Rethink Hierarchical Distributional Learning for Fine-grained Loneliness Characterization from Reddit Posts

Stanford CS224N Custom Project

Yunfan Jiang
Department of Electrical Engineering
Stanford University
yjiang05@stanford.edu

Abstract

Loneliness is a multidimensional experience. It is crucial to understand and characterize young adults' expressions and copping strategies with various forms of loneliness. In this project, we rethink previous efforts on building language models using hierarchical distributional learning for fine-grained loneliness characterization. To improve it, we first adapt a pre-trained language model to the data distribution we are interested in by pre-training it on 190K loneliness-related Reddit posts using unsupervised objectives. We then attempt to mitigate the *negative transfer* identified in previous work by learning to adaptively weight different sub-tasks. Our experiments show that the second pre-training stage greatly helps to improve the performance of characterizing fine-grained loneliness. By looking into learned loss weights, we analyze how different sub-tasks are weighted during the training and discuss various ways that are potentially more efficient to reduce the negative transfer.

1 Key Information to include

- Mentor: Elaine Sui
- External Collaborators (if you have any): No
- Sharing project: No

2 Introduction

Loneliness is a multidimensional experience and encapsulates different manifestations and forms [1]. Understanding fine-grained loneliness expression has a wide range of applications, from mental health screening to intervention and prevention of severe consequences including depression, bipolar disorder, or even mortality. It has been a goal for long time to enable machines to understand emotions such as loneliness [2].

Large-scale pre-trained models have become prevalent in the research of natural language processing (NLP) since the introduction of the Transformer model [3, 4, 5, 6, 7, 8]. Previous work leverage pre-trained language models for fine-grained emotion classification [9], emotion measurement during the pandemic [10], affective response detection [11], multilingual sentiment analysis and emotion detection [12], and so on. Despite the growing interest in examining emotion expressions in online discourses using large-scale language models, research on loneliness classification is limited [13], not even to mention fine-grained loneliness characterization. [14] introduces a dataset based on Reddit posts, annotates loneliness-related posts with different fine-grained categories, and builds a BERT-based [4] model for fine-grained loneliness characterization. Nevertheless, we argue that it

is worth rethinking [14] in terms of ways to improve performances and mitigate *negative transfer* [15, 16, 17] that is present but not investigated or analyzed.

Built on the top of [14], we improve it by pre-training the language model on loneliness-related Reddit posts using unsupervised objectives and attempt to alleviate the negative transfer by learning to weight different sub-tasks. Our contributions are twofold:

1. We improve the original performance by a large margin by pre-training the language model on 190K loneliness-related Reddit posts using unsupervised objectives.
2. We investigate the phenomenon of negative transfer and attempt to reduce it by learning to adaptively weight different sub-tasks. Based on that, we analyze how sub-tasks are weighted during the training and discuss various ways that are potentially more efficient to alleviate it.

3 Related Work

3.1 NLP for Emotion Analysis

There is a growing interest in NLP to study emotion analysis since the introduction of the first emotion recognition benchmark [18]. [19] summarized and unified several emotion datasets before the era of Transformer models [3]. [20] used automatic weak labeling to build an emotion dataset based on Twitter hashtags and proposed a model based on gated recurrent neural network [21] that achieved a state-of-the-art in classifying fine-grained emotions. With the success of large-scale pre-trained language models such as BERT [4], top-performing models in the EmotionX Challenge all leveraged feature representations from pre-trained BERT models [22]. [9] introduced a manually labeled dataset containing 58k English Reddit comments annotated for 27 emotion categories or Neutral. They proposed a BERT-based model and suggested room for improvement. Despite the granular emotion taxonomy in [9], loneliness-related emotion still requires comprehensive studies. A recent work instead studies the granular categories of loneliness emotion [14]. They built a dataset using Reddit posts annotated manually for binary and fine-grained loneliness. They also proposed two BERT-based models for loneliness classification and characterization. Our work is built on the top of it with improved performance and investigation into negative transfer.

3.2 Multi-task Learning and Negative Transfer

Multi-task learning [23, 24, 25] is a learning paradigm in which machine learning models are trained with data from multiple different tasks simultaneously. Common ideas behind various related tasks can be learned by using shared representations. It is prevalent in many machine learning areas such as computer vision [26, 27, 28, 29, 30, 31, 32], reinforcement learning [33, 34, 35, 36, 37, 38, 39], and NLP [40, 6, 41, 42, 43, 44]. In the thread of using language model for fine-grained loneliness characterization, Jiang et al. [14] uses a model that simultaneously learns six sub-tasks. One is the binary loneliness classification task. Each of the remaining five tasks corresponds to characterize one fine-grained category of loneliness.

