

An Information-theoretic Multi-task Representation Learning Framework for Natural Language Understanding

Dou Hu^{1,2}, Lingwei Wei^{1*}, Wei Zhou¹, Songlin Hu^{1,2*}

¹ Institute of Information Engineering, Chinese Academy of Sciences

² School of Cyber Security, University of Chinese Academy of Sciences
{hudou, weilingwei, zhouwei, husonglin}@iie.ac.cn

Abstract

This paper proposes a new principled multi-task representation learning framework (InfoMTL) to extract noise-invariant sufficient representations for all tasks. It ensures sufficiency of shared representations for all tasks and mitigates the negative effect of redundant features, which can enhance language understanding of pre-trained language models (PLMs) under the multi-task paradigm. Firstly, a shared information maximization principle is proposed to learn more sufficient shared representations for all target tasks. It can avoid the insufficiency issue arising from representation compression in the multi-task paradigm. Secondly, a task-specific information minimization principle is designed to mitigate the negative effect of potential redundant features in the input for each task. It can compress task-irrelevant redundant information and preserve necessary information relevant to the target for multi-task prediction. Experiments on six classification benchmarks show that our method outperforms 12 comparative multi-task methods under the same multi-task settings, especially in data-constrained and noisy scenarios. Extensive experiments demonstrate that the learned representations are more sufficient, data-efficient, and robust.

Introduction

Multi-task learning (MTL) has become a promising paradigm in deep learning to obtain language representations from large-scale data (Liu et al. 2019a). By leveraging supervised data from related tasks, multi-task learning approaches reduce expensive computational costs and provide a shared representation which is also more efficient for learning over multiple tasks (Wu, Zhang, and Ré 2020; Royer, Blankevoort, and Bejnordi 2023).

Most works (Kendall, Gal, and Cipolla 2018; Chennupati et al. 2019; Liu, Johns, and Davison 2019; Yu et al. 2020; Liu et al. 2021; Lin et al. 2022) on MTL mainly focus on balancing learning process across multiple tasks such as loss-based and gradient-based methods. However, in real-world scenarios, the labeled data resource is limited and contains a certain amount of noise. The fact leads the above task-balanced MTL methods to perform suboptimally and struggle to achieve promising task prediction results.

In recent years, some works (Qian, Chen, and Gechter 2020; de Freitas et al. 2022) introduce the information bottleneck (IB) principle (Tishby, Pereira, and Bialek 1999; Tishby and Zaslavsky 2015) into the information encoding process of multi-task learning to enhance the adaptability to noisy data. Different from vanilla MTL, IB-based MTL methods explicitly compress task-irrelevant redundant information by minimizing the mutual information between the input and the task-agnostic representations during the multi-task encoding process. However, in multi-task scenarios, redundant information often differs across tasks, leading to situations where information beneficial for one task may become redundant for another. For instance, target features that are highly relevant to stance detection may be irrelevant in emotion recognition. Directly applying the IB principle to compress redundancy for one task is prone to losing necessary information for other tasks. As a result, the learned shared representations would not only retain some redundant features but also face the task-specific insufficiency issue.

In this paper, we propose a new principled multi-task representation learning framework, named InfoMTL, to extract noise-invariant sufficient representations for all tasks. It ensures sufficiency of shared representations for all tasks and mitigates the negative effect of redundant features, which can enhance language understanding of pre-trained language models (PLMs) under the multi-task paradigm.

Firstly, we propose a shared information maximization (SIMax) principle to learn more sufficient shared representations for all target tasks. The SIMax principle simultaneously maximizes the mutual information between the input X and the shared representations Z for all tasks, as well as the mutual information between the shared representations Z and the target Y_t for each task t . It can preserve key information from the input and retain the necessary information for all target tasks during the network’s implicit compression process. In the implementation of SIMax, we utilize noise-contrastive estimation (Gutmann and Hyvärinen 2010; van den Oord, Li, and Vinyals 2018) to optimize the lower bound of $I(X; Z)$ in the multi-task encoding process. As shown in Figure 1(b) and (c), unlike IB-based MTL methods (Qian, Chen, and Gechter 2020; de Freitas et al. 2022) that minimize the input-representation information, our InfoMTL with SIMax handles the information in an opposite manner to avoid the insufficiency issue arising from repre-

*Corresponding author.

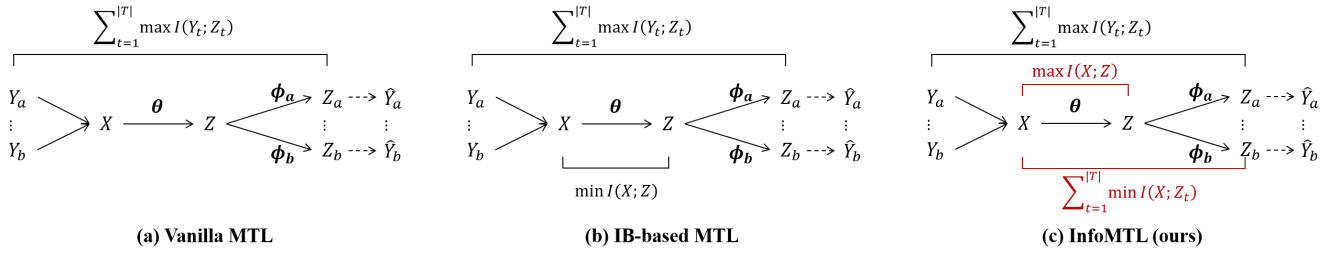


Figure 1: Comparison of different learning principles under Markov constraints in MTL paradigm. Given the input variable X , shared representations Z , task-specific output representations Z_t , and the prediction variable \hat{Y}_t , the Markov chain for each task t is $Y_t \rightarrow X \rightarrow Z \rightarrow Z_t \rightarrow \hat{Y}_t$.

sensation compression in the multi-task paradigm.

Besides, we design a task-specific information minimization (TIMin) principle to mitigate the negative effect of potential redundant features in the input for each task. The TIMin principle minimizes the mutual information between the input X and the task-specific output representations Z_t , while maximizing the mutual information between the output representations Z_t and the corresponding target task Y_t for each task t . It can compress task-irrelevant redundant information and preserve the necessary information relevant to the target for multi-task prediction. In the implementation of TIMin, we perform probabilistic embedding (Vilnis and McCallum 2015; Hu et al. 2024a) and task prediction in the multi-task decoding process. As shown in Figure 1(b) and (c), unlike IB-based MTL methods that reduce redundancy in the task-agnostic (shared or intermediate) representations, our TIMin focuses on alleviating redundancy in the task-specific output representations to avoid redundant interference across tasks. In this way, compressing task-irrelevant information does not interfere with the sufficiency of shared representations for other tasks.

We conduct experiments on six natural language understanding benchmarks. The results show that our InfoMTL outperforms 12 representative MTL methods across different PLMs under the same multi-task architecture. For example, with the RoBERTa backbone, InfoMTL improves the average performance by **+2.33%** and achieves an average relative improvement (Δp) of **+3.97%** over the EW baseline. Compared to single task learning baselines, InfoMTL also achieves better results on most tasks with the same scale of model parameters. Extensive experiments demonstrate that InfoMTL offers significant advantages in data-constrained and noisy scenarios, with learned representations being more sufficient, data-efficient, and robust.

