask layers, whereas the y-axis denotes values attributed to each dimension.

	Amazon→			
Models	**Yelp**	**IMDB**	**TweetEval**	**Avg.**
IMO-1k	92.21	87.29	85.18	88.22
IMO-10k	94.82	89.11	88.43	90.78
IMO-100k	94.90	90.24	89.01	91.38
IMO-1M	94.95	90.29	89.20	91.48
IMO-3.6M	95.14	90.86	91.08	92.36
IMO- w/o $\boldsymbol{am}$ -1k	70.62	68.61	66.07	68.43
IMO- w/o $\boldsymbol{am}$ -10k	84.88	79.02	75.19	79.70
IMO- w/o $\boldsymbol{am}$ -100k	87.05	84.95	80.48	84.16
IMO- w/o $\boldsymbol{am}$ -1M	91.38	87.06	81.59	86.68
IMO- w/o $\boldsymbol{am}$ -3.6M	92.20	88.19	82.69	87.69

Table 6: Domain generalization experiment with different training sizes in the source domain.

sistently achieve better performance than encoder-only models (*i.e.,* BERT). Compared with variants that remove both the attention module and mask layers, IMO with the attention module or mask module has a significant performance improvement in terms of accuracy or F1 on average, which justifies the usefulness of both modules.

Additionally, we compare IMO with various sparsity methods to implement mask layers, including **STR** (Kusupati et al., 2020), **STE** (Bengio et al., 2013; Liu et al., 2020), and **Scalar**, which uses a learnable single scalar instead of the threshold vector $\boldsymbol{s}$. All those alternative methods lead to a significant drop, as seen in Table 5.

To explore the influence of source domain training data size on performance within target domains, we train models based on BART with and without our method on the Amazon review dataset with varying sizes of training data (*i.e.,* 1k, 10k, 100k, 1M, and 3.6M). The results in Table 6 show that our method depends significantly less on training data size, though more training data can improve the performance overall. Notably, 1k training data yields a remarkable decline for the models without using IMO, while the corresponding performance reduction is significantly less by using our method.

5 Conclusion

This paper presents a novel method, coined IMO, which is a greedy layer-wise representation learning method aiming to improve single-source domain generalization on pre-trained deep encoders for text classification tasks. The key idea is to retain invariant features through trainable mask layers and incorporate a token-level attention module to focus on the tokens that directly lead to the prediction of labels. Through extensive experiments, we demonstrate that IMO achieves superior OOD performance over competitive baselines on multiple datasets. The visualization of masks and attention weights empirically justifies the effectiveness of identified invariant sparse representations.

Limitations

Our work focuses on the text classification task, intending to investigate how to learn invariant features to improve out-of-domain generalization. However, the proposed method has promising potential for domain generalization in various NLP tasks, such as question answering and text generation tasks. Future work may consider more tasks beyond text classification.

It is worth noting that IMO needs to be trained in a large source domain. The size of the source domain should ideally exceed 10,000 samples to achieve consistently good performance. However, this requirement may pose challenges in low-resource learning scenarios.

Ethics Statement

This research is dedicated to augmenting the reliability and safety of text classification models, particularly in the context of domain shifts, as highlighted by Ribeiro et al. (2020). By focusing on the learning of invariant features across diverse domains, our approach aims to provide tangible benefits to applications that serve a wide array of user groups. From a user-centric perspective, the implementation of our methodology is expected to bolster the trustworthiness and diminish potential biases in language models.

It is pertinent to note that our study does not involve human subjects, nor does it contravene any legal or ethical standards. We foresee no detrimental impacts arising from our research endeavors. The experimental work underpinning this study was exclusively conducted using datasets that are publicly accessible. Our overarching goal is to foster enhanced academic and societal consciousness regarding the challenges of domain generalization in the field of natural language processing.

Acknowledgement

This material is based on research sponsored by DARPA under agreement number HR001122C0029. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization.

Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. 2022. Masktune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation.

