# Interpreting and Exploiting Functional Specialization in Multi-Head Attention under Multi-task Learning

**Anonymous ACL submission**

## Abstract

Transformer-based models, even though achieving super-human performance on several downstream tasks, are often regarded as a black box and used as a whole. It is still unclear what mechanisms they have learned, especially their core module: multi-head attention. Inspired by functional specialization in the human brain, which helps to efficiently handle multiple tasks, this work attempts to figure out whether the multi-head attention module will evolve similar function separation under multi-tasking training. If it is, can this mechanism further improve the model performance? To investigate these questions, we introduce an interpreting method to quantify the degree of functional specialization in multi-head attention. We further propose a simple multi-task training method to increase functional specialization and mitigate negative information transfer in multi-task learning. Experimental results on seven pre-trained transformer models have demonstrated that multi-head attention does evolve functional specialization phenomenon after multi-task training which is affected by the similarity of tasks. Moreover, the multi-task training strategy based on functional specialization boosts performance in both multi-task learning and transfer learning without adding any parameters.

## 1 Introduction

Transformer, based on the multi-head attention module, has been the dominant model for downstream applications due to its impressive results (Devlin et al., 2019; Brown et al., 2020; Dosovitskiy et al., 2021). However, it is still being utilized as a whole black-box model, and little is known about the functions of each sub-module on the final prediction. Simultaneously, although controversy still exists, there is overwhelming evidence that supports the idea of functional specialization in the human brain (Finger, 2001; Kanwisher, 2010). Such a functional specialization mechanism makes it easier for the human brain to handle multiple tasks and solve new problems. It can reuse existing resources and at the same time evolve specific regions to avoid the huge cost of redesigning.

Considering the benefits of functional specialization to human learning ability, it is interesting to explore whether a transformer model, especially its central module multi-head attention, would evolve a similar mechanism under multi-task training. If so, which factors will impact the degree of functional specialization in the multi-head attention module? And how to exploit this phenomenon to improve the generalization ability of Transformer-based models?

To investigate these questions, we first propose a method, called Important Attention-head Pruning (IAP), to quantify the degree of functional specialization in the multi-head attention of Transformer-based models. IAP first calculates the importance scores of each attention head on different tasks, then prunes the top important heads for each task to determine their impact on task performance. We apply our method to five different tasks with seven pre-trained transformers. Results show that the multi-head attention module has evolved distinct functional specialization phenomena across different sizes of BERT and pre-training methods. Further quantitative analysis indicates that there is a negative correlation between task similarity and the functional specialization phenomenon.

Moreover, we propose a multi-task learning method, namely Important Attention-head Training (IAT), to promote the segregation of functions in the multi-head attention module by training only the most important part of attention heads for each task. Experimental results on the GLUE dataset have demonstrated that our method alleviates the negative transfer among tasks and improves the performance of Transformer-based models on both multi-task learning and transfer learning without additional parameters.

To summarize, our main contributions are twofold:

- We propose an interpretation method called IAP and find that the functional specialization phenomenon has evolved in multi-head attention after multi-task learning. Furthermore, empirical quantitative experiments show that such a phenomenon is influenced by the similarity between tasks: the more similar tasks are, the weaker the functional specialization phenomenon is.

- We propose an exploiting method called IAT to promote the degree of functional specialization. Experiments on multi-task learning and transfer learning validate that IAT is able to improve both the performance and generalization ability of multi-task learning models without adding any parameters.

## 2 Related Work

### 2.1 Interpreting Neural Networks

**Interpreting attention module** Analogous to visual attention, the distribution of attention weight over input is often used to interpret the final decision of attention-based model (Clark et al., 2019; Vig and Belinkov, 2019). Therefore, a lot of work has been done to study the interpretability of attention distribution (Jain and Wallace, 2019; Serrano and Smith, 2019; Jacovi and Goldberg, 2020) or design better explanation methods (Brunner et al., 2020; Kobayashi et al., 2020; Bai et al., 2021; Liu et al., 2022).

Our work can be classified into another line of study: investigating the individual attention head in the multi-head attention module. Voita et al. (2019) argued that there are redundant heads in Transformer by pruning less important heads and analyzing the resulting performance, which is confirmed by Michel et al. (2019). Jo and Myaeng (2020) analyzed the linguistic properties of the sentence representations from attention heads by ten linguistic probing tasks. Hao et al. (2021) only retained the important heads in BERT and constructed an attribution tree to interpret the information interactions inside Transformer.

Through pruning attention heads, we study the role they play in different tasks, rather than show redundancy in the multi-head attention module (Michel et al., 2019).

**Interpretation inspired by neuroscience** With more understanding of the functional specialization of the human brain, researchers attempt to interpret deep learning models with brain activities in specialized regions (Wehbe et al., 2014; Toneva and Wehbe, 2019; Zhuang et al., 2021; Bakhtiari et al., 2021). For example, Toneva and Wehbe (2019) studied the representations of NLP models across different layers by aligning with two groups of brain areas among the language network.

Unlike the existing works, we investigate whether the brain-like functional specialization phenomenon occurs in NLP models, and how to exploit this phenomenon to improve models.

### 2.2 Mitigating Negative Information Transfer in Multi-task Learning

By joint learning multiple tasks, the performance of a model on the target task can be boosted with regularization or sharing parameters among tasks (Collobert et al., 2011; Ruder, 2017; Liu et al., 2019a). However, multi-task learning models in NLP often suffer from negative information transfer and are inferior to the single task learning ones (Martínez Alonso and Plank, 2017; Bingel and Søgaard, 2017).

Our method aims to subdivide task-important modules in parameters shared to mitigate negative transfer among tasks, which is different from previous sampling or additional task-specific adapter methods (Wu et al., 2020; Pilault et al., 2021). We only need to preserve mask variables for each attention head rather than all parameters during training (Sun et al., 2020; Lin et al., 2021; Xie et al., 2021; Liang et al., 2021), which significantly reduce memory costs.

## 3 Background

### 3.1 Multi-Head Attention Module

Transformer (Vaswani et al., 2017) extended single head attention function to Multi-Head Attention (MHA) module, which aims at capturing information from different representation subspaces in parallel. Given input $X \in \mathbb{R}^{n \times d}$, this module linearly transforms it into $n_h$ subspaces and then applies attention separately:

$$A_h(X) = \text{Attention}(XW_h^Q, XW_h^K, XW_h^V)$$
$$\text{with Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where $Q, K \in \mathbb{R}^{n \times d_k}$ and $V \in \mathbb{R}^{n \times d_v}$. The outputs of all heads are concatenated and linearly trans-