The phenomenon of *negative transfer* [16, 25] (also named *destructive interference*) refers to the decrease of model's performance on a task caused by the increased performance on another task with different needs. How to mitigate negative transfer remains an active area in the research of multi-task learning. Numerous efforts have been made on neural network architecture [45, 46, 47, 48], optimization [26, 49, 50, 51, 52, 53, 54], and the design of better learning curriculum [55, 56, 57, 58, 59, 60].

As for multi-task learning for fine-grained loneliness characterization, although results from [14] show the evidence of negative transfer, they do not investigate in details. Therefore, in this work we rethink the method in [14] by examining the cause of negative transfer.

3.3 Hierarchical Multi-label Classification

In the hierarchical multi-label classification (HMC) problem, classes are structured hierarchically. An object can be assigned to multiple paths of the hierarchical tree [61, 62]. Algorithms designed to solve HMC problems either optimize losses globally or locally [63]. [64] proposed a method in which each local classifier predicts a particular node in the hierarchy tree. On the other end of

global optimization, [65, 66] designed global classifiers that can process the entire hierarchy and all nodes as a whole. To leverage advantages of both global and local optimization, [67] proposed a novel nerual network architecture named HMCN for HMC problem. HMCN includes both local and global classifiers and incorporates label hierarchy into the model architecture. Each local and global classifiers are independently learned during training. The predictions used in evaluation and inference are linear combinations of these two types of classifiers. [14] built on the top of HMCN and proposed a similar model for hierarchical fine-grained loneliness characterization.

4 Approach

In this section, we first formulate the problem of hierarchical distributional learning for fine-grained loneliness characterization as introduced in [14]. We then describe the neural network model we used. Finally, we elaborate two ways we rethink the work in [14]. One is to improve it by adapting a pre-trained language model to the data distribution we are interested in through fine-tuning on 190K loneliness-related Reddit posts using unsupervised objectives. The other attempts to mitigate the negative transfer by learning to adaptively weight different sub-tasks.

4.1 Hierarchical Distributional Learning for Fine-grained Loneliness Characterization

The problem of fine-grained loneliness characterization is essentially a text classification problem. Each sentence is associated with a structured label hierarchy as shown in Figure 1. Concretely, there are two degrees of granularity for each Reddit post. The coarse one measures if a Reddit post expresses loneliness or not. The fine-grained one measures the loneliness in multiple different dimensions, namely “duration", “context", “interpersonal", and “interaction". Table 1 shows labels associated with fine-grained categories.

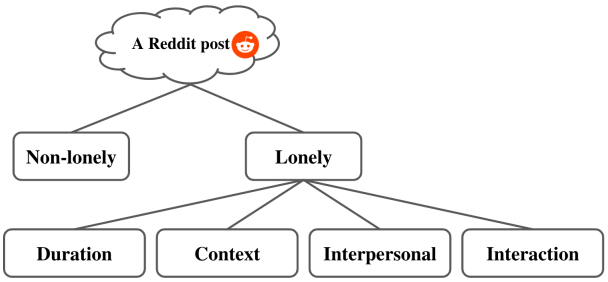


Figure 1: Structured label hierarchy.

Category	Labels
Duration	Transient, Enduring, Ambiguous, NA
Context	Social, Physical, Somatic, Romantic, NA
Interpersonal	Romantic, Friendship, Family, Peers, NA
Interaction	Seek Advice, Provide Support, Seek Val. & Aff., Reach Out, Non Directed

Table 1: Fine-grained loneliness categories and their labels [14]. “Seek Val. & Aff.” denotes "Seek Validation & Affirmation". NA indicates not applicable because posts can be irrelevant to such categories or labels.