Peter Bühlmann. 2018. Invariance, causality and robustness.

Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. 2020. Learning to balance specificity and invariance for in and out of domain generalization. In *Computer Vision – ECCV 2020*, pages 301–318, Cham. Springer International Publishing.

Clue-AI. 2023. Chatyuan. https://github.com/clue-ai/ChatYuan.

Gianna M. Del Corso, Antonio Gullí, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, page 97–106, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Ding, Lei Wang, Bin Liang, Shuming Liang, Yang Wang, and Fang Chen. 2022. Domain generalization by learning and removing domain-specific features. In *Advances in Neural Information Processing Systems*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Tao Feng, Lizhen Qu, and Gholamreza Haffari. 2023. Less is More: Mitigate Spurious Correlations for Open-Domain Dialogue Response Generation Models by Causal Discovery. *Transactions of the Association for Computational Linguistics*, 11:511–530.

Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. 2022. *Generalized but not Robust?* comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2705–2718, Dublin, Ireland. Association for Computational Linguistics.

A. Gulli. 2005. The anatomy of a news search engine. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, page 880–881, New York, NY, USA. Association for Computing Machinery.

Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ziniu Hu, Zhe Zhao, Xinyang Yi, Tiansheng Yao, Lichan Hong, Yizhou Sun, and Ed Chi. 2022b. Improving multi-task generalization via regularizing spurious correlation. In *Advances in Neural Information Processing Systems*, volume 35, pages 11450–11466. Curran Associates, Inc.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. 2022. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, volume 35, pages 38516–38532. Curran Associates, Inc.

Chen Jia and Yue Zhang. 2022. Prompt-based distribution alignment for domain generalization in text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10147–10157, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization.

Aditya Kusupati, Vivek Ramanujan, Raghav Somani, Mitchell Wortsman, Prateek Jain, Sham Kakade, and Ali Farhadi. 2020. Soft threshold weight reparameterization for learnable sparsity. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5544–5555. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3304–3313, Hong Kong, China. Association for Computational Linguistics.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. 2018a. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018b. Deep domain generalization via conditional invariant adversarial networks. In *Computer Vision – ECCV 2018*, pages 647–663, Cham. Springer International Publishing.

Chen Liang, Simiao Zuo, Minshuo Chen, Haoming Jiang, Xiaodong Liu, Pengcheng He, Tuo Zhao, and Weizhu Chen. 2021. Super tickets in pre-trained language models: From model compression to improving generalization. In *Annual Meeting of the Association for Computational Linguistics*.

Junjie Liu, Zhe XU, Runbin SHI, Ray C. C. Cheung, and Hayden K.H. So. 2020. Dynamic sparse training: Find efficient sparse network from scratch with trainable masked layers. In *International Conference on Learning Representations*.

Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, and Jonghye Woo. 2022. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality inspired representation learning for domain generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8036–8046, Los Alamitos, CA, USA. IEEE Computer Society.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page I–10–I–18. JMLR.org.

Brady Neal. 2020. Introduction to causal inference. *Course Lecture Notes (draft)*.

OpenAI. 2023. Gpt-4 technical report.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. 2023. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(5):947–1012.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Fengchun Qiao, Long Zhao, and Xi Peng. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565.

Francesco Quinzan, Ashkan Soleymani, Patrick Jaillet, Cristian R Rojas, and Stefan Bauer. 2023. Drcfs: Doubly robust causal feature selection. In *International Conference on Machine Learning*, pages 28468–28491. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75.

Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. 2019. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10015–10023.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey.

Fatemeh Shiri, Terry Yue Zhuo, Zhuang Li, Van Nguyen, Shirui Pan, Weiqing Wang, Reza Haffari, and Yuan-Fang Li. 2023. Paraphrasing techniques for maritime qa system.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

THUDM. 2023. Chatglm-6b. `https://github.com/THUDM/ChatGLM-6B`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*.

Riccardo Volpi and Vittorio Murino. 2019. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989.