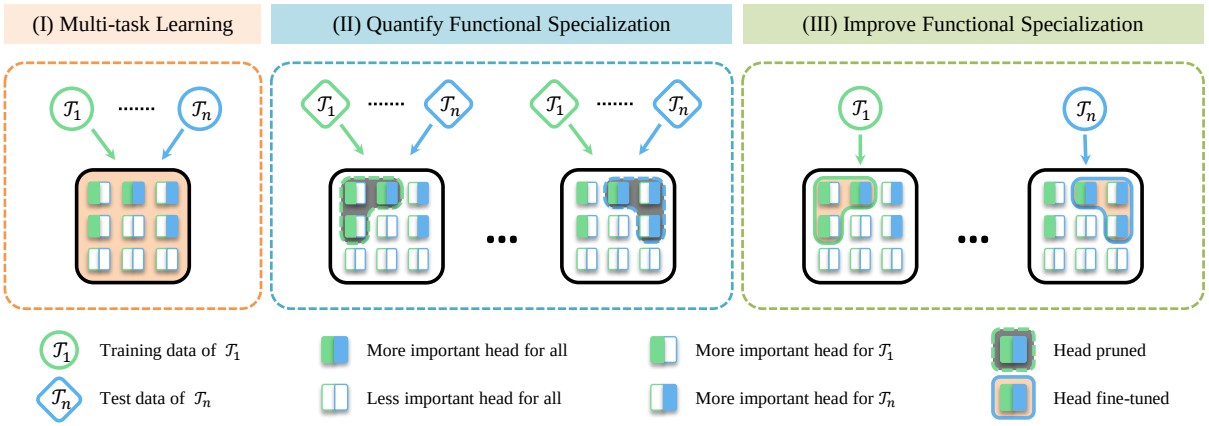


Figure 1: Illustration of how to quantify and improve the degree of functional specialization in multi-head attention for Transformer-based models. Only attention heads, which are our research target, are depicted in the model for simplicity. (I). Multi-task learning using Transformer-based models. (II). Quantify the functional specialization phenomenon by determining and pruning the important heads for each task. (III). Improve the functional specialization phenomenon by only fine-tuning the important heads for each task in the last part of multi-task learning process.

formed into the output space of this module:

$$MHA(X) = [A_1(X); ...; A_{n_h}(X)]W^O \quad (2)$$

### 3.2 Head Importance Score

Michel et al. (2019) proposed an effective method to prune attention heads and evaluate the importance of attention heads for a task. In order to prune the attention head $h$, they incorporated a mask variable $\xi_h \in [0, 1]$ into the attention function:

$$\widetilde{A}_h(X) = \xi_h \cdot A_h(X) \quad (3)$$

and set it to a zero value. When $\xi_h$ equals 1, Equation (3) is the same with the vanilla attention (Eq. (1)). The head importance score $I_h^{(i)}$ of task $\mathcal{T}_i$ is approximated by the expected sensitivity of loss function to the mask variable $\xi_h$:

$$I_h^{(i)} = \mathbb{E}_{(x,y)\sim\mathcal{D}^{(i)}} \left| \frac{\partial \mathcal{L}^{(i)}(x, y)}{\partial \xi_h} \right| \quad (4)$$

where $\mathcal{D}^{(i)}$ is the data distribution of task $\mathcal{T}_i$ and $\mathcal{L}^{(i)}(x, y)$ is the loss of task $\mathcal{T}_i$ on sample $(x, y)$.

Different from Michel et al. (2019) which prune the least important attention heads to prove the redundancy of attention heads, this paper focuses on exploring the functional specialization phenomenon after training, thus we prune the most important heads for each task.

## 4 Method

Figure 1 illustrates the general procedure of our methods. Firstly, Transformer-based models are utilized for multi-task learning and may arise segregation of functions in the multi-head attention module. Subsequently, the important attention heads are determined and pruned to quantify the functional specialization in multi-head attention (Section 4.1). Lastly, the roles of important heads in each task are enhanced to promote the degree of functional specialization by important attention-head training (Section 4.2).

### 4.1 Interpreting: Important Attention-head Pruning

We introduce a two-step method, namely *Important Attention-head Pruning* (IAP), to quantify the degree of functional specialization in multi-head attention. First, the top $\alpha \in [0, 1]$ percentage important heads $H_i^\alpha$ for task $\mathcal{T}_i$, e.g., the ones circled by dashed lines in Figure 1(II), are found after dual-task or multi-task training by their head importance scores. Specifically, we calculate the head importance score $I_h^{(i)}$, defined by Eq. (4), on training samples to approximate the contribution of head $h$ to task $\mathcal{T}_i$.

Second, dissociation experiments are conducted to determine the degree of functional specialization in multi-head attention. Given a model $f_\theta$ after dual-task training on tasks $\mathcal{T}_A$ and $\mathcal{T}_B$, for example, the relative performance on $\mathcal{T}_A$ after pruning the top $\alpha$ important attention heads for $\mathcal{T}_B$, denoted by $H_B^\alpha$, is calculated as follows:

$$RP_A(H_B^\alpha) = \frac{\mathcal{P}\left(f_{\theta \setminus H_B^\alpha}(X_A), Y_A\right)}{\mathcal{P}(f_\theta(X_A), Y_A)} \quad (5)$$

where $\mathcal{P}(\cdot)$ is the performance metric used, e.g., Accuracy, and $(X_A, Y_A)$ is the test samples of Task $\mathcal{T}_A$. Then, we estimate the degree of functional specialization by the relative performance difference after top $\alpha$ important heads for each task are

pruned, called dissociation score:

$$\begin{aligned} D_A(\alpha) &= RP_A(H_B^\alpha) - RP_A(H_A^\alpha), \\ D_B(\alpha) &= RP_B(H_A^\alpha) - RP_B(H_B^\alpha), \\ D(\alpha) &= \frac{D_A(\alpha)+D_B(\alpha)}{2} \end{aligned} \quad (6)$$

where $D_A(\alpha)$ denotes the dissociation score of task $\mathcal{T}_A$, and $D(\alpha)$ is the average dissociation score of this dual-task learning. Given an appropriate $\alpha$, a larger dissociation score implies a higher degree of functional specialization.

Similarly, the dissociation score of task $\mathcal{T}_i$ under multi-task learning is measured via:

$$\begin{aligned} D_i(\alpha) &= \frac{\sum_{j=1,j\neq i}^{n} RP_i(H_j^\alpha)}{n-1} - RP_i(H_i^\alpha), \\ D(\alpha) &= \frac{\sum_{i=1}^{n} D_i(\alpha)}{n} \end{aligned} \quad (7)$$

To clearly illustrate the functional specialization phenomenon, we summarize two representative cases under dual-task learning:

- **Double dissociation** when $D_A(\alpha) > 0$ and $D_B(\alpha) > 0$. This is a significant indicator of functional specialization. That is, each task requires a unique group of heads, which can be selectively masked. To eliminate the accidental functional specialization phenomenon, we argue that a distinct one occurs if the average dissociation score is higher than or equal to 10%, i.e., $D(\alpha) \geq 10\%$, in which 10% is chosen according to the definition of double dissociation in neuroscience (Shallice, 1988).

- **Single dissociation** when $D_A(\alpha) > 10\%$ and $D_B(\alpha) < 0$, or $D_A(\alpha) < 0$ and $D_B(\alpha) > 10\%$. One significant positive dissociation score suggests functional specialization may only arise in this task.