The main contributions are summarized as follows: 1) We propose a shared information maximization (SIMax) principle to learn more sufficient shared representations for all target tasks. It can avoid the insufficiency issue arising from representation compression in the multi-task paradigm. 2) We design a task-specific information minimization (TIMin) principle to mitigate the negative effect of potential redundant features in the input for each task. It can compress task-irrelevant redundant information and preserve necessary information relevant to the target for multi-task prediction. 3)

We design a new principled multi-task learning framework (InfoMTL) to extract noise-invariant sufficient representations for all tasks. It can enhance language understanding of PLMs under the multi-task paradigm. 4) Experiments on six benchmarks show that our method outperforms comparative MTL approaches under the same multi-task settings, especially in data-constrained and noisy scenarios. Extensive experiments demonstrate that the learned representations are more sufficient, data-efficient, and robust.¹

Preliminary

Scope of the Study. This paper follows the line of multi-task optimization that typically employs a hard parameter-sharing pattern (Caruana 1993), where several lightweight task-specific heads are attached to a heavyweight task-agnostic backbone. Another orthogonal line of MTL research focuses on designing network architectures that usually employ a soft parameter-sharing pattern. The details of these two lines are listed in the Appendix.

Notations. Suppose there are T tasks and task t has its corresponding dataset \mathcal{D}_t . An MTL model typically involves two parametric modules, i.e., a shared encoder with parameters θ , and T task-specific decoders with parameters $\{\phi_t\}_{t=1}^{|T|}$, where ϕ_t represents the parameters for task t . Let $\ell_t(\mathcal{D}_t; \theta, \phi_t)$ be the average loss on \mathcal{D}_t for t . $\{\lambda_t\}_{t=1}^{|T|}$ are task-specific loss weights with a constraint that $\lambda_t \geq 0$ for t .

MTL Baseline. Since there are multiple losses in MTL, they usually are aggregated as a single one via loss weights, $\mathcal{L}(\mathcal{D}; \theta, \{\phi_t\}_{t=1}^{|T|}) = \sum_{t=1}^{|T|} \lambda_t \ell_t(\mathcal{D}_t; \theta, \phi_t)$. Apparently, the most simple method for loss weighting is to assign the same weight to all the tasks in the training process, i.e., $\lambda_t = \frac{1}{|T|}$ for task t in every iteration. The method is a common baseline in MTL named EW in this paper.

Information Flow in Neural Networks. Let X be an input random variable, Y_t be a target variable given task t , and $p(x, y_t)$ be their joint distribution for task t . The universal representations Z shared by all tasks is a function of X by a mapping $p_\theta(z|x)$. For task t , the output representations Z_t in the output space can be obtained by a task-specific

¹The source code is available at <https://github.com/zerohd4869/InfoMTL>

head $p_{\phi_t}(z_t|z)$, and the corresponding prediction variable \hat{Y}_t is non-parametric mapping of Z_t . Then, define information flow (Shwartz-Ziv and Tishby 2017; Goldfeld et al. 2019) in neural networks as a Markov chain shared by all tasks, i.e., $Y_t \rightarrow X \rightarrow Z \rightarrow Z_t \rightarrow \hat{Y}_t$ for task t .

Methodology

To ensure sufficiency for target tasks and mitigate the negative effects of redundant features, we propose a principled multi-task representation learning framework (InfoMTL) that extracts noise-invariant sufficient representations for all tasks. It contains two learning principles: shared information maximization (SIMax) and task-specific information minimization (TIMin), which constrain the amount of shared and task-specific information in the multi-task learning process.

Shared Information Maximization Principle

In the MTL paradigm, the mutual information between the input X and the shared representations Z is typically reduced implicitly (Figure 1(a)) or explicitly (Figure 1(b)) under the information bottleneck (IB) theory (Shwartz-Ziv and Tishby 2017; Kawaguchi et al. 2023). However, directly reducing redundancy for one task is prone to losing necessary information for others. As a result, the learned shared representations Z usually suffer from task-specific insufficiency.

To alleviate the insufficiency issue, a shared information maximization (SIMax) principle is proposed to learn more sufficient shared representations for all target tasks. The SIMax principle simultaneously maximizes the mutual information between the input X and the shared representations Z for all tasks, as well as the mutual information between the shared representations Z and the target task Y_t for each task t . It can be formulated as the maximization of the following Lagrangian,

$$\max \sum_{t=1}^{|T|} [I(Y_t; Z)] + \alpha I(X; Z), \quad (1)$$

subject to the Markov constraint, i.e., $Y_t \rightarrow X \rightarrow Z \rightarrow Z_t \rightarrow \hat{Y}_t$. α is a parameter that balances the trade-off between the informativeness of Z for X and Y_t .

In Equation (1), the second term promotes the task-agnostic shared representations Z preserves as much information as possible about the input X . It ensures the sufficiency of the shared representations Z for all potential targets Y . The first term encourages the shared representations Z to capture the necessary information relevant to the target Y_t for each task t . As shown in Figure 1(b) and (c), unlike IB-based MTL methods (Qian, Chen, and Gechter 2020; de Freitas et al. 2022) that explicitly minimize input-representation information, our InfoMTL with SIMax handles the information in an opposite manner to avoid the insufficiency issue arising from representation compression in the multi-task paradigm.

Implementation of SIMax The implementation of SIMax contains two terms, i.e., maximizing $\sum_{t=1}^{|T|} [I(Y_t; Z)]$ and $I(X; Z)$. Firstly, we maximize the lower bound of $I(Y_t; Z)$

by estimating the conditional entropy of the target Y_t given the shared representations Z . Following Kolchinsky, Tracey, and Wolpert (2019) and Hu et al. (2024a), we use cross-entropy (CE) as the estimator for each classification task t . Secondly, to maximize $I(X; Z)$, we use the InfoNCE estimator (Gutmann and Hyvärinen 2010; van den Oord, Li, and Vinyals 2018) to optimize the lower bound of $I(X; Z)$ during the multi-task encoding process. According to the information flow of $Z \leftarrow X \rightarrow Z'$, the Markov chain rule states that $I(X; Z) \geq I(Z; Z')$. For the InfoNCE estimation of maximizing $I(Z; Z')$, the selection of positive and negative samples as well as the implementation of noise-contrastive loss are consistent with Hu et al. (2024b). The optimization objective of SIMax for MTL can be:

$$\begin{aligned} \mathcal{L}_{\text{SIMax}} = & \mathbb{E}_{z \sim p_{\theta}(z|x)} \{ \mathbb{E}_{t \sim T} [-\log q_{\phi_t}(y_t|z)] \\ & - \alpha \log \frac{\exp(\text{sim}(z, z^+)/\tau)}{\sum_{z' \in \mathcal{B}^+} \exp(\text{sim}(z, z')/\tau)} \}, \end{aligned} \quad (2)$$

where $p_{\theta}(z|x)$ is a shared encoder with parameters θ . $q_{\phi_t}(y_t|z)$ is a task-specific decoder with parameters ϕ_t for task t , and its output distribution is adapted for task prediction by a non-parametric function (e.g., Softmax operation for classification). z^+ refers to the positive key of z , generated by dropout in $p_{\theta}(z|x)$. \mathcal{B}^+ represents the set of positive keys in the current batch \mathcal{B} . $\text{sim}(\cdot)$ is a pairwise similarity function, i.e., cosine similarity. $\tau > 0$ is a scalar temperature parameter that controls the sharpness of the probability distribution, which is applied during the Softmax operation.

Task-specific Information Minimization Principle

In multi-task scenarios, redundant information often differs across tasks, such that information beneficial for one task may become redundant for another. Directly applying the IB principle (Tishby, Pereira, and Bialek 1999; Tishby and Zaslavsky 2015) to compress the task-specific redundancy for one task can result in the loss of necessary information for other tasks. Consequently, both the shared representations Z and the task-specific output representations Z_t for a given task t often contain redundant features that are irrelevant to the specific task.