Following literature [14], we formulate this problem under the framework of *label distributional learning* (LDL) [68]. Instead of learning models to predict one-hot labels, we learn models that are able to predict *distributions* over labels. Recent advances also show that adopting distributional labels yields better generalization performances in natural-image classification, image-based diagnostic, and age estimation [68, 69, 70]. For a concrete example, given a Reddit post as shown in Figure 2, the model is supposed to first predict that this post expresses loneliness (the coarser loneliness), then for the fine-grained dimension of “context", it should predict a distribution over five labels of that

category (namely “social", “physical", “somatic", “romantic", and “NA" as shown in Table 1) and assign probabilities of $\frac{2}{3}$ to label “social" and $\frac{1}{3}$ to label “romantic". The model repeats the same procedure for other three fine-grained categories.

“Have you ever feel bad when your friend talking about her crush? I think I am a introvert but I have being alone. I want someone beside whom I can go out or brag everything. But after I break up with my girlfriend, I feel like no one beside me anymore. And now there’s that friends and they are sharing their feeling about their crush. And it make me feel something I don’t know. I know I am not in love with them but hearing them talking about someone make me feel hurt. What should I do? Is there something wrong with me?”

Figure 2: An example Reddit post.

With the goal of characterizing fine-grained loneliness under the LDL framework, we now formulate this problem. Formally, we denote the i-th Reddit post as $\mathbf{x}^{(i)}$ and corresponding (tree) labels as $\mathcal{P}^{(i)}$. Note that $\mathcal{P}^{(i)}$ includes $\mathcal{P}^{(i)}_{lonely}$ and $\mathcal{P}^{(i)}_{f.g.}$, where $\mathcal{P}^{(i)}_{lonely}$ represents the distribution of loneliness itself (i.e., lonely or non-lonely) and $\mathcal{P}^{(i)}_{f.g.}$ instead represents fine-grained ones. $\mathcal{P}^{(i)}_{f.g.}$ contains four parts with each part $\mathcal{P}^{(i)}_{c}$ corresponding to each fine-grained category $c \in \mathcal{C} = \{\text{duration, context, interpersonal, interaction}\}$. Denoting model predictions as $\hat{\mathcal{P}}$ and total number of samples as N, we learn models by minimizing

$$\frac{1}{N}\sum_{i=1}^{N} \ell(\mathcal{P}^{(i)}_{lonely}, \widehat{\mathcal{P}}^{(i)}_{lonely}) + \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \ell(\mathcal{P}^{(i)}_{c}, \widehat{\mathcal{P}}^{(i)}_{c}). \tag{1}$$

Note that $\ell(\cdot)$ measures the distance between two *distributions*. It can be KL divergence or Hellinger divergence or simply the cross-entropy.

4.2 BERT-Based Hierarchical Distributional Learning Network

We follow [14] to use a BERT-based hierarchical distributional learning network (HDLN). It is built on the top of HMCN for HMC problem as proposed in [67]. As shown in Figure 3, leveraging text representation extracted from pre-trained BERT model [4], HDLN incorporates the label hierarchy and graph of conditionality into model architecture and learns one global classifier and five local classifiers to predict distributions we are interested in. The global classifier predicts the concatenation of all distributional labels $\hat{\mathcal{P}}_G$ without any hierarchy imposed. Instead, five local classifiers model the graph of conditionality and individually predict one out of five distributions $\hat{\mathcal{P}}_L$ with hierarchy imposed. Final predictions $\hat{\mathcal{P}}_F$ are linear combinations $\hat{\mathcal{P}}_F = \beta\hat{\mathcal{P}}_L + (1-\beta)\hat{\mathcal{P}}_G$, where $\beta \in [0, 1]$ is a hyperparameter.