The dissociation scores may be both negatives, which arise from the wrong evaluation of the important heads for each task. It can be summarized into the double dissociation case under the correct evaluation and pruning. In the other cases, e.g, the dissociation scores of both tasks are relatively small, we argue that there is no functional specialization in the multi-head attention module. Specifically, the influence on all tasks will be almost identical when pruning different groups of heads.

### 4.2 Exploiting: Important Attention-head Training

Motivated by the high degree of functional specialization in human brain, it is interesting to investigate whether a higher degree of functional specialization could improve the performance of the model on multi-task learning or transfer learning.

To promote the degree of functional specialization in multi-head attention, we design a multi-task training method, named *Important Attention-head Traning* (IAT). Specifically, only the top $\alpha \in [0, 1]$ important attention heads for task $\mathcal{T}_i$ are tuned at the last $\delta \in [0, 1]$ multi-task training process, and the parameters other than the multi-head attention module are trained as before. To achieve this, we introduce a mask variable $M_i \in \{0, 1\}^{n_h}$ for task $\mathcal{T}_i$, where 1 indicates to fine-tune this attention head for $\mathcal{T}_i$. For example in Figure 1(III), only the mask variables of heads circled by the solid blue line are set to 1 for $\mathcal{T}_n$. When $\alpha = 1$ or $\delta = 0$, our method is the same as the normal multi-task learning method.

We expect to consolidate the roles of important heads for each task and facilitate the functional separation of multi-head attention in this way.

## 5 Experimental Setup

### 5.1 Datasets

We select a topic classification datasets (Zhang et al., 2015), eight natural language understanding datasets of GLUE (Williams et al., 2018; Rajpurkar et al., 2016; Wang et al., 2019), and two datasets (Maas et al., 2011; Khot et al., 2018) for transfer learning in this study. To avoid an extreme ratio of training samples between tasks, only five large datasets in different tasks, which contain more than 10k training samples, are preserved in dual-task and multi-task learning interpretation experiments. Like Karimi Mahabadi et al. (2021), SciTail and IMDB are used only in transfer learning. Statistics of all datasets used are shown in Table 1.

| Task | Dataset | #Class | #Train |
|---|---|---|---|
| Topic Classification | AG's News$^\star$ | 4 | 120,000 |
| Acceptability | CoLA | 2 | 8,551 |
| Natural Language Inference | MNLI$^\star$ | 3 | 392,702 |
| Paraphrase | QQP$^\star$ | 2 | 363,846 |
| Paraphrase | MRPC | 2 | 3,668 |
| Question Answering | QNLI$^\star$ | 2 | 104,743 |
| Sentiment Analysis | SST-2$^\star$ | 2 | 67,349 |
| Entailment | RTE | 2 | 2,490 |
| Textual Similarity | STS-B | - | 5,749 |
| Natural Language Inference | SciTail | 2 | 23,596 |
| Sentiment Analysis | IMDB | 2 | 25,000 |

Table 1: Statistic of datasets used. $^\star$ denotes dataset used in dual-task and multi-task interpretation experiments.


Figure 2: Average dissociation scores of different Transformer-based models (y-axis) after ten dual-task learning tasks (x-axis) with $\alpha = 30\%$. The larger dissociation score implies a higher degree of functional specialization in multi-head attention (Section 4.1). All dissociation scores are reported in Table 8. * indicates that the parameters of $\text{BERT}_{\text{BASE}}$ encoder are **frozen**, i.e., the output layers are fine-tuned only.

### 5.2 Models

As shown in Table 2, seven Pre-trained Transformer Models (PTMs), including GPT family models, different sizes of BERT and different pre-training methods (Radford et al., 2018, 2019; Devlin et al., 2019; Liu et al., 2019b; Jiao et al., 2020; He et al., 2021), are investigated in this paper. These models are all initialized from the transformer library of HuggingFace (Wolf et al., 2019). Hyperparameters are reported in Appendix A. Codes will be published to facilitate future work after acceptance.

| Model | #L | #A | #L × #A | Parameters |
|---|---|---|---|---|
| GPT | 12 | 12 | 144 | 110M |
| GPT-2 | 24 | 16 | 384 | 355M |
| TinyBERT | 6 | 12 | 72 | 67M |
| $\text{BERT}_{\text{BASE}}$ | 12 | 12 | 144 | 110M |
| $\text{RoBERTa}_{\text{BASE}}$ | 12 | 12 | 144 | 125M |
| $\text{DeBERTaV3}_{\text{BASE}}$ | 12 | 12 | 144 | 184M |
| $\text{BERT}_{\text{LARGE}}$ | 24 | 16 | 384 | 340M |

Table 2: Statistic of models used. #L=the number of layers, #A=the number of attention heads per layer.

## 6 Experiments and Results

### 6.1 Functional Specialization Does Evolve in Multi-head Attention

**Dual-task Learning** Based on the pairwise combination of five datasets, there are ten groups of dual-task learning tasks. We observe that the dissociation scores of models without frozen in dual-task learning are all positive, i.e., double dissociation phenomenon appears in all task-pairs (details are shown in Appendix B). As illustrated in Figure 2, $\text{BERT}_{\text{BASE}}$ shows a distinct functional specialization phenomenon ($D(\alpha) > 10\%$) in four dual-task learning tasks. Moreover, distinct functional specialization phenomena are also found in the other two sizes of BERT and GPT models. The other two base-size models, $\text{RoBERTa}_{\text{BASE}}$ and $\text{DeBERTV3}_{\text{BASE}}$, even show a higher degree of functional specialization, in which average dissociation scores among ten dual-task learning tasks are 13.44% and 10.88% respectively.

| Prune Task | MNLI | QQP | QNLI | AG | SST-2 |
|---|---|---|---|---|---|
| $\text{MNLI}^{\dagger}$ | 58.23 | 71.52 | 61.39 | 91.99 | 85.32 |
| $\text{QQP}^{\dagger}$ | 62.54 | 69.43 | 60.80 | 91.54 | 85.13 |
| $\text{QNLI}^{\dagger}$ | 59.29 | 70.96 | 57.50 | 91.88 | 86.35 |
| $\text{AG}^{\dagger}$ | 65.88 | 76.28 | 69.35 | 80.01 | 85.09 |
| $\text{SST-2}^{\dagger}$ | 69.50 | 77.40 | 73.96 | 86.51 | 82.45 |
| $\text{Random}^{\dagger}$ | 80.68 | 85.42 | 85.05 | 93.65 | 91.23 |
| **Base** | 83.91 | 87.64 | 90.26 | 94.50 | 92.05 |
| $D_i(\alpha)$ | 7.28 | 5.26 | 11.07 | 9.84 | 3.28 |

Table 3: Performance(%) of the pruned and base model on each task using $\text{BERT}_{\text{BASE}}$ with $\alpha = 30\%$. $\mathcal{T}^{\dagger}$ denotes top $\alpha$ important heads for this task are pruned. The lowest value is underlined.