To better alleviate the redundancy issue, a task-specific information minimization (TIMin) principle is designed to mitigate the negative effect of potential redundant features in the inputs for the target task. It can compress task-irrelevant redundant information and preserve necessary information relevant to the target for multi-task prediction. The TIMin principle minimizes the mutual information between the input X and the task-specific output representations Z_t , while maximizing the mutual information between the output representations Z_t and the corresponding target task Y_t for each task t . The principle can be formulated as the maximization of the following Lagrangian,

$$\max \sum_{t=1}^{|T|} [I(Y_t; Z_t) - \beta I(X; Z_t)], \quad (3)$$

subject to $Y_t \rightarrow X \rightarrow Z \rightarrow Z_t \rightarrow \hat{Y}_t$ for each task t . β is a trade-off parameter of the compression of Z_t from the input X and the informativeness of Z_t for Y_t .

In Equation (3), the first term encourages the output representations Z_t to preserve task-relevant information necessary for multi-task prediction. The second term compresses task-irrelevant redundant information of Z_t for each task. These two terms make the learned output representations Z_t approximately the minimal sufficient task-specific representations for each task t . As shown in Figure 1(b) and (c), unlike IB-based MTL methods that reduce redundancy in the task-agnostic (shared or intermediate) representations, our TIMin focuses on alleviating redundancy in the task-specific output representations to avoid redundant interference across tasks. In this way, compressing task-irrelevant information does not interfere with the sufficiency of shared representations for other tasks.

Implementation of TIMin To achieve the principle of TIMin, we perform probabilistic embedding (Vilnis and McCollum 2015; Hu et al. 2024a) and task prediction in the multi-task decoding process. Following Hu et al. (2024a), we perform variational inference (Hoffman et al. 2013) to minimize the mutual information between the input X and task-specific output representations Z_t for each task t . It maps the shared representations Z to a set of different Gaussian distributions in the output space, i.e., $\mathbb{R}^{|Y_t|}$. Additionally, we can maximize the lower bound of $I(Y_t; Z_t)$ by estimating the conditional entropy $H(Y_t|Z_t)$.

Given the input x and its task-agnostic representations z , the task-specific output representations $z_t \sim p_{\theta, \phi_t}(z_t|x)$ can be learned by the shared encoder with parameters θ and task-specific head with parameters ϕ_t . The true posterior $p_{\theta, \phi_t}(z_t|x)$ can be approximated as $p_{\phi_t}(z_t|z)$ where $z \sim p_{\theta}(z|x)$. Let $r(z_t) \sim \mathcal{N}(z_t; \mathbf{0}, \mathbf{I})$ be an estimate of the prior $p(z_t)$ of z_t . Let $q_{\phi_t}(z_t|z)$ be a variational estimate of the intractable true posterior $p(z_t|z)$ of z_t given z , and learned by the t -th stochastic head parametrized by ϕ_t . Then, we have $I(Z; Z_t) = \int dz dz_t p(z, z_t) \log \frac{p(z_t|z)}{p(z_t)} \lesssim \int dz dz_t p(z) q(z_t|z) \log \frac{q(z_t|z)}{r(z_t)}$. And the optimization objective of TIMin for MTL can be:

$$\mathcal{L}_{\text{TIMin}} = \mathbb{E}_{t \sim T, z \sim p_{\theta}(z|x)} \left\{ \mathbb{E}_{z_t \sim q_{\phi_t}(z_t|z)} [-\log s(y_t|z_t)] + \beta KL(q_{\phi_t}(z_t|z); r(z_t)) \right\}, \quad (4)$$

where z_t is randomly sampled from $p_{\phi_t}(z_t|z)$ and $s(y_t|z_t)$ is a non-parametric operation on z_t . $KL(\cdot)$ denotes the analytic KL-divergence term, serving as the regularization that forces the variational posterior $q_{\phi_t}(z_t|z)$ to approximately converge to the Gaussian prior $r(z_t)$. $\beta > 0$ controls the trade-off between the sufficiency of z_t for predicting y_t , and the compression of z_t from x .

We assume the variational posterior $q_{\phi_t}(z_t|z)$ be a multi-variate Gaussian with a diagonal covariance structure, i.e.,

$$q_{\phi_t}(z_t^i|z^i) = \mathcal{N}(z_t^i; \mu_t(z^i), \Sigma_t(z^i)), \quad (5)$$

where $\mu_t(z^i)$ and $\Sigma_t(z^i)$ denote the mean and diagonal covariance of sample z^i for task t . Following Hu et al. (2022, 2024a), both of their parameters are input-dependent and can be learned by an MLP (a fully-connected neural network with a single hidden layer) for each task, respectively. Next, we sample z_t^i from the approximate posterior

$q_{\phi_t}(z_t^i|z^i)$, and obtain the prediction value by $s(y_t^i|z_t^i)$. Since the sampling process of z_t^i is stochastic, we use the re-parameterization trick (Kingma and Welling 2014) to ensure it trainable, i.e., $z_t^i = \mu_t(z^i) + (\Sigma_t(z^i))^{1/2} \odot \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where \odot refers to an element-wise product. Then, the KL term can be calculated by: $KL(q_{\phi_t}(z_t^i|z^i); r(z_t^i)) = -\frac{1}{2} (1 + \log \Sigma_t(z^i) - (\mu_t(z^i))^2 - \Sigma_t(z^i))$.

InfoMTL Framework

We incorporate the TIMin principle into the SIMax principle, and design a new information-theoretic multi-task learning framework (InfoMTL) to extract noise-invariant sufficient representations for all tasks. According to data processing inequality, in the Markov chain $Y_t \rightarrow X \rightarrow Z \rightarrow Z_t$, we have $I(Y_t; Z) \geq I(Y_t; Z_t)$. For simplification, we use $\max I(Y_t; Z_t)$ in TIMin to compute the lower bound of $\max I(Y_t; Z)$ in SIMax. As shown in Figure 1(c), the total learning principle of InfoMTL is,

$$\max \sum_{t=1}^{|T|} [I(Y_t; Z_t) - \beta I(X; Z_t)] + \alpha I(X; Z). \quad (6)$$

Firstly, maximizing $I(X; Z)$ can learn more sufficient shared representations for all target tasks. It preserves as much information as possible about the input X and ensures sufficiency of the task-agnostic shared representations for all targets Y . Then, minimizing $I(X; Z_t)$ mitigates the negative effect of potential redundant features in the input for each task. It can compress task-irrelevant redundant information. Finally, maximizing $I(Y_t; Z_t)$ captures necessary information relevant to the target Y_t from the output representations Z_t . It ensures sufficiency of the task-agnostic shared representations and task-specific output representations for multi-task prediction. Totally, InfoMTL can preserve necessary information in the shared representations for all tasks, and eliminate redundant information in the task-specific representations for each task.

Experiments

Experimental Setups

Datasets and Downstream Tasks Since this paper mainly focuses on MTL in natural language understanding, we experiment on six text classification benchmarks (Barbieri et al. 2020), i.e., *EmotionEval* (Mohammad et al. 2018) for emotion detection, *HateEval* (Basile et al. 2019) for hate speech detection, *IronyEval* (Hee, Lefever, and Hoste 2018) for irony detection, *OffensEval* (Zampieri et al. 2019) for offensive language detection, *SentiEval* (Rosenthal, Farra, and Nakov 2017) for sentiment analysis, and *StanceEval* (Mohammad et al. 2016) for stance detection. The details are listed in the Appendix.