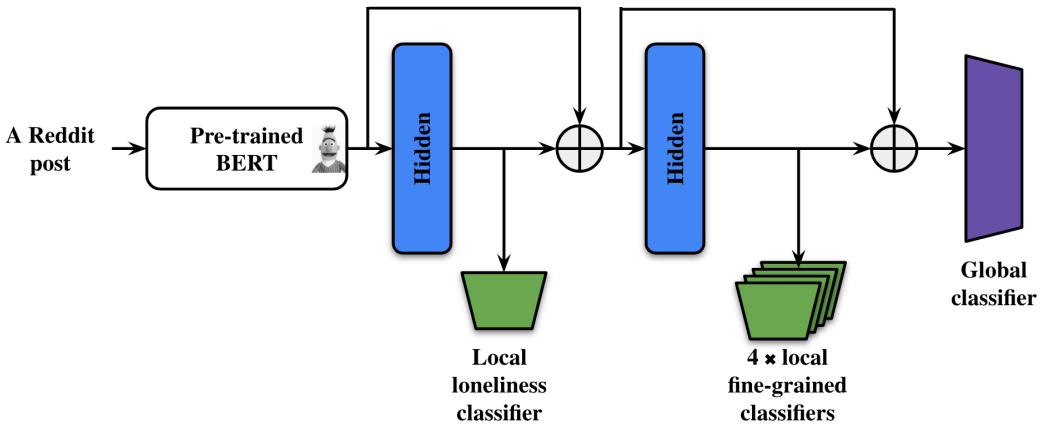


Figure 3: Architecture of HDLN. Symbol $\oplus$ denotes vector concatenation.

4.3 Rethinking Hierarchical Distributional Learning for Loneliness Characterization

4.3.1 Unsupervised Adaptation of Pre-trained BERT

Literature suggests that the performance of a BERT model can be improved by simply training for longer time using more data [71]. We hypothesize that performances in [14] can also be improved in such a way by pre-training a pre-trained BERT model on loneliness-related Reddit posts. As labels are not available for large-scale Reddit posts, we instead propose to use unsupervised objectives such as masked language modelling (MLM) [4, 71]. For a certain input Reddit post sequence, one token is randomly sampled and replaced with the special token `[MASK]`. Then BERT model is then trained to predict those masked tokens by optimizing the cross-entropy loss.

4.3.2 Learn to Weight Sub-tasks

One line of previous efforts towards mitigating negative transfer centers around loss weighting. Researchers have proposed to weight losses of sub-tasks by uncertainty [26], by learning speed [50], by performance [53], and so on. Here we explore a simpler method to learn to weight adaptively. In [14] losses from multiple sub-tasks are equally weighted globally and locally. Then aggregated global and local losses are equally weighted again. We argue that this may not be a proper weighting scheme. Instead, we propose to use three learnable weight vectors $\mathbf{\Gamma}_G$, $\mathbf{\Gamma}_L$, and $\mathbf{\Lambda}$ to weight sub-losses of global classifier and local classifiers and aggregated global and local losses. To prevent degenerated case in which all weights are optimized to zero, we impose the following constraints.

$$\|\mathbf{\Gamma}_G\|_1 = 1 \tag{2}$$

$$\|\mathbf{\Gamma}_L\|_1 = 1 \tag{3}$$

$$\|\mathbf{\Lambda}\|_1 = 1 \tag{4}$$

5 Experiments

In this section, we start with discussing the dataset we used. We then elaborate our evaluation metrics. We then provide experimental details. Finally we report results.

5.1 Data

We use the FIG-Loneliness dataset[1] introduced in [14]. In the first experiment of pre-training a pre-trained BERT model, we use 190K unlabeled loneliness-related Reddit posts. For the second experiment of investigating the negative transfer, we use the same training set, dev set, test set, and splitting as in [14]. The entire labeled dataset includes 6K posts cross-annotated by six raters. A concrete example of an input and its expected outputs is given in Section 4.1.

5.2 Evaluation method

For binary loneliness classification (i.e., the first level in Figure 1), we use “Accuracy", “Precision", “Recall", and “F1". For fine-grained loneliness characterization, we use distributional metrics adopted from [68] including “Clark distance", “Canberra metric", “Cosine similarity", and “intersection similarity". Defining $\boldsymbol{P} \in \mathbb{R}^K$ and $\hat{\boldsymbol{P}} \in \mathbb{R}^K$ as the ground-truth and predicted distributions with K supports, Table 2 summarizes the computation of distributional metrics. Note that the distributional version of metric “Accuracy" is similar to “mode matching".

5.3 Experimental details

There are two stages involved in the first experiment of pre-training a pre-trained BERT model, i.e., 1) pre-training using MLM on 140K loneliness-related Reddit posts and 2) fine-tuning on FIG-Loneliness labeled dataset. In the first stage, we pre-train a `bert-base-cased` model for 3 epochs