To eliminate the accidental functional specialization phenomenon, we train another dual-task model using a frozen $\text{BERT}_{\text{BASE}}$ encoder for comparison. As shown in the fifth row of Figure 2, most of the dissociation scores are relatively small and only one dual-task pair, "MNLI and AG", shows a mild functional specialization phenomenon ($D(\alpha) > 5\%$). The average dissocia-

| | | Performance on Task A | | | Performance on Task B | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Task A | Task B | Base Acc. | Task A$^{\dagger}$ | Task B$^{\dagger}$ | Base Acc. | Task A$^{\dagger}$ | Task B$^{\dagger}$ | $D_A(30\%)$ | $D_B(30\%)$ | $D(30\%)$ |
| AG | QNLI | 94.13 | 85.94 | 92.44 | 91.13 | 65.04 | 52.95 | 6.905 | 13.270 | 10.088 |
| AG-Pair | QNLI | 94.46 | 56.17 | 64.52 | 90.72 | 64.85 | 53.66 | 8.842 | 12.337 | 10.590 |
| AG | SST-2 | 94.29 | 89.51 | 92.32 | 92.47 | 89.76 | 86.01 | 2.982 | 4.051 | 3.517 |
| AG-Pair | SST-2 | 94.68 | 67.65 | 71.80 | 92.66 | 89.33 | 85.78 | 4.387 | 3.837 | 4.112 |

Table 4: Comparison between different input paradigm combinations. The input of AG-Pair is a pair of sentences from AG, and the label is whether they belong to the same topic.

tion score of these ten task pairs spontaneously increases by 6.32% if we fine-tune the shared encoder.

**Multi-task Learning** We further conduct multi-task learning experiments using all five tasks in dual-task learning. In addition to all positive dissociation scores, we find that the performance of one task decreases more when pruning the top 30% important attention heads of this task compared with other tasks (Table 3). It shows that the functional specialization phenomenon has evolved after multi-task learning, i.e., there is a unique group of heads more important to one specific task. Otherwise, the influence on all tasks would be similar when pruning the most important attention heads for different tasks.

The absolute performances on the first three tasks (MNLI, QQP, and QNLI) suffer a drastic drop after pruning only 30% attention heads. For example, the lowest drop is 14.41% on MNLI when the top 30% important heads for SST-2 are pruned, while the highest one is only 14.49% among the AG and SST-2 when pruning the same amount of attention heads. It indicates that tasks taking two sequences as input, e.g., natural language inference and question answering, depend on attention mechanism more than one sequence input task, which is in line with the finding of Vashishth et al. (2019). See Appendix C for more details and analyses.

### 6.2 Task Similarity Affects Functional Specialization

After observing the functional specialization phenomenon in the multi-head module, it is interesting to study how this phenomenon is affected. In this section, we empirically explore two factors: task similarity and input paradigm.

**Task Similarity** The task similarity metric *Cognitive-Neural Mapping* (CNM), which is found less sensitive to underlying models (Luo et al., 2022), is utilized to quantify the similarity of task-pair in this section.

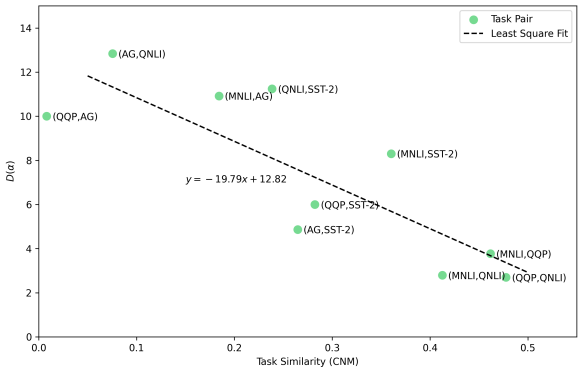


Figure 3: The average dissociation score and similarity of each task-pair in multi-task learning.

As shown in Figure 3, we observe that there is a significant negative correlation between the average dissociation score of task-pair and the similarity between tasks. In other words, the more similar the tasks are, the lower the average dissociation score is, which suggests the weaker the functional head specialization phenomenon is. The other three task similarity metrics used and fitting results refer to Appendix D, where this negative relationship is also found.

**Input Paradigm** There are two different input paradigms, sentence pair (MNLI, QNLI, and QQP) and single sentence (AG and SST-2), among these five tasks. We notice the average dissociation score of two tasks in different input paradigms is higher than the same input paradigm ones in Figure 2 ($\text{BERT}_{\text{BASE}}$: $12.654\% > 3.016\%$). Thus, experiments are conducted to investigate the effect of input paradigm on the degree of functional specialization in multi-head attention. Specifically, we construct a dataset named "AG-Pair" using the sentences of AG dataset, which aims to identify whether a pair of input sentences belong to the same topic. The number of samples in AG-Pair is the same as AG, which is 120k, and each sample in AG occurs twice in the AG-Pair dataset. The generation method and statistics of AG-Pair are reported in Appendix E.