Comparison Methods To fairly compare our method with different multi-task methods, we reproduce and compare with the following 12 representative MTL methods under the same experimental settings (e.g., network architecture). Comparison methods include Equal Weighting (EW),

Methods	BERT backbone		RoBERTa backbone	
	Avg.	$\Delta p \uparrow$	Avg.	$\Delta p \uparrow$
EW (baseline)	65.62	0.00	66.17	0.00
<i>Task-balanced Methods</i>				
SI	65.67	+0.06	67.16	+1.75
TW	65.68	+0.11	67.08	+1.55
UW	66.97	+2.22	67.11	+1.92
GLS	66.05	+0.60	67.32	+1.67
DWA	65.56	-0.09	66.94	+1.35
PCGrad	65.45	-0.50	67.42	+1.96
IMTL-L	66.18	+0.86	66.54	+0.67
RLW	66.76	+1.86	67.07	+1.63
<i>Probabilistic Methods</i>				
MT-VIB	65.80	+0.66	67.14	+2.00
VMTL	65.80	+0.65	67.05	+1.81
Hierarchical MTL	66.42	+1.76	66.84	+1.60
InfoMTL (ours)	67.51*	+3.70	68.50*	+3.97

Table 1: Multi-task performance (%) on six benchmarks. For all methods with BERT/RoBERTa backbone, we run three random seeds and report the average result on test sets. Best results are highlighted in bold. * represents statistical significance over scores of the baseline under the t -test ($p < 0.05$).

Task Weighting (TW), Scale-invariant Loss (SI), Uncertainty Weighting (UW) (Kendall, Gal, and Cipolla 2018), Geometric Loss Strategy (GLS) (Chennupati et al. 2019), Dynamic Weight Average (DWA) (Liu, Johns, and Davison 2019), Projecting Conflicting Gradient (PCGrad) (Yu et al. 2020), IMTL-L (Liu et al. 2021), Random Loss Weighting (RLW) (Lin et al. 2022), MT-VIB (Qian, Chen, and Gechter 2020), VMTL (Shen et al. 2021), and Hierarchical MTL (de Freitas et al. 2022). MT-VIB, VMTL, Hierarchical MTL are probabilistic MTL series. We use two pre-trained language models, i.e., BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019b), as the backbone model. Concretely, we use *bert-base-uncased*² and *roberta-base*² to initialize BERT and RoBERTa for fine-tuning. For each method, we fine-tune key parameters following the original paper to obtain optimal performance. We also compare with the single task learning baseline (STL) and the large language model GPT-3.5³. See the Appendix for more details.

Evaluation Metrics We use the same evaluation metric as the original tasks. The macro-averaged F1 over all classes is applied in EmotionEval, HatEval, and OffenseEval. The F1 score of ironic class is applied in IronyEval. The macro-averaged F1 score of favor and against classes is applied in StanceEval. The macro-averaged recall score is applied in SentiEval. Following Barbieri et al. (2020); Hu et al. (2024a), we report a global metric based on the average of all task-specific metrics, denoted as Avg.. Following Maninis, Radosavovic, and Kokkinos (2019); Lin et al. (2022), we also report the average relative improvement over the EW baseline, denoted as Δp . In addition, the t -test (Kim 2015) is used to verify the statistical significance of the differences between the results of our method and the baseline.

²<https://huggingface.co/>

³<https://chat.openai.com>

Implementation Details All experiments are conducted on a single NVIDIA Tesla A100 80GB card. The validation sets are used to tune hyperparameters and choose the optimal model. For each method, we run three random seeds and report the average result of the test sets. Besides, we experiment using an epoch number of 20, a total batch size of 128, and a maximum token length of 256. The maximum patience for early stopping is set to 3 epochs. The network parameters are optimized by using Adamax optimizer (Kingma and Ba 2015). The dropout is searched from $\{0, 0.2\}$. The parameters α and β are searched from $\{0.001, 0.01, 0.1, 1\}$. τ is searched from $\{0.1, 1\}$. See the Appendix for more details.

Main Results

Overall Results for MTL The overall results on six benchmarks with the BERT and RoBERTa backbone are summarized in Table 1. For each backbone, the top row shows the performance of the widely used EW, and we use it as a baseline to measure the relative improvement of different methods as shown in the definition of Δp . From the results, our InfoMTL achieves the best performance in terms of Avg and Δp on different backbones. With the BERT/RoBERTa backbone, InfoMTL enhances the average performance by **+1.89%/+2.33%** and achieves Δp of **+3.70%/+3.97%** over the EW baseline.

Fine-grained Results for Probabilistic MTL Table 2 shows the comparison of InfoMTL and representative probabilistic MTL methods such as MT-VIB, VMTL, and Hierarchical MTL. Our InfoMTL consistently outperforms EW on all tasks and achieves the best fine-grained results on most tasks, which confirms the effectiveness of our method.

Comparison with STL and LLM We compare our InfoMTL with the single-task learning (STL) baseline and the large language model (LLM) GPT-3.5. For STL, each task is trained with a separate model. For GPT-3.5, predictions are made under the zero-shot setting using task descriptions and instructions. As shown in Table 3, our InfoMTL outperforms GPT-3.5 on all tasks significantly. Compared to the STL baselines, our method also achieves better results on most tasks with the same scale of model parameters.

Ablation Study

We conduct ablation studies by removing the loss of shared information maximization (w/o SIMax) and task-specific information minimization (w/o TIMin) in our InfoMTL. As shown in Table 4, the full InfoMTL achieves the best results in terms of the average performance and Δp . See the Appendix for the fine-grained results with RoBERTa backbone. When removing either SIMax or TIMin, the ablated methods obtain inferior performance on most tasks. When further removing both components, the ablation w/o SIMax & TIMin would be equivalent to EW. The declining performance reveals the effectiveness of both principles.

Representation Evaluation and Analysis

Mutual Information Analysis We analyze the mutual information between different variables in the information

Methods	EmotionEval M-F1	HatEval M-F1	IronyEval F1(i.)	OffensEval M-F1	SentiEval M-Recall	StanceEval M-F1 (a. & f.)	Avg.	$\Delta p \uparrow$
EW (baseline)	74.37 \pm 0.56	44.08 \pm 5.26	65.32 \pm 1.84	79.04 \pm 1.43	70.64 \pm 1.71	63.59 \pm 2.43	66.17 \pm 0.43	0.00
MT-VIB	74.74 \pm 0.38	48.06 \pm 4.79	66.09 \pm 3.38	78.17 \pm 1.39	70.95 \pm 0.99	64.83 \pm 1.56	67.14 \pm 0.87	+2.00
VMTL	74.07 \pm 0.72	47.44 \pm 3.42	68.55 \pm 2.80	77.95 \pm 0.22	70.52 \pm 1.04	63.76 \pm 2.86	67.05 \pm 1.06	+1.81
Hierarchical MTL	74.09 \pm 1.73	48.52 \pm 4.26	64.92 \pm 6.14	78.26 \pm 1.63	71.45 \pm 0.44	63.82 \pm 0.54	66.84 \pm 1.68	+1.60
InfoMTL (ours)	76.90 \pm 0.62	48.44 \pm 2.15	68.94 \pm 1.86	79.78 \pm 0.86	71.92 \pm 0.36	65.02 \pm 1.81	68.50 \pm 0.58	+3.97

Table 2: Fin-grained results (%) of probabilistic multi-task methods with RoBERTa backbone. * represents statistical significance over scores of the baseline under the t -test ($p < 0.05$).

