

Rethinking Misalignment in Vision-Language Model Adaptation from a Causal Perspective

Yanan Zhang^{*}, Jiangmeng Li^{*}, Lixiang Liu, Wenwen Qiang





Motivation

There exists a two-level misalignment between CLIP and downstream tasks, which hinders its adaptation to these tasks.



| Abyssinian | 0.33 |
|------------|------|
| table | 0.25 |
| | |

(a) A motivating example of task misalignment.



(b) A motivating experiment on data misalignment

Two-level misalignment:

(a) task misalignment: caused by the discrepancy between the pre-training objectives of CLIP and the objectives of downstream tasks.

(b) data misalignment: inconsistency exists between the training and testing data.

Problem Analysis



Confounder: the set of task-irrelevant generative factors that are incorrectly retained

Front-Door Adjustment:

$$P(\hat{Y} = y | do(x)) = \sum_{s} \underbrace{P(s|x) \sum_{x'} P(y|x', s) P(x')}_{\text{first term}} \underbrace{\sum_{x'} P(y|x', s) P(x')}_{\text{second term}}.$$

Causality-Guided Semantic Decoupling and Classification



Figure Framework of CDC.

Visual-Language Dual Semantic Decoupling

• Visual: data augmentations

• Language:

$$\mathcal{L}_{de} = \frac{1}{M-1} \frac{1}{C} \sum_{m'=1,m' \neq m}^{M} \sum_{c=1}^{C} \sum_{\bar{c}=1}^{C} P(\bar{c}|w_c^m, w^{m'}) \log P(\bar{c}|w_c^m, w^{m'})$$

$$\mathcal{L}_{con} = -\frac{1}{C} \sum_{m=1}^{M} \sum_{c=1}^{C} \log P(c|w_c^m, w^0)$$

Causality-Guided Semantic Decoupling and Classification



Figure Framework of CDC.

Decoupled Semantic Trusted Classification

$$\begin{split} B^m_c &= \begin{cases} b^1_c, & \text{if } m = 1 \\ \frac{1}{1-C} (B^{m-1}_c b^m_c + B^{m-1}_c u^m + b^m_c U^{m-1}), & \text{if } 1 < m \leq M \end{cases}, \\ U^m &= \begin{cases} u^1, & \text{if } m = 1 \\ \frac{1}{1-C} U^{m-1} u^m, & \text{if } 1 < m \leq M \end{cases}. \end{split}$$

Evaluation

Table 1: The comparison with baseline methods on base-to-novel generalization setting.