| Model | Type | #Params | CoLA Mcc | MNLI-(m/mm) Acc | MRPC F1 | QNLI Acc | QQP F1 | RTE Acc | SST-2 Acc | STS-B $r^s$ | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TinyBERT‡ | ST | 9.0× | 46.3 | 83.0/82.4 | 85.1 | 90.0 | 70.7 | 65.6 | 92.9 | 84.6 | 77.8 |
| TinyBERT | MTL | 1.0× | 35.2 | **82.6**/**81.9** | 83.4 | **90.5** | 70.2 | 74.0 | 92.5 | 83.5 | 77.1 |
| +IAT | MTL | 1.0× | **39.3** | 82.5/**81.9** | **85.4** | 90.3 | **70.5** | **74.1** | **92.7** | **84.2** | **77.9** |
| $\text{BERT}^1_{\text{BASE}}$ | ST | 9.0× | 52.1 | 84.6/83.4 | 88.9 | 90.5 | 71.2 | 66.4 | 93.5 | 85.8 | 79.6 |
| PALs[2] | MTL | 1.13× | 51.2 | 84.3/83.5 | 88.7 | 90.0 | 71.5 | 76.0 | 92.6 | 85.8 | 80.4 |
| $\text{CA-MTL}^3_{\text{BASE}}$ | MTL | 1.12× | 53.1 | 85.9/85.8 | 88.6 | 90.5 | 69.2 | 76.4 | 93.2 | 85.3 | 80.9 |
| $\text{Ticket-Share}^‡_{\text{BASE}}$ | MTL | 1.0× | 50.3 | 83.7/83.0 | 88.0 | 90.5 | 70.5 | 76.6 | 93.7 | 84.8 | 80.1 |
| $\text{BERT}_{\text{BASE}}$ | MTL | 1.0× | 49.8 | **83.9**/**83.4** | 86.4 | 89.9 | 70.3 | 76.0 | 93.2 | 85.7 | 79.8 |
| +IAT | MTL | 1.0× | **51.6** | **83.9**/83.1 | **87.6** | **90.6** | **71.2** | **76.8** | **94.1** | **86.2** | **80.6** |
| $\text{BERT}^1_{\text{LARGE}}$ | ST | 9.0× | 60.5 | 86.7/85.9 | 89.3 | 92.7 | 72.1 | 70.1 | 94.9 | 86.5 | 82.1 |
| Adapters-256[4] | ST | 1.3× | 59.5 | 84.9/85.1 | 89.5 | 90.7 | 71.8 | 71.5 | 94.0 | 86.9 | 80.0 |
| $\text{CA-MTL}^3_{\text{LARGE}}$ | MTL | 1.12× | 59.5 | 85.9/85.4 | 89.3 | 92.6 | 71.4 | 79.0 | 94.7 | 87.7 | 82.8 |
| $\text{Ticket-Share}^‡_{\text{LARGE}}$ | MTL | 1.0× | 56.2 | 86.0/85.6 | 88.7 | 92.7 | 71.4 | 78.8 | 94.5 | 85.6 | 82.2 |
| $\text{BERT}_{\text{LARGE}}$ | MTL | 1.0× | 56.8 | **85.6**/84.9 | 86.6 | **92.4** | 71.3 | 79.0 | 94.3 | 86.2 | 81.9 |
| +IAT | MTL | 1.0× | **60.0** | 85.5/**85.3** | **88.4** | 92.1 | **71.5** | **79.1** | **94.5** | **86.8** | **82.6** |
| $\text{RoBERTa}^‡_{\text{BASE}}$ | ST | 9.0× | 60.0 | 87.2/86.7 | 90.8 | 93.1 | 72.1 | 71.9 | 95.7 | 88.2 | 82.9 |
| $\text{RoBERTa}_{\text{BASE}}$ | MTL | 1.0× | 55.3 | **87.2**/86.7 | 89.6 | 92.3 | 71.4 | 80.0 | 95.1 | **87.1** | 82.7 |
| +IAT | MTL | 1.0× | **59.9** | 86.9/**86.8** | **90.9** | **92.4** | **71.8** | **80.5** | **95.4** | **87.1** | **83.5** |
| $\text{DeBERTaV3}^‡_{\text{BASE}}$ | ST | 9.0× | 67.1 | 90.0/89.2 | 90.6 | 94.4 | 73.9 | 81.5 | 96.2 | 88.9 | 85.8 |
| $\text{DeBERTaV3}_{\text{BASE}}$ | MTL | 1.0× | 63.1 | **89.9**/**89.2** | 89.4 | **93.8** | 73.7 | 86.7 | 95.5 | 89.6 | 85.7 |
| +IAT | MTL | 1.0× | **67.2** | 89.7/**89.2** | **90.9** | **93.8** | **74.0** | **86.9** | **95.8** | **89.7** | **86.4** |

Table 5: GLUE test set results using the GLUE evaluation server. "ST" stands for the single task fine-tuned model, whereas "MTL" denotes the multi-task learning model. The multi-task learning models we tested are not further fine-tuned on each task, so there is only one model for all tasks (1.0× in #Params). Results from: Devlin et al. (2019)[1], Stickland and Murray (2019)[2], Pilault et al. (2021)[3], Houlsby et al. (2019)[4]. ‡ indicates our implement result for a fair comparison. The highest performance in the last two conditions of each model is displayed in **bold**.

As shown in Table 4, there is no significant dissociation score difference between "AG + QNLI" and "AG-Pair + QNLI" dual-task learning tasks, which also holds for "AG + SST-2" and "AG-Pair + SST-2". We note that the absolute performances on the AG-Pair dataset suffer a drastic drop after pruning only 30% attention heads, which is similar to the other tasks taking pair of sentences as input (Table 3).

According to the experimental results presented above, we observe that task similarity plays a more important role than the input paradigm in the functional specialization of the multi-head attention module.

### 6.3 Improving Multi-Task Models by Training Important Attention Heads

Once the importance of attention heads for each task is figured out, we should be able to consolidate their roles by only finetuning them. Thus, Important Attention-head Training (IAT) (Section 4.2) is applied to the multi-task learning models on 9 GLUE datasets and compared against vanilla multi-task learning. We observe that the degree of functional specialization in the multi-head attention module is improved by training the top important attention heads during the last part of multi-task learning (details refer to Appendix F).

Table 5 reports on a comparison result of single task fine-tuning models, multi-task learning models as well as the models using adapters on GLUE test set.[1] GPT and GPT-2 are not incorporated due to their inferior performance on GLUE. With important attention-head training, the average performances of five multi-task learning models are increased by 0.76% on average over the vanilla multi-task learning baseline. These transformer family models for multi-task learning even surpass their single task fine-tuning counterparts, which consist of 9 task-specific models.

In most cases, multi-task learning models with IAT receive a performance gain on the four small datasets (CoLA, MRPC, RTE, and STS-B), among which the improvement on CoLA is the most significant (+3.6% on average). It comes from the allevi-

[1]For a fair comparison, we treat MNLI-m and MNLI-mm as two tasks, which is the same as Houlsby et al. (2019) and Pilault et al. (2021).

| Model | # Samples of SciTail | | | | # Samples of IMDB | | | |
|---|---|---|---|---|---|---|---|---|
| | 4 | 16 | 32 | 100 | 4 | 16 | 32 | 100 |
| MT-DNN | $71.83_{\pm 6.5}$ | $81.24_{\pm 3.8}$ | $82.59_{\pm 2.3}$ | $85.90_{\pm 2.0}$ | $77.65_{\pm 4.7}$ | $80.76_{\pm 3.1}$ | $82.98_{\pm 0.9}$ | $83.65_{\pm 0.6}$ |
| Ticket-Share$_{\text{BASE}}$ | $73.17_{\pm 6.1}$ | $82.07_{\pm 4.0}$ | $83.05_{\pm 2.4}$ | $86.22_{\pm 1.5}$ | $78.43_{\pm 4.0}$ | $81.57_{\pm 1.6}$ | $83.07_{\pm 0.6}$ | $83.84_{\pm 0.5}$ |
| BERT$_{\text{BASE}}$ | $69.44_{\pm 8.9}$ | $79.41_{\pm 4.7}$ | $81.52_{\pm 3.0}$ | $85.65_{\pm 1.6}$ | $72.21_{\pm 6.2}$ | $78.67_{\pm 3.5}$ | $82.10_{\pm 1.0}$ | $83.39_{\pm 0.5}$ |
| +IAT | $\mathbf{75.66}_{\pm 4.0}$ | $\mathbf{82.11}_{\pm 3.2}$ | $\mathbf{83.82}^{*}_{\pm 1.9}$ | $\mathbf{86.60}^{*}_{\pm 1.3}$ | $\mathbf{80.50}^{*}_{\pm 2.6}$ | $\mathbf{82.03}^{*}_{\pm 1.9}$ | $\mathbf{83.33}^{*}_{\pm 0.5}$ | $\mathbf{84.08}^{*}_{\pm 0.3}$ |

Table 6: Few-shot transfer learning results on development sets across 30 seeds (* indicates statistically significant improvements of 5% level). All models use BERT$_{\text{BASE}}$ as encoder and are initialized from their multi-task learning models on GLUE.

ation of negative transfer in CoLA under multi-task learning. For example, compared with fine-tuning on CoLA (60.5%), the performance of BERT$_{\text{LARGE}}$ drops to 56.8% under multi-task learning, while it increases to 60.0% after using IAT. The performances of multi-task learning models on two large datasets, QQP and SST-2, are also improved by our method. More results, including different sampling methods and performances on GLUE development sets, are shown in Appendix F.