Eric J. Wang. 2023. Alpaca-lora. `https://github.com/tloen/alpaca-lora`.

Haohan Wang, Zeyi Huang, and Eric Xing. 2021a. Learning robust models by countering spurious correlations.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021b. Generalizing to unseen domains: A survey on domain generalization. pages 4627–4635. Survey Track.

Yixin Wang and Michael Jordan. 2022. Representation learning as finding necessary and sufficient causes. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study.

Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. 2021c. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. 2022. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Zhe Xu and Ray C. C. Cheung. 2019. Accurate and compact convolutional neural networks with trained binarization. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 19. BMVA Press.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, et al. 2024. Renovi: A benchmark towards remediating norm violations in socio-cultural conversations. *arXiv preprint arXiv:2402.11178*.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. Socialdial: A benchmark for socially-aware dialogue systems.

Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron Courville. 2021. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pages 12356–12367. PMLR.

Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P. Xing. 2022. Towards principled disentanglement for domain generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8014–8024.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. page 649–657.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. 2020. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

A Appendix

A.1 Background of Causal Representation Learning

Causal representation learning aims to learn and leverage the causal relations within data to enhance model generalization and robustness against distribution shifts in the data generation process (Shen et al., 2021). This approach differs from traditional machine learning methods that predominantly focus on correlational patterns without distinguishing causation and correlation. Causation refers to the underlying mechanisms that connect variables, implying that alterations in a causal variable will consequentially affect the associated effect variable, a process known as an intervention. In contrast, correlation does not necessarily indicate a direct mechanistic link. For instance, a model might infer ’it is raining’ upon observing people with open umbrellas, which recognizes a correlation between these events. However, the act of closing umbrellas does not influence the weather. This example shows the difference between correlation and causation. Predictions relying on correlation may yield erroneous outcomes when the environment (*i.e.,* data distributions) change. For example, if umbrellas are opened due to sunlight rather than rain, a model trained on the correlation might inaccurately predict rain. This indicates the importance of making predictions based on causation rather than correlation. A causal model should predict the weather based on temperature, humidity, air pressure, etc. Prediction based on causation can enhance out-of-distribution performance, which is supported by the assumption that causal relations remain constant across diverse environments (Bühlmann, 2018).

However, learning the complete causal structure is ambitious and may not be realized in practice. A more feasible approach involves identifying invariant features that reliably predict target variables across varying environments. A series of methods (Muandet et al., 2013; Chattopadhyay et al., 2020; Asgari et al., 2022; Izmailov et al., 2022; Hu et al., 2022b) have been proposed by leveraging the invariance between environments. They leverage the fact that when conditioning all direct causes of a target variable, the conditional distribution of the target will not change when interventions are applied to all other variables in the model except the target itself. Building upon this foundational idea, our work seeks to identify and utilize these direct causes (*i.e.,* invariant features across environments) for the accurate prediction of target variables in the out-of-distribution setting.

A.2 Experiment Datasets

AG News (Gulli, 2005; Del Corso et al., 2005; Zhang et al., 2015b) is a collection of news articles used for topic classification, which contains news titles, and news descriptions assigned to four topic classes. Titles and descriptions are employed as different domains. For social factor prediction, we use SocialDial (Zhan et al., 2023), which is a Chinese socially-aware dialogue corpus consisting of synthetic conversations generated by ChatGPT and human-written conversations. Both are annotated with social factors such as location, social distance, and social relation. Synthetic conversations and human-written conversations are considered as different domains. The statistics of datasets are listed in Table 7.

Binary Classification				
Dataset	**Domain**	**#Train**	**#Dev**	**#Test**
Amazon	Review of products	3.6M	0	40k
IMDB	Review of movies	25k	0	25k
Yelp	Review of businesses	560k	0	38k
TweetEval	Tweet	25k	1k	6k
Yahoo	Questions from Yahoo! Answers	4k	2k	1k

Multi-class Classification				
Dataset	**Domain**	**#Train**	**#Dev**	**#Test**
AG News	Title of news articles	120k	0	7k
AG News	Description of news articles	120k	0	7k
SocialDial	Synthetic conversations by ChatGPT	68k	7k	7k
SocialDial	Human-written conversations	0	0	5k

Table 7: Statistics of datasets.