Methods	# Param.	EmotionEval M-F1	HatEval M-F1	IronyEval F1(i.)	OffensEval M-F1	SentiEval M-Recall	StanceEval M-F1 (a. & f.)	Avg.
GPT-3.5	(LLMs)	73.23	48.30	66.81	63.71	40.40	39.45	55.32
STL with CNN	110M+6 \times 2M	59.11	47.61	52.10	77.80	70.85	57.58	60.84
InfoMTL	110M	76.90	48.44	68.94	79.78	71.92	65.02	68.50
STL	6 \times 110M	74.49	45.26	53.27	79.20	72.43	66.70	65.23
InfoMTL*	< 6 \times 110M	78.55	52.38	68.14	80.80	72.40	68.56	70.14

Table 3: Comparison results (%) with different learning paradigms. We experiment with RoBERTa backbone for all methods except for GPT-3.5. STL means single-task learning with a cross-entropy loss. STL with CNN indicates fine-tuning task-specific CNN classifiers with a frozen RoBERTa backbone. InfoMTL and InfoMTL* indicate the model trained on six and pair-wise tasks, respectively. # Param. refers to the number of model parameters for all tasks excluding the task-specific linear head.

Methods	BERT backbone		RoBERTa backbone	
	Avg.	$\Delta p \uparrow$	Avg.	$\Delta p \uparrow$
InfoMTL	67.51	+3.70	68.50	+3.97
w/o SIMax	66.01	+0.71	67.59	+2.72
w/o TIMin	67.02	+2.59	67.55	+2.34
w/o SIMax & TIMin	65.62	0.00	66.17	0.00

Table 4: Ablation results (%) of our InfoMTL.

flow during training. From Figure 2, 1) compared to the EW baseline and MT-VIB, both InfoMTL and the ablation without TIMin gain larger mutual information between X and Z , given the same epoch. This indicates SIMax principle can promote the shared representations Z to preserve more information about the input X . 2) Compared to the EW baseline, InfoMTL without SIMax obtains less information Z from X , and larger information Z_t from Z , given the same epoch. The same trends can be observed for InfoMTL when compared to InfoMTL without TIMin. This indicates TIMin principle can compress the redundant features in the shared representations Z , and ensure sufficiency of Z_t for task t .

Representation Quality Evaluation To evaluate the quality of representations learned by different MTL methods, we measure the sufficiency of the learned representations on the test set for both the input data and the target task. Following Hu et al. (2024b), we use the uniformity (Uni.) metric (Wang and Isola 2020) to measure the preserved maximal information of the shared representations from the input, and the adjusted rand index (ARI) score to assess the preserved maximal information of output representations for label structure. Table 5 shows Uni. and ARI of the representations learned by different MTL methods on all benchmarks. Our InfoMTL achieves better performance on both metrics across all tasks. This implies that InfoMTL can learn both sufficient shared representations for the input and sufficient task-specific out-

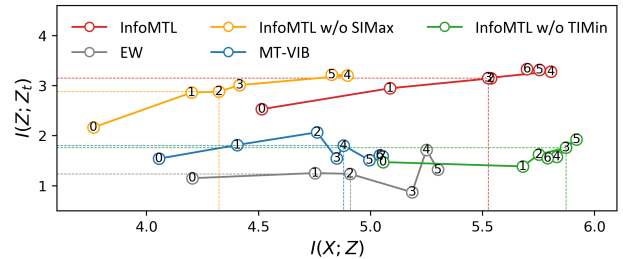


Figure 2: Mutual information analysis results. The X-axis refers to the mutual information between the shared representations Z and the input X , i.e., $I(X; Z)$. Y-axis represents the mutual information between the shared and output representations, i.e., $I(Z; Z_t)$. Each number on the line is the training epoch, and the optimal epochs are marked with dashed lines.

put representations for the target task.

Evaluation under Data-constrained Conditions

We experiment under different ratios of the training set to evaluate generalization when training with limited data. Following Hu et al. (2024a,b), all methods are trained on randomly sampled subsets from the original training set with different seeds, and we report the average results on the test set. Table 6 shows results of different techniques against different sizes of training set. InfoMTL achieves superior performance against different ratios of the training set under most settings. With only 20% training data, InfoMTL achieves Δp of +4.07% over the EW baseline, showing better generalization of InfoMTL under data-constrained conditions. This indicates InfoMTL can learn more sufficient representations from the inputs and enhance the efficiency of utilizing limited data.

Methods	EmotionEval		HatEval		IronyEval		OffensEval		SentiEval		StanceEval	
	ARI \uparrow	Uni. \downarrow	ARI \uparrow	Uni. \downarrow	ARI \uparrow	Uni. \downarrow	ARI \uparrow	Uni. \downarrow	ARI \uparrow	Uni. \downarrow	ARI \uparrow	Uni. \downarrow
EW	5.23	-2.30	-0.32	-1.54	1.06	-2.18	5.54	-2.19	6.76	-2.15	0.08	-2.35
UW	8.89	-2.21	-0.82	-1.54	0.68	-2.21	0.87	-2.18	4.73	-2.13	1.60	-2.21
PCGrad	7.21	-2.36	-0.03	-1.60	-0.28	-2.35	4.02	-2.29	2.97	-2.29	2.13	-2.48
MT-VIB	1.98	-2.25	0.14	-1.60	-0.09	-2.26	-0.17	-2.21	1.54	-2.13	0.01	-2.44
Hierarchical MTL	5.89	-2.35	-0.01	-1.60	0.04	-2.38	1.13	-2.32	0.55	-2.21	-0.00	-2.66
InfoMTL	51.38	-2.54	0.29	-1.83	11.40	-2.65	41.58	-2.69	28.66	-2.71	17.04	-2.71

Table 5: Quality evaluation of the learned representations by different MTL methods. Adjusted Rand Index (ARI, %) assesses preserved maximal information of output representations for label structure. Uniformity (Uni.) measures preserved maximal information of hidden representations from input. The lower uniformity means the better sufficiency for the input data, and the higher ARI means the better sufficiency for target task. RoBERTa is the default backbone.

Methods	Data per	Avg.	$\Delta p\uparrow$
EW	20%	62.43	0.00
UW	20%	61.78	-1.59
PCGrad	20%	62.75	+1.48
MT-VIB	20%	60.00	-4.18
Hierarchical MTL	20%	61.11	-2.30
InfoMTL	20%	64.83	+4.07
EW	40%	66.01	0.00
UW	40%	64.35	-2.82
PCGrad	40%	64.28	-2.93
MT-VIB	40%	63.58	-3.90
Hierarchical MTL	40%	62.86	-5.07
InfoMTL	40%	67.10	+2.16
EW	60%	66.38	0.00
UW	60%	66.17	-0.45
PCGrad	60%	65.82	-1.21
MT-VIB	60%	66.31	+0.04
Hierarchical MTL	60%	65.00	-1.95
InfoMTL	60%	66.71	+0.50
EW	80%	66.34	0.00
UW	80%	66.93	+1.30
PCGrad	80%	66.48	+0.81
MT-VIB	80%	65.34	-1.57
Hierarchical MTL	80%	65.35	-1.33
InfoMTL	80%	67.88	+2.53

Table 6: Results (%) against different sizes of training set. RoBERTa is the default backbone.