**Few-shot Transfer Learning** Furthermore, we investigate whether a multi-task learning model with a more specialized multi-head attention module will be better at transfer learning. Table 6 presents the few-shot transfer learning results using different amounts of training samples from SciTail (natural language inference task) and IMDB (sentiment analysis task). We find that the model initialized from a multi-task learning model using IAT achieves a higher accuracy on the new task, especially when fewer samples are provided. IAT degrades to the multi-task learning method proposed by Liang et al. (2021) when $\delta = 1$, and often obtains a worse performance in multi-task learning and transfer learning (Ticket-Share in Table 5 and Table 6). It may come from the weak functional specialization phenomenon in the original pre-trained models (e.g., the frozen BERT$_{\text{BASE}}$ encoder in Figure 2), which makes it harder to correctly determine the most important attention heads for each task at the beginning of multi-task training.

**Ablation Study** To take a deep look into the improvements contributed by important attention-head training, we conduct an ablation study on GLUE dev set using BERT$_{\text{BASE}}$ (Table 7). After pruning the least important 30% heads, there is a performance gain on three tasks (MRPC, SST-2, and STS-B), which is in line with the previous finding that Transformer can be improved by pruning some redundant attention heads (Michel et al., 2019).

| Model | Avg | # Tasks Improved |
|---|---|---|
| BERT$_{\text{BASE}}$ | $82.64_{\pm 0.09}$ | - |
| w/ Prune the least important 30% heads | $82.23_{\pm 0.46}$ | 3 |
| w/ Randomly train 30% heads | $82.71_{\pm 0.10}$ | 6 |
| w/ Train the most important 30% heads | $\mathbf{83.41}_{\pm 0.20}$ | 8 |

Table 7: Ablation study of different multi-task methods on GLUE dev set with $\delta = 10\%$.

It is interesting to find that multi-task models can benefit from randomly training 30% attention heads for each task, which may arise from the mitigation of gradient interference by subdividing the parameters shared. Compared with randomly training 30% attention heads, training the most important part of attention heads can further improve the average performance and benefit more tasks.

## 7 Conclusions and Future Work

In this paper, we conduct extensive dissociation experiments and observe that the brain-like functional specialization phenomenon does evolve in multi-head attention after dual-task or multi-task learning. Furthermore, experimental results show that the performance and generalization ability of multi-task models can both be improved by the multi-task training method based on functional specialization. This work, inspired by neuroscience findings, studies the interpretation and improvement of neural networks, which we hope will promote more efforts on interdisciplinary work combining neuroscience and artificial intelligence.

In the future, we plan to investigate more neural network modules that may arise the functional specialization phenomenon under multi-task learning. Another direction is to design better methods exploiting this phenomenon to further improve multi-task learning models.

## Limitations

Firstly, we conduct extensive experiments on multiple natural language understanding tasks only, and multi-modal tasks could be investigated further.

In addition, only one approach is utilized to estimate the importance of each attention head, and the most important attention heads are pruned at once. Because of this choice, our results can be seen as a lower bound on the estimation of functional specialization in multi-head attention. We acknowledge that there might be methods to show higher dissociation scores, such as adopting other attention head importance estimation methods (Hao et al., 2021; Li et al., 2021) or iterative pruning.

We note that the four similarity metrics used in this study are model-dependent, and recognize that results might be different for other Transformer-based models.

Lastly, there are two hyper-parameters introduced in our multi-task training method, which may need extra tuning when adapted to other multi-task learning settings.

## References


Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why attentions may not be interpretable? In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 25–34. ACM.

Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. 2021. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 25164–25178. Curran Associates, Inc.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Stanley Finger. 2001. *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press, USA.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963–12971.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417, Online. Association for Computational Linguistics.

Nancy Kanwisher. 2010. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25):11163–11170.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 565–576, Online. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2021. Differentiable subset pruning of transformer heads. *Transactions of the Association for Computational Linguistics*, 9:1442–1459.

Chen Liang, Simiao Zuo, Minshuo Chen, Haoming Jiang, Xiaodong Liu, Pengcheng He, Tuo Zhao, and Weizhu Chen. 2021. Super tickets in pre-trained language models: From model compression to improving generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6524–6538, Online. Association for Computational Linguistics.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305, Online. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. 2022. Rethinking attention-model explainability through faithfulness violation test. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13807–13824. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yifei Luo, Minghui Xu, and Deyi Xiong. 2022. CogTaskonomy: Cognitively inspired task taxonomy is beneficial to transfer learning in NLP. In *Proceedings of the 60th Annual Meeting of the Association*

*for Computational Linguistics (Volume 1: Long Papers)*, pages 904–920, Dublin, Ireland. Association for Computational Linguistics.

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

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53, Valencia, Spain. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13.

Jonathan Pilault, Amine El hattami, and Christopher Pal. 2021. Conditionally adaptive multi-task learning: Improving transfer learning in NLP using fewer parameters & less data. In *International Conference on Learning Representations*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Tim Shallice. 1988. *From neuropsychology to mental structure*. Cambridge University Press.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

Tianxiang Sun, Yunfan Shao, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Learning sparse sharing architectures for multiple tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8936–8943.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Sen Wu, Hongyang R. Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5725–5737, Online. Association for Computational Linguistics.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. 2021. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118.

# A Hyperparameters

## A.1 Dual-task and Multi-task Learning

To fine-tune the pre-trained models on dual-task or multi-task learning, we use Adam optimizer (Kingma and Ba, 2015), in which $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a learning rate of 2e-5. We also use a linear warm-up schedule and set the warm-up proportion to 0.1. The number of epochs is empirically set to 5 for a fair comparison. The only exception is the distillation of TinyBERT, which contains *intermediate layer distillation* and *prediction layer distillation*. Under the supervision of a fine-tuned $\text{BERT}_{\text{BASE}}$, these distillation methods are performed for 2 and 3 epochs without augmented data, respectively. Unless otherwise specified, the proportional sampling method is utilized in multi-task learning.

Similar to the difference in area between cortical regions, the best $\alpha$ for each task may be different in dissociation experiments. We acknowledge that higher dissociation scores can be obtained by fine-tuning $\alpha$ in each dual-task learning task. For a fair comparison, $\alpha$ is empirically set to 30% in all dissociation experiments to show the extent of functional specialization in the multi-head attention module. All experiments are repeated under three random seeds and average results are reported.