A.3 Training details

We use the encoder of BART (Lewis et al., 2020) as the default pre-trained language model. All models are trained up to 100 epochs with a minibatch size of 32 in the source domain. We use the Adam (Kingma and Ba, 2015) optimizer with hyperparameters tuned on the validation sets. As a result, we run Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is 5×10^{-5}. We use a linear learning rate scheduler that dynamically decreases the learning rate after a warm-up period. All experiments are conducted on NVIDIA A40 GPU.

The process of model selection in domain generalization is inherently a learning problem. In this approach, we employ training-domain validation, which is one of the three selection methods introduced by Gulrajani and Lopez-Paz (2021). We divide each training domain into separate training and validation sets. Models are trained on the training set, and the model that achieves the highest accuracy on the validation set is chosen as the selected model.

When using large language models to predict target classification labels, the query template for sentiment analysis is: “There are some examples about sentiment analysis: {examples}. Given text: {sentence}, what is the sentiment conveyed? Please select the answer from ‘positive’ or ‘negative’.”. The query template for AG News topic classification is "There are some examples for topic classification: {examples}. Given text: {sentence}, what is the topic of this text? Please select the answer from ‘Business’, ‘Sci/Tech’, ‘World’ or ‘Sports’.” The query templates for SocialDial are “There are some examples for classification: {examples}. Given conversation: {conversation}, what's the location/social distance/social relation of this conversation? Please select the answer from {labels}"[3] (Min et al., 2022; Wang et al., 2023; Yang et al., 2023).

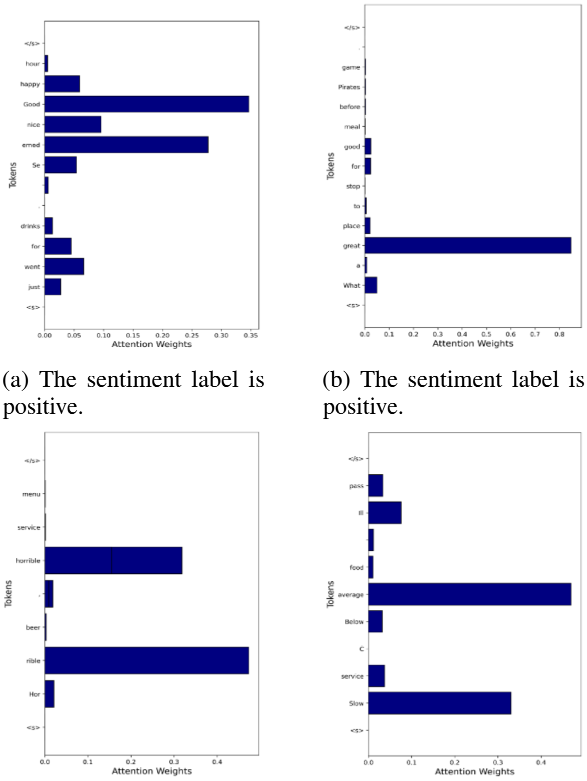
A.4 Visual Explanation

To intuitively show how the attention module and mask module work in models, we visualize attention weights on tokens and mask vectors in Figure 4 and 3, respectively. We also demonstrate cosine similarities between mask vectors $\boldsymbol{m}$ trained on different source domains and Jaccard similarities between binary vectors $\boldsymbol{q}$ trained on different source domains on Table 8 and Table 9, respectively.

From Figure 4, we can find that our model primarily focuses its attention on sentiment-indicative tokens. Notably, positive reviews exhibit high attention weights for tokens like 'good,' 'great,' and 'nice,' indicating their significance. Conversely, negative reviews assign high attention weights to tokens such as 'horrible' and 'slow,' highlighting their importance in expressing negativity.