Robustness Evaluation on Noisy Data

To assess the adaptability to noisy data (Fang et al. 2022; Hu et al. 2024b), we evaluate the model’s robustness under various optimization objectives during multi-task learning. We adjust different strengths of random and adversarial perturbations on the test set. The random perturbations are from a multivariate Gaussian, and the adversarial perturbations are produced by a fast gradient method (Miyato, Dai, and Goodfellow 2017). These perturbations are scaled by the L_2 norm and then applied to the embedding layer in the testing process. Following Carlini and Wagner (2017); Hu et al. (2024b), we report the robust scores in terms of original evaluation metrics on noise samples generated from original test sets for each task. From Figure 3 (see the Appendix for results against different random perturbation strengths), InfoMTL gains better robust scores over other objectives on all tasks. Compared to EW, InfoMTL achieves an average

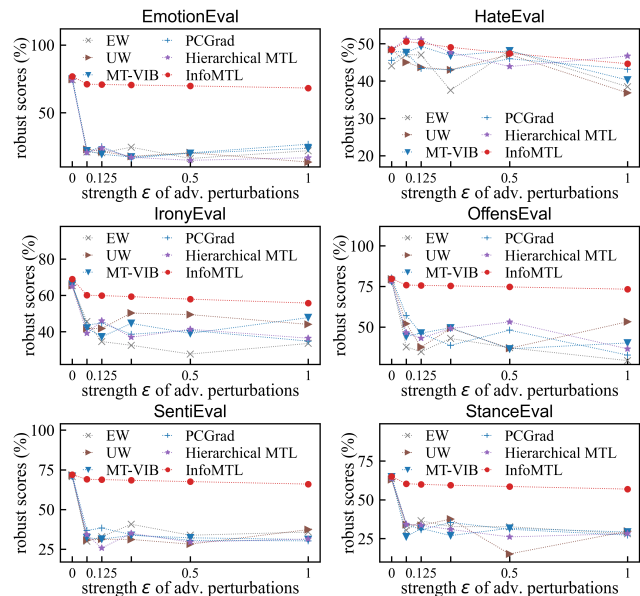


Figure 3: Robust scores (%) against adversarial perturbation strengths. RoBERTa is the default backbone.

increase of **+29.2%** and **+27.8%** in robust scores under random and adversarial noise, respectively. This indicates InfoMTL extracts noise-invariant representations for all tasks, which can enhance the model’s adaptability to noisy data.

Conclusion

We propose an information-theoretic multi-task learning framework (InfoMTL) to extract noise-invariant sufficient representations for all tasks, which can enhance language understanding of PLMs under the multi-task paradigm. It ensures sufficiency of shared representations for all tasks and mitigates the negative effect of redundant features. Firstly, the SIMax principle is proposed to learn more sufficient shared representations for all target tasks. In addition, the TIMin principle is designed to mitigate the negative effect of potential redundant features in the input for each task. Experiments on six benchmarks show that InfoMTL outperforms 12 comparative MTL methods under the same settings, especially in data-constrained and noisy scenarios.

Appendix Overview

In this supplementary material, we provide: (i) the related work, (ii) detailed description of experimental setups, and (iii) supplementary results.

Related Work

Recent works on Multi-task Learning (MTL) mainly come from two aspects: multi-task optimization and network architecture design.

Multi-task Optimization

The line of multi-task optimization works usually employ a hard parameter-sharing pattern (Caruana 1993), where several light-weight task-specific heads are attached upon the heavy-weight task-agnostic backbone.

Task-balanced Methods Task-balanced methods mainly focus on balancing learning process across multiple tasks for all tasks, such as loss-based and gradient-based methods. Loss-based methods (Kendall, Gal, and Cipolla 2018; Chennupati et al. 2019; Liu, Johns, and Davison 2019; Liu et al. 2021; Lin et al. 2022) focus on aligning the task loss magnitudes by rescaling loss scales. Gradient-based methods (Sener and Koltun 2018; Chen et al. 2018; Yu et al. 2020) aims to find an aggregated gradient to balance different tasks. Moreover, Liu et al. (2021); Lin et al. (2022) provide general task balancing strategies that can simultaneously balance the loss and gradient of different tasks. In real-world scenarios, the labeled data resource is limited and contains a certain amount of noise. The fact leads the above task-balanced MTL methods to perform suboptimally and struggle to achieve promising task prediction results.

Probabilistic Methods Probabilistic methods (Yousefi, Smith, and Álvarez 2019; Kim et al. 2022; Qian, Chen, and Gechter 2020; Shen et al. 2021; de Freitas et al. 2022) have been widely developed to explore shared priors for all tasks. The relationships among multiple tasks are investigated by designing priors over model parameters (Yu, Tresp, and Schwaighofer 2005; Titsias and Lázaro-Gredilla 2011; Archambeau, Guo, and Zoeter 2011; Bakker and Heskes 2003) under the Bayesian framework, or sharing the covariance structure of parameters (III 2009). In addition to the above methods that mainly focus on task relatedness or shared prior, some works (Vera, Vega, and Piantanida 2017; Qian, Chen, and Gechter 2020; de Freitas et al. 2022) introduce the information bottleneck (IB) principle (Tishby, Pereira, and Bialek 1999; Tishby and Zaslavsky 2015) into the information encoding process of multi-task learning. These IB-based methods typically enhance the adaptability to noisy data by compressing task-irrelevant redundant information and learning compact intermediate representations. Specifically, Qian, Chen, and Gechter (2020) use the variational IB (Alemi et al. 2017) to learn probabilistic representations for multiple tasks. de Freitas et al. (2022) propose a hierarchical variational MTL method that restricts information individual tasks can access from a task-agnostic representation.

In multi-task scenarios, redundant information often differs across tasks, leading to situations where information

Dataset	# Label	# Train	# Val	# Test	# Total
EmotionEval	4	3,257	374	1,421	5,502
HatEval	2	9,000	1,000	2,970	12,970
IronyEval	2	2,862	955	784	4,601
OffensEval	2	11,916	1,324	860	14,100
SentiEval	3	45,389	2,000	11,906	59,295
StanceEval	3	2,620	294	1,249	4,163

Table 7: Dataset statistics.

Hyperparameter	BERT	RoBERTa
Trade-off weight β	1	0.01
Trade-off weight α	0.01	0.1
Temperature τ	0.1	1
Number of epochs	20	20
Patience	3	3
Batch size	128	128
Learning rate	$5e^{-5}$	$5e^{-5}$
Weight decay	0	0
Dropout	0	0.2
Maximum token length	256	256

Table 8: Hyperparameters of InfoMTL.

beneficial for one task may become redundant for another. Directly applying the IB principle to compress redundancy for one task is prone to losing necessary information for other tasks. As a result, the learned shared representations would not only retain some redundant features but also face the task-specific insufficiency issue. To solve this, this paper proposes a new principled MTL framework InfoMTL to extract noise-invariant sufficient representations for all tasks. The proposed InfoMTL can ensure sufficiency for all target tasks and mitigate the negative effect of redundant features.

Architectures for MTL

Orthogonal to the line of multi-task optimization, another line of MTL focuses on designing network architectures that mitigate task interference by optimizing the allocation of shared versus task-specific parameters (Misra et al. 2016; Hashimoto et al. 2017; Ruder et al. 2019; Liu, Johns, and Davison 2019; Liu et al. 2019a). Among them, some methods by soft parameter-sharing can share parameters among tasks to a large extent, but usually lead to high inference cost. The scope of our work is complementary to the architecture for MTL, as we mainly focus on learning better multi-task representations rather than designing better architectures.