| Dataset | CoOp [4] | | CoCoOp [5] | | | MaPLe [6] | | | CDC | | | | |
|----------|--|---|--|--|--|--|--|---|--|--|---|--|--|
| Duiuser | Base | New | HM | Base | New | HM | Base | New | HM | Base | New | HM | Δ |
| Avg | 82.69 | 63.22 | 71.66 | 80.47 | 71.69 | 75.83 | 82.28 | 75.14 | 78.55 | 83.34 | 77.38 | 80.25 | +1.70 |
| ImageNet | 76.47 | 67.88 | 71.92 | 75.98 | 70.43 | 73.10 | 76.66 | 70.54 | 73.47 | 77.50 | 71.73 | 74.51 | +1.04 |
| Caltech | 98.00 | 89.91 | 93.73 | 97.96 | 93.81 | 95.84 | 97.74 | 94.36 | 96.02 | 98.20 | 94.37 | 96.25 | +0.23 |
| Pets | 93.67 | 95.29 | 94.47 | 95.20 | 97.69 | 96.43 | 95.43 | 97.76 | 96.58 | 96.07 | 98.00 | 97.02 | +0.44 |
| Cars | 78.12 | 60.40 | 68.13 | 70.49 | 73.59 | 72.01 | 72.94 | 74.00 | 73.47 | 73.80 | 73.97 | 73.88 | +0.41 |
| Flowers | 97.60 | 59.67 | 74.06 | 94.87 | 71.75 | 81.71 | 95.92 | 72.46 | 82.56 | 96.93 | 75.07 | 84.61 | +2.05 |
| Food | 88.33 | 82.26 | 85.19 | 90.70 | 91.29 | 90.99 | 90.71 | 92.05 | 91.38 | 90.87 | 92.33 | 91.59 | +0.21 |
| Aircraft | 40.44 | 22.30 | 28.75 | 33.41 | 23.71 | 27.74 | 37.44 | 35.61 | 36.50 | 37.47 | 37.50 | 37.48 | +0.98 |
| SUN | 80.60 | 65.89 | 72.51 | 79.74 | 76.86 | 78.27 | 80.82 | 78.70 | 79.75 | 82.37 | 80.03 | 81.18 | +1.43 |
| DTD | 79.44 | 41.18 | 54.24 | 77.01 | 56.00 | 64.85 | 80.36 | 59.18 | 68.16 | 82.70 | 64.10 | 72.22 | +4.06 |
| SAT | 92.19 | 54.74 | 68.90 | 87.49 | 60.04 | 71.21 | 94.07 | 73.23 | 82.35 | 95.10 | 82.33 | 88.26 | +5.91 |
| UCF | 84.69 | 56.05 | 67.46 | 82.33 | 73.45 | 77.64 | 83.00 | 78.66 | 80.77 | 85.70 | 81.73 | 83.67 | +2.90 |
| | Dataset Avg ImageNet Caltech Pets Cars Flowers Food Aircraft SUN DTD SAT UCF | Dataset C Base Base Avg 82.69 ImageNet 76.47 Caltech 98.00 Pets 93.67 Cars 78.12 Flowers 97.60 Food 88.33 Aircraft 40.44 SUN 80.60 DTD 79.44 SAT 92.19 UCF 84.69 | DatasetCoOp [2]BaseNewAvg82.6963.22ImageNet76.4767.88Caltech98.0089.91Pets93.6795.29Cars 78.12 60.40Flowers 97.60 59.67Food88.3382.26Aircraft 40.44 22.30SUN80.6065.89DTD79.4441.18SAT92.1954.74UCF84.6956.05 | CoOp [4]DatasetCoOp [4]BaseNewHMAvg82.6963.2271.66ImageNet76.4767.8871.92Caltech98.0089.9193.73Pets93.6795.2994.47Cars 78.12 60.4068.13Flowers 97.60 59.6774.06Food88.3382.2685.19Aircraft 40.44 22.3028.75SUN80.6065.8972.51DTD79.4441.1854.24SAT92.1954.7468.90UCF84.6956.0567.46 | CoOp [4]CoDatasetCoOp [4]CoBaseNewHMBaseAvg82.6963.2271.6680.47ImageNet76.4767.8871.9275.98Caltech98.0089.9193.7397.96Pets93.6795.2994.4795.20Cars 78.12 60.4068.1370.49Flowers 97.60 59.6774.0694.87Food88.3382.2685.1990.70Aircraft 40.44 22.3028.7533.41SUN80.6065.8972.5179.74DTD79.4441.1854.2477.01SAT92.1954.7468.9087.49UCF84.6956.0567.4682.33 | CoOp [4]CoCoOpBaseNewHMBaseNewAvg82.6963.2271.6680.4771.69ImageNet76.4767.8871.9275.9870.43Caltech98.0089.9193.7397.9693.81Pets93.6795.2994.4795.2097.69Cars 78.12 60.4068.1370.4973.59Flowers 97.60 59.6774.0694.8771.75Food88.3382.2685.1990.7091.29Aircraft 40.44 22.3028.7533.4123.71SUN80.6065.8972.5179.7476.86DTD79.4441.1854.2477.0156.00SAT92.1954.7468.9087.4960.04UCF84.6956.0567.4682.3373.45 | CoOp [4]CoCoOp [5]BaseNewHMBaseNewHMAvg82.6963.2271.6680.4771.6975.83ImageNet76.4767.8871.9275.9870.4373.10Caltech98.0089.9193.7397.9693.8195.84Pets93.6795.2994.4795.2097.6996.43Cars 78.12 60.4068.1370.4973.5972.01Flowers 97.60 59.6774.0694.8771.7581.71Food88.3382.2685.1990.7091.2990.99Aircraft 40.44 22.3028.7533.4123.7127.74SUN80.6065.8972.5179.7476.8678.27DTD79.4441.1854.2477.0156.0064.85SAT92.1954.7468.9087.4960.0471.21UCF84.6956.0567.4682.3373.4577.64 | CoOp [4]CoOp [5]MDatasetNewNewHMBaseNewHMBaseAvg82.6963.2271.6680.4771.6975.8382.28ImageNet76.4767.8871.9275.9870.4373.1076.66Caltech98.0089.9193.7397.9693.8195.8497.74Pets93.6795.2994.4795.2097.6996.4395.43Cars 78.12 60.4068.1370.4973.5972.0172.94Flowers 97.60 59.6774.0694.8771.7581.7195.92Food88.3382.2685.1990.7091.2990.9990.71Aircraft 40.44 22.3028.7533.4123.7127.7437.44SUN80.6065.8972.5179.7476.8678.2780.82DTD79.4441.1854.2477.0156.0064.8580.36SAT92.1954.7468.9087.4960.0471.2194.07UCF84.6956.0567.4682.3373.4577.6483.00 | DatasetCoOp [4]CoCoOp [5]MaPLe [BaseNewHMBaseNewHMBaseNewAvg82.6963.2271.6680.4771.6975.8382.2875.14ImageNet76.4767.8871.9275.9870.4373.1076.6670.54Caltech98.0089.9193.7397.9693.8195.8497.7494.36Pets93.6795.2994.4795.2097.6996.4395.4397.76Cars 78.12 60.4068.1370.4973.5972.0172.94 74.00 Flowers 97.60 59.6774.0694.8771.7581.7195.9272.46Food88.3382.2685.1990.7091.2990.9990.7192.05Aircraft 40.44 22.3028.7533.4123.7127.7437.4435.61SUN80.6065.8972.5179.7476.8678.2780.8278.70DTD79.4441.1854.2477.0156.0064.8580.3659.18SAT92.1954.7468.9087.4960.0471.2194.0773.23UCF84.6956.0567.4682.3373.4577.6483.0078.66 | $\begin{array}{c c c c c c c c c c c c c c c c c c c $ | Dataset CoOp [4] CoCoOp [5] MaPLe [6] Base New HM Base Pase Pase | Dataset CoOp [4] CoCoOp [5] MaPLe [6] CI Base New HM Base New Base New HA Pase Pase Pase Pase Pase Pase Pa | Dataset CoOp [4] CoCoOp [5] MaPLe [6] CDC Base New HM Avg 82.69 63.22 71.66 80.47 71.69 75.83 82.28 75.14 78.55 83.34 77.38 80.25 ImageNet 76.47 67.88 71.92 75.98 70.43 73.10 76.66 70.54 73.47 77.50 71.73 74.51 Caltech 98.00 89.91 93.73 97.96 93.81 95.44 97.74 94.36 96.02 98.20 94.37 96.25 Pets 93.67 95.29 94.47 95.20 97.69 96.43 95.43 97.76 96.58 96.07 98.00 97.02 Cars |