[1] `https://huggingface.co/datasets/FIG-Loneliness/FIG-Loneliness`

Metric	Formulation
Accuracy ↑	$\mathbb{1}\left(arg\,max(\hat{\boldsymbol{P}}) \in arg\,max(\boldsymbol{P})\right)$
Clark ↓	$\sqrt{\sum_k^K \frac{\left(P_k-\hat{P}_k\right)^2}{\left(P_k+\hat{P}_k\right)^2}}$
Canberra ↓	$\sum_k^K \frac{\left\|P_k-\hat{P}_k\right\|}{P_k+\hat{P}_k}$
Cosine ↑	$\frac{\sum_k^K P_k\hat{P}_k}{\sqrt{\sum_k^K P_k^2}\sqrt{\sum_k^K \hat{P}_k^2}}$
Intersection ↑	$\sum_k^K \min\left(P_k, \hat{P}_k\right)$

Table 2: Distributional metrics calculation. ↑ indicates higher values are better. ↓ indicates lower values are better.

on 190K loneliness-related Reddit posts. We use AdamW [72] with Cosine schedule [73] of learning rate warm up to 5×10^{-5}. We use a mask probability of 0.15. We train with 2 NVIDIA V100 GPUs with an effective batch size of 16. Training takes approximately 4 hours. In the second stage, we train our model for 3 different random seeds. We use AdamW again with a learning rate of 2×10^{-5} to train for at most 20 epochs. The model for evaluation from each run is chosen to be the checkpoint with the highest dev score. We train using a NVIDIA P100 GPU with a batch size of 16. The training takes approximately 5 hours.

The experimental details in the second experiment of investigating negative transfer are similar to the second stage of fine-tuning as described above, except that we use learnable loss weights to balance the learning of different sub-tasks.

5.4 Results

Tables 3 and 4 shows results on binary loneliness classification and fine-grained characterization, respectively. We denote our method with the adapted BERT as "Ours w/ Adapted BERT" and our method with the adapted BERT and learned sub-loss weights as "Ours w/ Learned Weights". We highlight values in row "Ours w/ Adapted BERT" if they are better than values in row "Baseline (HDLN)" because of the same model architecture used. We highlight values in row "Ours w/ Learned Weights" if they are better than values in row "Baseline (BERT + MLP)" because we are interested in if negative transfer is mitigated.

Our method with the adapted BERT generally outperforms the baseline. It is expected because by pre-training the pre-trained BERT model on unlabelled loneliness-related Reddit posts using MLM objective, we adapt it to the data distribution we are interested in and it hence can provide better text representations for downstream tasks. As for the thread of reducing the negative transfer, improvements are not consistent. Some results only improve over the baseline marginally and some do not improve at all. We thus analyze the effects of the learned weights in the next section.

6 Analysis

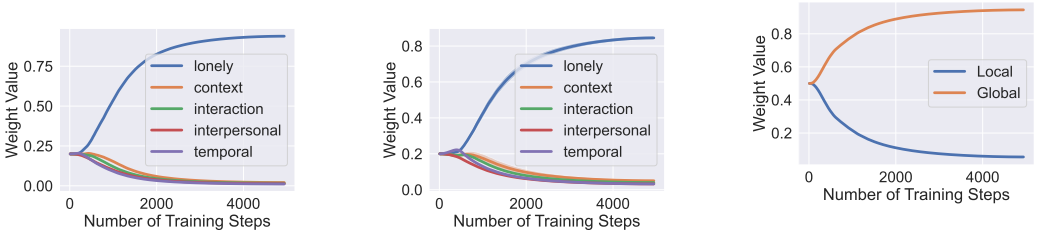
We now analyze how learned loss weights adapt during the entire training to shed light on the cause of negative transfer. Figure 4 shows the change of learnable loss weights during the entire training. Regarding learning loneliness expression with different degrees of granularity, we find that the model tends to learn the coarse one, i.e., lonely vs non-lonely in the first level of hierarchy as shown in Figure 1. The learning signal from coarse loneliness gradually dominates over learning signals from fine-grained loneliness. This phenomenon happens in both global and local classifiers. Regarding the learning of global classifier and local classifiers, the learning signal from the global classifier gradually dominates. It raises two questions: 1) "Does the label hierarchy really exist?" and 2) "If the label hierarchy exists, does the model architecture properly model the conditionality?". We leave the answering to these two open-ended questions as future work.