Experimental Setups

Datasets and Downstream Tasks This study primarily focuses on MTL in the field of natural language understanding, and proposes a new MTL approach to better handle real-world scenarios with data noise and limited labeled data. To effectively validate the proposed method, we selected six text classification benchmarks (Barbieri et al. 2020) from social media, which naturally contain some noise, to evaluate multi-task performance. The statistics are listed in Table 7. *EmotionEval* (Mohammad et al. 2018) involves detecting the emotion evoked by a tweet and is based on the

Methods	EmotionEval M-F1	HatEval M-F1	IronyEval F1(i.)	OffensEval M-F1	SentiEval M-Recall	StanceEval M-F1 (a. & f.)	Avg.	$\Delta p \uparrow$
InfoMTL	76.90 \pm 0.62	48.44 \pm 2.15	68.94 \pm 1.86	79.78 \pm 0.86	71.92 \pm 0.36	65.02 \pm 1.81	68.50 \pm 0.58	+3.97
w/o SIMax	76.02 \pm 1.10	49.30 \pm 3.75	64.63 \pm 3.14	79.44 \pm 1.93	71.67 \pm 1.25	64.47 \pm 1.65	67.59 \pm 1.06	+2.72
w/o TIMin	76.37 \pm 0.37	46.82 \pm 4.98	64.12 \pm 7.01	80.25 \pm 0.89	71.39 \pm 0.49	66.34 \pm 0.74	67.55 \pm 0.31	+2.34
w/o SIMax & TIMin	74.37 \pm 0.56	44.08 \pm 5.26	65.32 \pm 1.84	79.04 \pm 1.43	70.64 \pm 1.71	63.59 \pm 2.43	66.17 \pm 0.43	0.00

Table 9: Fin-grained ablation results (%) of InfoMTL with RoBERTa backbone.

Affects in Tweets conducted during SemEval-2018. Following Barbieri et al. (2020); Hu et al. (2024a), the four most common emotions (i.e., anger, joy, sadness, and optimism) are selected as labels. *HatEval* (Basile et al. 2019) stems from SemEval-2019 HatEval challenge and is used to predict whether a tweet is hateful towards immigrants or women. *IronyEval* (Hee, Lefever, and Hoste 2018) is from SemEval-2018 Irony Detection and consists of identifying whether a tweet includes ironic intents or not. *OffensEval* (Zampieri et al. 2019) is from SemEval-2019 OffensEval and involves predicting whether a tweet contains any form of offensive language. *SentiEval* (Rosenthal, Farra, and Nakov 2017) comes from SemEval-2017 and is designed for the task of determining whether a tweet expresses a positive, negative, or neutral sentiment. *StanceEval* (Mohammad et al. 2016) involves determining if the author of a piece of text has a favorable, neutral, or negative position towards a proposition or target.

Description of Comparison Methods we compare with the following 12 representative MTL methods. **Equal Weighting (EW)** is a typical baseline that applies equal weights for each task. MT-DNN (Liu et al. 2019a) is a version of EW baseline with the BERT backbone. **Task Weighting (TW)** assigns loss weights to each task based on the ratio of task examples. **Scale-invariant Loss (SI)** is invariant to rescaling each loss with a logarithmic operation. **Uncertainty Weighting (UW)** (Kendall, Gal, and Cipolla 2018) uses the homoscedastic uncertainty quantification to adjust task weights. **Geometric Loss Strategy (GLS)** (Chennupati et al. 2019) uses the geometric mean of task losses to the weighted average of task losses. **Dynamic Weight Average (DWA)** (Liu, Johns, and Davison 2019) sets the loss weight of each task to be the ratio of two adjacent losses. **Projecting Conflicting Gradient (PCGrad)** (Yu et al. 2020) removes conflicting components of each gradient w.r.t the other gradients. **IMTL-L** (Liu et al. 2021) dynamically reweighs the losses such that they all have the same magnitude. **Random Loss Weighting (RLW)** (Lin et al. 2022) with normal distribution, scales the losses according to randomly sampled task weights. **MT-VIB** (Qian, Chen, and Gechter 2020) is a variational MTL method based on information bottleneck. **VMTL** (Shen et al. 2021) is a variational MTL framework that uses Gumbel-Softmax priors for both representations and weights. **Hierarchical MTL** (de Freitas et al. 2022) is a hierarchical variational MTL method with compressed task-specific representations based on information bottleneck. We also compare with GPT-3.5, an enhanced generative pre-trained transformer

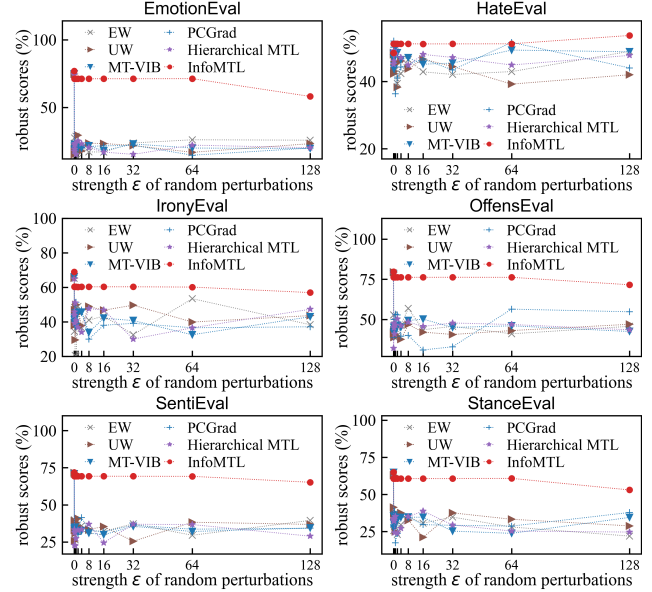


Figure 4: Robust scores (%) against random perturbation strengths. RoBERTa is the default backbone.

model based on text-davinci-003⁴, optimized for chatting.

Details of Evaluation Metrics Following Barbieri et al. (2020); Hu et al. (2024a), the global metric based on the average of all dataset-specific metrics, computed as $\text{Avg.} = \frac{1}{|T|} \sum_{t=1}^{|T|} \frac{1}{N_t} \sum_{n=1}^{N_t} M_{t,n}$, where $M_{t,n}$ denotes the performance for the n -th metric in task t . N_t denotes the number of metrics in task t . $|T|$ refers to the number of tasks. Following Maninis, Radosavovic, and Kokkinos (2019); Lin et al. (2022), the average relative improvement of each method over the EW baseline as the multi-task performance measure, denoted as $\Delta p = \frac{1}{|T|} \sum_{t=1}^{|T|} \frac{1}{N_t} \sum_{n=1}^{N_t} \frac{(-1)^{p_{t,n}} (M_{t,n} - M_{t,n}^{EW})}{M_{t,n}^{EW}}$, where $M_{t,n}^{EW}$ is the n -th metric score for EW on task t . $p_{t,n} = 0$ if a higher value is better for the n -th metric in task t and 1 otherwise.

Implementation Details Table 8 shows the best parameters of our InfoMTL with RoBERTa and BERT backbones.

Supplementary Results

Table 9 presents the fine-grained ablation results. Figure 4 illustrates the results against random perturbation strengths.

⁴We present the zero-shot results of the GPT-3.5-turbo snapshot from June 13th 2023.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. U24A20335), and the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (No. GZC20232969). The authors thank the anonymous reviewers and the meta-reviewer for their helpful comments on the paper.