## A.2 Transfer Learning

Since only a small part of training samples are used in transfer learning experiments, we increase the number of training epochs to 20, and conduct a paired bootstrap statistical test under 30 random seeds (Dror et al., 2018).

# B Dual-task Learning Experiments

In this section, we present the results of all multi-head attention based models investigated in dual-task learning tasks.

As reported in Table 8, the dissociation scores of Transformer-based models in dual-task learning are all positive when fine-tuning the pre-trained encoder, i.e., double dissociation phenomenon appears in all task-pairs. It further demonstrates that the functional specialization phenomenon does appear in the multi-head attention module after training on these dual-task learning tasks.

# C Multi-task Learning Experiments on $\text{BERT}_{\text{BASE}}$

We report more results of multi-task learning experiments conducted in Section 6.2. The pair-wise dissociation scores are reported in Table 9.

**Distribution of Heads Pruned** To gain more insights about the functional specialization in multi-task learning, we statistic the distribution of heads pruned for each task across layers in multi-task learning (Figure 4). The average number of attention heads pruned shows a trend of increasing first and then decreasing, which changes at the 4th layer.

| | MNLI$_A$ | | | | | | | | QQP$_A$ | | | | | | QNLI$_A$ | | | | AG$_A$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | QQP$_B$ | | QNLI$_B$ | | AG$_B$ | | SST-2$_B$ | | QNLI$_B$ | | AG$_B$ | | SST-2$_B$ | | AG$_B$ | | SST-2$_B$ | | SST-2$_B$ | |
| **Model** | $D_A$ | $D_B$ | $D_A$ | $D_B$ | $D_A$ | $D_B$ | $D_A$ | $D_B$ | $D_A$ | $D_B$ | $D_A$ | $D_B$ | $D_A$ | $D_B$ | $D_A$ | $D_B$ | $D_A$ | $D_B$ | $D_A$ | $D_B$ |
| GPT | 0.28 | 3.76 | 2.10 | 1.36 | 16.75 | 10.47 | 16.49 | 1.13 | 1.36 | 0.80 | 18.19 | 3.60 | 25.19 | 1.08 | 28.74 | 3.51 | 31.34 | 0.75 | 2.51 | 4.08 |
| GPT-2 | 4.76 | 3.26 | 3.09 | 0.71 | 15.44 | 3.29 | 17.69 | 5.79 | 12.64 | 0.66 | 19.24 | 2.39 | 50.19 | 4.18 | 18.62 | 4.69 | 16.49 | 6.81 | 3.60 | 6.61 |
| TinyBERT | 2.04 | 20.10 | 3.82 | 1.19 | 35.66 | 5.36 | 31.85 | 4.71 | 14.42 | 0.63 | 8.70 | 6.54 | 24.39 | 2.35 | 14.87 | 6.28 | 35.59 | 1.63 | 1.08 | 1.59 |
| BERT$_{\text{BASE}}$ | 2.78 | 2.60 | 2.34 | 2.63 | 11.36 | 16.26 | 13.81 | 0.21 | 1.92 | 4.82 | 28.97 | 10.39 | 11.34 | 3.18 | 13.27 | 6.91 | 32.87 | 3.28 | 2.98 | 4.05 |
| BERT$^*_{\text{BASE}}$ | −0.96 | 0.35 | −0.13 | 1.21 | 5.59 | 4.53 | 3.27 | 1.20 | 0.01 | 0.15 | 0.82 | 3.94 | −1.07 | 6.95 | 6.74 | 0.90 | 7.68 | 2.14 | 2.48 | 3.90 |
| BERT$_{\text{LARGE}}$ | 2.43 | 0.83 | 5.69 | 3.55 | 17.31 | 6.16 | 19.83 | 9.81 | 0.02 | 5.90 | 22.40 | 2.88 | 21.20 | 6.52 | 17.90 | 9.56 | 20.44 | 8.73 | 7.72 | 7.23 |
| RoBERTa$_{\text{BASE}}$ | 8.56 | 10.86 | 1.31 | 4.32 | 17.43 | 38.23 | 5.48 | 8.40 | 2.43 | 0.54 | 23.27 | 22.73 | 22.19 | 5.59 | 15.96 | 47.73 | 16.09 | 5.01 | 6.62 | 6.16 |
| DeBERTaV3$_{\text{BASE}}$ | 6.21 | 4.64 | 14.21 | 0.02 | 24.11 | 8.47 | 27.03 | 2.65 | 9.50 | 0.02 | 11.65 | 9.50 | 20.51 | 13.33 | 3.42 | 16.92 | 20.36 | 2.33 | 4.97 | 17.78 |

Table 8: Results in dual-task learning experiments under $\alpha = 30\%$. * indicates that the parameters of BERT$_{\text{BASE}}$ encoder are **frozen**

| **Task** | MNLI$_A$ | QQP$_A$ | QNLI$_A$ | AG$_A$ | SST-2$_A$ |
|---|---|---|---|---|---|
| MNLI$_B$ | - | 2.385 | 4.321 | **12.671** | 3.115 |
| QQP$_B$ | 5.161 | - | 3.657 | 12.198 | 2.907 |
| QNLI$_B$ | 1.275 | 1.746 | - | 12.555 | **4.236** |
| AG$_B$ | 9.169 | 7.819 | 13.135 | - | 2.865 |
| SST-2$_B$ | **13.496** | **9.091** | **18.246** | 6.874 | - |

Table 9: $D_A(\alpha)$ between task-pairs, which is calculated on the pruning results of multi-task learning with $\alpha = 30\%$. The highest dissociation score in each task A is displayed in **bold**, and the lowest one is underlined.

| **Task** | MNLI$_A$ | QQP$_A$ | QNLI$_A$ | AG$_A$ | SST-2$_A$ |
|---|---|---|---|---|---|
| MNLI$_B$ | - | **81.40** | **83.70** | 58.14 | **68.99** |
| QQP$_B$ | 81.40 | - | 78.29 | 63.57 | 62.79 |
| QNLI$_B$ | **83.70** | 78.29 | - | 62.02 | 62.02 |
| AG$_B$ | 58.14 | 63.57 | 62.02 | - | 64.34 |
| SST-2$_B$ | 68.99 | 62.79 | 62.02 | **64.34** | - |

Table 10: The overlapping percentage of important heads pruned in multi-task learning under $\alpha = 30\%$. The highest overlapping in each task A is displayed in **bold**, and the lowest one is underlined.

The two layers with the greatest difference among tasks are the first layer ($\sigma = 2.39$) and the sixth layer ($\sigma = 2.08$) of BERT$_{\text{BASE}}$ after fine-tuning 5 epochs on these five tasks.