Regarding more efficient way to mitigate negative transfer, we hypothesize that weighting losses by performances is a promising direction. Instead of imposing naive constraints that weights sum up to

	Acc. ↑	Precis. ↑	Rec. ↑	F1 ↑
Baseline (BERT + MLP)	0.9722 ± 0.0046	0.9538 ± 0.0117	0.9870 ± 0.0036	0.9700 ± 0.0048
Baseline (HDLN)	0.9763 ± 0.0041	0.9609 ± 0.0045	0.9883 ± 0.0063	0.9744 ± 0.0045
Ours w/ Adapted BERT	**0.9763** ± **0.0008**	**0.9632** ± **0.0044**	0.9857 ± 0.0048	**0.9734** ± **0.0009**
Ours w/ Learned Weights	**0.9781** ± **0.0008**	**0.9634** ± **0.0044**	**0.9896** ± **0.0036**	**0.9763** ± **0.0008**

Table 3: Results for loneliness binary classification. "Acc.": Accuracy. "Precis.": Precision. "Rec.": Recall. ↑ indicates higher values are better.

one, we can adaptively change the contribution of each sub-task to the model update such that all tasks have similar impacts on the learning dynamics [53]. Another promising but computationally expensive method is to leverage population-based training [74], where loss weights are adjusted in the direction of maximizing the overall performance. Admittedly, searching for better neural network architectures that are more suitable for multi-task learning is still a feasible direction. Future work in this thread can leverage tools from neural architecture search [75, 76].



(a) Change of global weights $\mathbf{\Gamma}_G$. (b) Change of local weights $\mathbf{\Gamma}_L$. (c) Change of global and local weights $\mathbf{\Lambda}$.

Figure 4: Change of learnable loss weights.

7 Conclusion

In this work, we rethink previous efforts made for fine-grained loneliness characterization. We improve the baseline method by a large margin by adapting a pre-trained language model to the data distribution we are interested in and then leverage it for downstream tasks with the same data distribution. We then investigate the phenomenon of negative transfer and attempt to reduce it by learning to adaptively weight different sub-tasks. Analysis on experiment results raises open-ended questions that are worth studying in the future. We then shed light on more approaches that are potentially helpful to mitigate the negative transfer.

References

[1] John T Cacioppo and William Patrick. *Loneliness: Human nature and the need for social connection*. WW Norton & Company, 2008.

[2] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.

[3] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

	Acc. ↑	Clark ↓	Canberra ↓	Cosine ↑	Intersection ↑
			Duration		
Baseline (BERT + MLP)	0.5992 ± 0.0114	1.4682 ± 0.0060	2.5670 ± 0.0065	0.7389 ± 0.0058	0.5713 ± 0.0113
Baseline (HDLN)	0.4539 ± 0.0073	1.4622 ± 0.0057	2.5515 ± 0.0112	0.6847 ± 0.0033	0.5293 ± 0.0017
Ours w/ Adapted BERT	**0.5058 ± 0.0331**	**1.4549 ± 0.0026**	**2.5304 ± 0.0079**	**0.7021 ± 0.0040**	**0.5447 ± 0.0062**
Ours w/ Learned Weights	0.4708 ± 0.01145	**1.4428 ± 0.0004**	**2.4896 ± 0.0033**	0.6996 ± 0.0034	0.5335 ± 0.0030
			Context		
Baseline (BERT + MLP)	0.8573 ± 0.0102	1.9507 ± 0.0028	3.9649 ± 0.0083	0.9065 ± 0.0051	0.7864 ± 0.0028
Baseline (HDLN)	0.8560 ± 0.0220	1.9590 ± 0.0018	3.9986 ± 0.0066	0.8920 ± 0.0093	0.7720 ± 0.0090
Ours w/ Adapted BERT	**0.8702 ± 0.0048**	**1.9479 ± 0.0011**	**3.9605 ± 0.0050**	**0.9082 ± 0.0034**	***0.7890 ± 0.0037***
Ours w/ Learned Weights	**0.8702 ± 0.0143**	1.9511 ± 0.0010	3.9831 ± 0.0165	0.9000 ± 0.0090	0.7731 ± 0.0140
			Interpersonal		
Baseline (BERT + MLP)	0.7976 ± 0.0031	1.9077 ± 0.0009	3.8672 ± 0.0034	0.8737 ± 0.0017	0.7229 ± 0.0006
Baseline (HDLN)	0.7795 ± 0.0191	1.9123 ± 0.0015	3.8911 ± 0.0102	0.8542 ± 0.0088	0.6943 ± 0.0148
Ours w/ Adapted BERT	**0.8093 ± 0.0055**	**1.9039 ± 0.0010**	**3.8604 ± 0.0038**	**0.8716 ± 0.0016**	**0.7150 ± 0.0078**
Ours w/ Learned Weights	0.7898 ± 0.0336	**1.9017 ± 0.0005**	**3.8647 ± 0.0107**	0.8647 ± 0.0113	0.6872 ± 0.0192
			Interaction		
Baseline (BERT + MLP)	0.8352 ± 0.0120	1.9643 ± 0.0029	4.0096 ± 0.0112	0.8852 ± 0.0084	0.7612 ± 0.0044
Baseline (HDLN)	0.7237 ± 0.0138	1.9817 ± 0.0014	4.1003 ± 0.0063	0.8154 ± 0.0072	0.6721 ± 0.0063
Ours w/ Adapted BERT	**0.7859 ± 0.0063**	**1.9709 ± 0.0002**	**4.0439 ± 0.0001**	**0.8567 ± 0.0001**	**0.7269 ± 0.0024**
Ours w/ Learned Weights	0.7341 ± 0.0066	1.9776 ± 0.0056	4.0912 ± 0.0241	0.8265 ± 0.0147	0.6712 ± 0.0184