In Figure 3, we visualize mask vectors $\boldsymbol{m}$ and binary vectors $\boldsymbol{q}$ trained on different source domains across dimensions. It can be observed that certain dimensions are consistently assigned zero (or non-zero) values across different training domains, indicating our mask layers can capture some features that are irrelevant (or invariant) across domains. We quantify invariant features across domains by computing vector similarity. We calculate cosine similarities between different mask vectors $\boldsymbol{m}$. The results are shown in Table 8. We can find that most mask vector pairs have over 0.75 similarity, except the Yelp-TweetEval pair, which is probably because of a larger divergence between Yelp and TweetEval domains. Table 9 shows Jaccard similarities between binary vectors $\boldsymbol{q}$. Most binary vector pairs have similarities of over 0.5, except the Yelp-TweetEval pair, with a similarity of 0.45.

A.5 Additional Experimental Results



(a) The sentiment label is positive.

(b) The sentiment label is positive.

(c) The sentiment label is negative.

(d) The sentiment label is negative.

Figure 4: Visualization of attention weights on tokens in Yelp dataset reviews.

	Yelp	Amazon	IMDB	TweetEval
Yelp	1.0	0.7930	0.7533	0.6838
Amazon	-	1.0	0.8458	0.7687
IMDB	-	-	1.0	0.8069
TweetEval	-	-	-	1.0

Table 8: Cosine similarities between mask vectors $\boldsymbol{m}$ trained on different source domains.

	Yelp	Amazon	IMDB	TweetEval
Yelp	1.0	0.5869	0.5231	0.4504
Amazon	-	1.0	0.6513	0.5614
IMDB	-	-	1.0	0.6139
TweetEval	-	-	-	1.0

Table 9: Jaccard similarities between binary vectors $\boldsymbol{q}$ trained on different source domains.

	SocialDial			
Models	**Loc (Synthetic) → Loc (Human)**	**SD (Synthetic) → SD(Human)**	**SR (Synthetic) → SR(Human)**	**Avg- F1**
CHATYUAN w/o IMO	19.12*	37.75*	34.07*	30.31*
IMO-CY	**23.22**	**46.04**	**42.71**	**37.32**
IMO-CY w/o $\boldsymbol{m}$	22.47*	41.86*	38.95*	34.43*
IMO-CY w/o $\boldsymbol{a}$	21.05*	39.88*	37.28*	32.73*
IMO-CY w/o $\mathcal{L}_{dist}$	20.17*	39.26*	39.41*	32.95*
BART-zh w/o IMO	15.86*	35.92*	31.04*	27.61*
IMO-BART-zh	19.94*	41.39*	39.27*	33.53*
BERT-zh w/o IMO	10.34*	30.17*	19.87*	20.12*
IMO-BERT-zh	14.68*	36.75*	27.41*	26.28*

Table 10: Ablation study on SocialDial datasets. CY represents the pre-trained language model CHATYUAN.

[3] Since SocialDial is a dataset in Chinese, we provided queries translated from Chinese to English for use with CHATGPT.

AG News			
Models	**Title → Desc**	**Desc → Title**	**Avg-F1**
BART w/o IMO	80.91*	73.89*	77.40*
IMO-BART	**89.40**	**81.97**	**85.68**
IMO-BART w/o $\boldsymbol{m}$	83.29*	77.08*	80.19*
IMO-BART w/o $\boldsymbol{a}$	82.72*	77.27*	79.99*
IMO-BART w/o $\mathcal{L}_{dist}$	87.79*	79.82*	83.81*
T5 w/o IMO	78.48*	71.26*	74.87*
IMO-T5	86.91*	79.75*	83.33*
BERT w/o IMO	75.12*	61.47*	68.29*
IMO-BERT	84.79*	75.38*	80.09*

Table 11: Ablation study on AG News dataset. ’Binary’ refers to the application of the proposed binary classification method on multi-label classification tasks.