Table 2: Comparison of CDC with recent approaches on cross-dataset evaluation.

| | | Source | Target | | | | | | | | | | | |
|----------------------|---------|----------|---------|-------|-------|---------|-------|----------|-------|-------|-------|-------|-------|--|
| • The cross- | | ImageNet | Caltech | Pets | Cars | Flowers | Food | Aircraft | SUN | DTD | SAT | UCF | Avg | |
| dataset | CoOp | 71.51 | 93.70 | 89.14 | 64.51 | 68.71 | 85.30 | 18.47 | 64.15 | 41.92 | 46.39 | 66.55 | 63.88 | |
| generalization | Co-CoOp | 71.02 | 94.43 | 90.14 | 65.32 | 71.88 | 86.06 | 22.94 | 67.36 | 45.73 | 45.37 | 68.21 | 65.74 | |
| generalization MaPLe | MaPLe | 70.72 | 93.53 | 90.49 | 65.57 | 72.23 | 86.20 | 24.74 | 67.01 | 46.49 | 48.06 | 68.69 | 66.30 | |
| | CDC | 71.76 | 94.47 | 90.77 | 66.27 | 72.67 | 86.27 | 24.50 | 68.07 | 46.60 | 49.13 | 68.60 | 66.73 | |

- We identify the two-level misalignment in adaptation of CLIP.
- We develop a Structural Causal Model and discover the confounder which hinders the estimation of true causal relationships between new samples and their categories.
- Empirically, CDC outperforms state-of-the-art methods on multiple experimental settings.

Thanks !