Table 4: Results for distributional fine-grained loneliness characterization. ↑ indicates higher values are better. ↓ indicates lower values are better.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[5] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[6] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[8] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683, 2020.

[9] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *ArXiv*, abs/2005.00547, 2020.

[10] Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. Measuring emotions in the covid-19 real world worry dataset. *ArXiv*, abs/2004.04225, 2020.

[11] Jane A. Yu and Alon Y. Halevy. The care dataset for affective response detection. *ArXiv*, abs/2201.11895, 2022.

[12] Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. Xed: A multilingual dataset for sentiment analysis and emotion detection. In *COLING*, 2020.

[13] Sharath Chandra Guntuku, Rachelle Schneider, Arthur Pelullo, Jami Young, Vivien Wong, Lyle Ungar, Daniel Polsky, Kevin G Volpp, and Raina Merchant. Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ Open*, 9(11), 2019.

[14] Yueyi Jiang, Yunfan Jiang, Liu Leqi, and Piotr Winkielman. Many ways to be lonely: Fine-grained characterization of loneliness and its potential changes in covid-19. *ArXiv*, abs/2201.07423, 2022.

[15] Giwoong Lee, Eunho Yang, and Sung Hwang. Asymmetric multi-task learning based on task relatedness and loss. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 230–238, New York, New York, USA, 20–22 Jun 2016. PMLR.

[16] Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. 2020.

[17] Partoo Vafaeikia, Khashayar Namdar, and Farzad Khalvati. A brief review of deep multi-task learning and auxiliary task learning. *ArXiv*, abs/2007.01126, 2020.

[18] Carlo Strapparava and Rada Mihalcea. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[19] Laura-Ana-Maria Bostan and Roman Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[20] Muhammad Abdul-Mageed and Lyle Ungar. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[21] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555, 2014.

[22] Chao-Chun Hsu and Lun-Wei Ku. SocialNLP 2018 EmotionX challenge overview: Recognizing emotions in dialogues. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 27–31, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[23] Rich Caruana. *Multitask Learning*, pages 95–133. Springer US, Boston, MA, 1998.

[24] Yu Zhang and Qiang Yang. A survey on multi-task learning. *ArXiv*, abs/1707.08114, 2017.

[25] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *ArXiv*, abs/2009.09796, 2020.

[26] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[27] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5454–5463, 2017.

[28] Yiyi Liao, Sarath Kodagoda, Yue Wang, Lei Shi, and Y. Liu. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2318–2325, 2016.

[29] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2014.

[30] Marvin Teichmann, Michael Weber, Johann Marius Zöllner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1013–1020, 2018.

[31] Ishan Misra, Abhinav Shrivastava, Abhinav Kumar Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003, 2016.

[32] Jonas Uhrig, Marius Cordts, Uwe Franke, and Thomas Brox. Pixel-level encoding and depth layering for instance-level semantic labeling. In *GCPR*, 2016.

[33] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *ArXiv*, abs/2104.08212, 2021.

[34] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex X. Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *ArXiv*, abs/1812.00568, 2018.

[35] Anusha Nagabandi, Kurt Konolige, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. *ArXiv*, abs/1909.11652, 2019.

[36] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *CoRL*, 2021.