References

- Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *ICLR*.
- Archambeau, C.; Guo, S.; and Zoeter, O. 2011. Sparse Bayesian Multi-Task Learning. In *NeurIPS*, 1755–1763.
- Bakker, B.; and Heskes, T. 2003. Task Clustering and Gating for Bayesian Multitask Learning. *J. Mach. Learn. Res.*, 4: 83–99.
- Barbieri, F.; Camacho-Collados, J.; Anke, L. E.; and Neves, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *EMNLP (Findings)*, 1644–1650.
- Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Pardo, F. M. R.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *SemEval@NAACL-HLT*, 54–63.
- Carlini, N.; and Wagner, D. A. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57.
- Caruana, R. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *ICML*, 41–48.
- Chen, Z.; Badrinarayanan, V.; Lee, C.; and Rabinovich, A. 2018. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In *ICML*, volume 80, 793–802.
- Chennupati, S.; Sistu, G.; Yogamani, S. K.; and Rawashdeh, S. A. 2019. MultiNet++: Multi-Stream Feature Aggregation and Geometric Loss Strategy for Multi-Task Learning. In *CVPR Workshops*, 1200–1210.
- de Freitas, J. M.; Berg, S.; Geiger, B. C.; and Mücke, M. 2022. Compressed Hierarchical Representations for Multi-Task Learning and Task Clustering. In *IJCNN*, 1–8.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.
- Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2022. ANIMC: A Soft Approach for Autoweighted Noisy and Incomplete Multiview Clustering. *IEEE Trans. Artif. Intell.*, 3(2): 192–206.
- Goldfeld, Z.; van den Berg, E.; Greenewald, K. H.; Melnyk, I.; Nguyen, N.; Kingsbury, B.; and Polyanskiy, Y. 2019. Estimating Information Flow in Deep Neural Networks. In *ICML*, volume 97, 2299–2308.
- Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, volume 9, 297–304.
- Hashimoto, K.; Xiong, C.; Tsuruoka, Y.; and Socher, R. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *EMNLP*, 1923–1933.
- Hee, C. V.; Lefever, E.; and Hoste, V. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *SemEval@NAACL-HLT*, 39–50.
- Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. W. 2013. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1): 1303–1347.
- Hu, D.; Hou, X.; Du, X.; Zhou, M.; Jiang, L.; Mo, Y.; and Shi, X. 2022. VarMAE: Pre-training of Variational Masked Autoencoder for Domain-adaptive Language Understanding. In *EMNLP (Findings)*, 6276–6286.
- Hu, D.; Wei, L.; Liu, Y.; Zhou, W.; and Hu, S. 2024a. Structured Probabilistic Coding. In *AAAI*, 12491–12501.
- Hu, D.; Wei, L.; Zhou, W.; and Hu, S. 2024b. Representation Learning with Conditional Information Flow Maximization. In *ACL*, 14088–14103.
- III, H. D. 2009. Bayesian Multitask Learning with Latent Hierarchies. In *UAI*, 135–142.
- Kawaguchi, K.; Deng, Z.; Ji, X.; and Huang, J. 2023. How Does Information Bottleneck Help Deep Learning? In *ICML*, volume 202, 16049–16096.
- Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *CVPR*, 7482–7491.
- Kim, D.; Cho, S.; Lee, W.; and Hong, S. 2022. Multi-Task Processes. In *ICLR*.
- Kim, T. K. 2015. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6): 540–546.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.
- Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.
- Kolchinsky, A.; Tracey, B. D.; and Wolpert, D. H. 2019. Nonlinear Information Bottleneck. *Entropy*, 21(12): 1181.
- Lin, B.; Ye, F.; Zhang, Y.; and Tsang, I. W. 2022. Reasonable Effectiveness of Random Weighting: A Litmus Test for Multi-Task Learning. *Trans. Mach. Learn. Res.*, 2022.
- Liu, L.; Li, Y.; Kuang, Z.; Xue, J.; Chen, Y.; Yang, W.; Liao, Q.; and Zhang, W. 2021. Towards Impartial Multi-task Learning. In *ICLR*.
- Liu, S.; Johns, E.; and Davison, A. J. 2019. End-To-End Multi-Task Learning With Attention. In *CVPR*, 1871–1880.
- Liu, X.; He, P.; Chen, W.; and Gao, J. 2019a. Multi-Task Deep Neural Networks for Natural Language Understanding. In *ACL*, 4487–4496.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.
- Maninis, K.; Radosavovic, I.; and Kokkinos, I. 2019. Attentive Single-Tasking of Multiple Tasks. In *CVPR*, 1851–1860.

- Misra, I.; Shrivastava, A.; Gupta, A.; and Hebert, M. 2016. Cross-Stitch Networks for Multi-task Learning. In *CVPR*, 3994–4003.
- Miyato, T.; Dai, A. M.; and Goodfellow, I. J. 2017. Adversarial Training Methods for Semi-Supervised Text Classification. In *ICLR (Poster)*.
- Mohammad, S. M.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 Task 1: Affect in Tweets. In *SemEval@NAACL-HLT*, 1–17.
- Mohammad, S. M.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *SemEval@NAACL-HLT*, 31–41.
- Qian, W.; Chen, B.; and Gechter, F. 2020. Multi-Task Variational Information Bottleneck. *CoRR*, abs/2007.00339.
- Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *SemEval@ACL*, 502–518.
- Royer, A.; Blankevoort, T.; and Bejnordi, B. E. 2023. Scalarization for Multi-Task and Multi-Domain Learning at Scale. In *NeurIPS*.
- Ruder, S.; Bingel, J.; Augenstein, I.; and Søgaard, A. 2019. Latent Multi-Task Architecture Learning. In *AAAI*, 4822–4829.
- Sener, O.; and Koltun, V. 2018. Multi-Task Learning as Multi-Objective Optimization. In *NeurIPS*, 525–536.
- Shen, J.; Zhen, X.; Worring, M.; and Shao, L. 2021. Variational Multi-Task Learning with Gumbel-Softmax Priors. In *NeurIPS*, 21031–21042.
- Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the Black Box of Deep Neural Networks via Information. *CoRR*, abs/1703.00810.
- Tishby, N.; Pereira, F. C.; and Bialek, W. 1999. The Information Bottleneck Method. In *Proc. of the 37th Allerton Conference on Communication and Computation*.
- Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *ITW*, 1–5.
- Titsias, M. K.; and Lázaro-Gredilla, M. 2011. Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning. In *NeurIPS*, 2339–2347.
- van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*, abs/1807.03748.
- Vera, M.; Vega, L. R.; and Piantanida, P. 2017. Compression-Based Regularization with an Application to Multi-Task Learning. *CoRR*, abs/1711.07099.
- Vilnis, L.; and McCallum, A. 2015. Word Representations via Gaussian Embedding. In *ICLR*.
- Wang, T.; and Isola, P. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *ICML*, volume 119, 9929–9939.
- Wu, S.; Zhang, H. R.; and Ré, C. 2020. Understanding and Improving Information Transfer in Multi-Task Learning. In *ICLR*.
- Yousefi, F.; Smith, M. T.; and Álvarez, M. A. 2019. Multi-task Learning for Aggregated Data using Gaussian Processes. In *NeurIPS*, 15050–15060.
- Yu, K.; Tresp, V.; and Schwaighofer, A. 2005. Learning Gaussian processes from multiple tasks. In *ICML*, volume 119, 1012–1019.
- Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient Surgery for Multi-Task Learning. In *NeurIPS*.
- Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *SemEval@NAACL-HLT*, 75–86.