[37] Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *ArXiv*, abs/1802.01561, 2018.

[38] Girish Joshi and Girish Chowdhary. Cross-domain transfer in reinforcement learning using target apprentice. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7525–7532, 2018.

[39] B. B. D. Silva, George Dimitri Konidaris, and Andrew G. Barto. Learning parameterized skills. In *ICML*, 2012.

[40] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher Joseph Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. *ArXiv*, abs/1804.00079, 2018.

[41] Jeremy Howard and Sebastian Ruder. Fine-tuned language models for text classification. *ArXiv*, abs/1801.06146, 2018.

[42] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *ArXiv*, abs/1806.08730, 2018.

[43] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *ACL*, 2019.

[44] Dani Yogatama, Cyprien de Masson d'Autume, Jerome T. Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. Learning and evaluating general linguistic intelligence. *ArXiv*, abs/1901.11373, 2019.

[45] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-yi Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[46] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *ArXiv*, abs/1605.05101, 2016.

[47] Anders Søgaard and Yoav Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany, August 2016. Association for Computational Linguistics.

[48] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. *ArXiv*, abs/1704.05742, 2017.

[49] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018.

[50] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880, 2019.

[51] Feng Zheng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, and Feiyue Huang. A coarse-to-fine pyramidal model for person re-identification via multi-loss dynamic training. *ArXiv*, abs/1810.12193, 2018.

[52] Sébastien Jean, Orhan Firat, and Melvin Johnson. Adaptive scheduling for multi-task learning. *ArXiv*, abs/1909.06434, 2019.

[53] Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech M. Czarnecki, Simon Schmitt, and H. V. Hasselt. Multi-task deep reinforcement learning with popart. In *AAAI*, 2019.

[54] Sumanth Chennupati, Ganesh Sistu, Senthil Kumar Yogamani, and Samir A. Rawashdeh. Multinet++: Multi-stream feature aggregation and geometric loss strategy for multi-task learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1200–1210, 2019.

[55] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[56] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China, July 2015. Association for Computational Linguistics.

[57] Victor Sanh, Thomas Wolf, and Sebastian Ruder. A hierarchical multi-task approach for learning embeddings from semantic tasks. *ArXiv*, abs/1811.06031, 2019.

[58] Sahil Sharma and Balaraman Ravindran. Online multi-task learning using active sampling. *ArXiv*, abs/1702.06053, 2017.

[59] Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. *CoRR*, abs/1511.06114, 2016.

[60] Ayan Sinha, Zhao Chen, Vijay Badrinarayanan, and Andrew Rabinovich. Gradient adversarial training of neural networks. *ArXiv*, abs/1806.08028, 2018.

[61] Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, 7(59):1601–1626, 2006.

[62] Nicolò Cesa-bianchi, Claudio Gentile, Andrea Tironi, and Luca Zaniboni. Incremental algorithms for hierarchical classification. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

[63] Carlos Silla and Alex Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72, 01 2011.

[64] Svetlana Kiritchenko and Fazel Famili. Hierarchical text categorization as a tool of associating genes with gene ontology codes. *Proceedings of the 2nd European Workshop on Data Mining and Text Mining in Bioinformatics*, 01 2004.

[65] Ricardo Cerri, Rodrigo C. Barros, and Andre C. P. L. F. de Carvalho. A genetic algorithm for hierarchical multi-label classification. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, page 250–255, New York, NY, USA, 2012. Association for Computing Machinery.

[66] Ricardo Cerri, Rodrigo Barros, A Carvalho, and Alex Freitas. A grammatical evolution algorithm for generation of hierarchical multi-label classification rules. pages 454–461, 06 2013.

[67] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5075–5084. PMLR, 10–15 Jul 2018.

[68] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[69] Joshua C. Peterson, R. Battleday, T. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625, 2019.

[70] Ali Akbari, Muhammad Awais, Manijeh Bashar, and Josef Kittler. How does loss function affect generalization performance of deep learning? application to human age estimation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 141–151. PMLR, 18–24 Jul 2021.

[71] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[72] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017.

[73] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv: Learning*, 2017.

[74] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. Population based training of neural networks. *ArXiv*, abs/1711.09846, 2017.

[75] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *J. Mach. Learn. Res.*, 20:55:1–55:21, 2019.

[76] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *ArXiv*, abs/1611.01578, 2017.