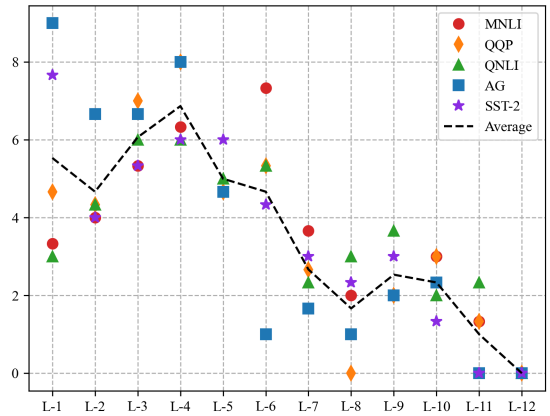


Figure 4: The number of important heads pruned among the layers of BERT$_{\text{BASE}}$ after multi-task learning. The average number of heads pruned in one layer is 3.6 ($\alpha = 30\%$).

**Overlapping of Heads Pruned** Table 10 reports the overlapping of attention heads pruned between tasks. It seems that the proportion of overlapping heads pruned does not completely correspond to the dissociation score of each task (Table 9). For example, as for the MNLI and AG tasks, the task with the highest overlapping of heads pruned is the same as the one with the lowest dissociation score. However, the highest overlapping of heads pruned for SST-2 comes to the second-highest dissociation score when combined with MNLI.

## D Task Similarity Metrics and Fitting Results

To verify the robustness of our finding in Section 6.2, the following four metrics are adopted to determine the similarity of each task pair:

**Direct Similarity Estimation (DSE)** This method approximates the similarity of task pairs by the average similarity of sentence representations from models fine-tuning on the corresponding task. Therefore, we randomly select 1000 sentences from the Wikipedia corpus and adopt cosine similarity to quantify the similarity of sentence representations. Results with DSE metric are shown in Figure 5.

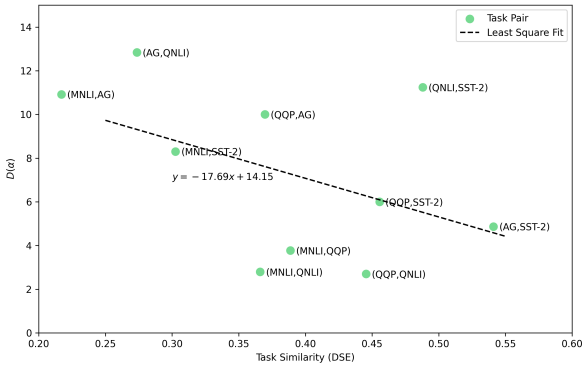


Figure 5: The average dissociation score and DSE similarity of each task pair in multi-task learning.

**Analytic Hierarchy Process (AHP)** On the other hand, the similarity of task pairs can be approximated from the pair-wise transfer learning results (Zamir et al., 2018). Given a target task, models transferred from different source tasks are compared on a hold-out dataset to determine the transferability of the target task, which is further used to approximate the similarity between tasks. Results using AHP are illustrated in Figure 6.

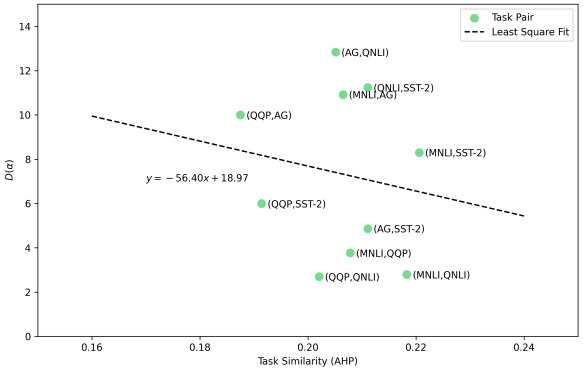


Figure 6: The average dissociation score and AHP similarity of each task pair in multi-task learning.

**Cognitive Representation Analytics (CRA)** Inspired by Representational Similarity Analysis (RSA) in cognitive neuroscience (Kriegeskorte et al., 2008), CRA first calculates the Representation Dissimilarity Matrix (RDM) by the dissimilarity of sentence representations, then approximates the similarity between tasks by the similarity between the corresponding RDMs (Luo et al., 2022). Figure 7 presents the results with CRA.

**Cognitive-Neural Mapping (CNM)** CNM calculates the task similarity by mapping sentence representations of fine-tuned models to fMRI data (Luo et al., 2022), which is recorded when 5 participants were intently reading presented 384 passages (Pereira et al., 2018). Different from randomly selecting 25k fMRI voxels, the most informative 5k fMRI voxels for each participant are used to predict the similarity among tasks. Results with CNM have been shown in Figure 3.

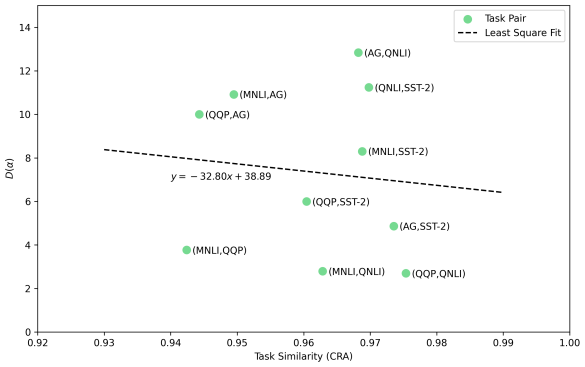


Figure 7: The average dissociation score and CRA similarity of each task pair in multi-task learning.

To sum up, we observe that there is a negative correlation between the average dissociation score and the task similarity, no matter which task similarity metric is adopted.

## E AG-Pair dataset

The AG-Pair dataset is built from the original dataset AG's News that contains 120k training samples from four topics. Given a pair of news as input, the model has to predict whether they are belonging to the same topic (Same) or not (Different).

To generate this dataset, samples in AG are iterated in random order and have an equal chance to combine a sample in the same topic or the other three topics. Thus the numbers of training samples in two classes are both 60k. Moreover, each news in AG's News occurs exactly twice in the AG-Pair dataset to keep the same word frequency.

## F Other Experimental Results on GLUE

In this section, we report more results and analyses of multi-task learning models on GLUE. Figure 8 illustrates that the average dissociation scores of five Transformer-based models are all improved by IAT as we expected.

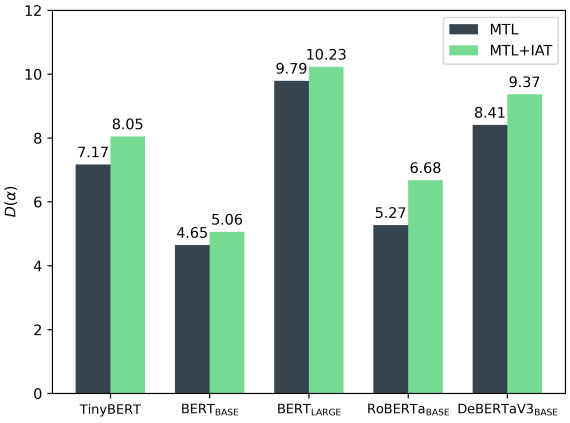


Figure 8: Average dissociation score of five multi-task learning models on GLUE dev set with $\alpha = 30\%$.

Figure 9 and 10 present the impact of two hyperparameters, $\delta$ and $\alpha$ in IAT, on the average performance of $\text{BERT}_{\text{BASE}}$. It is interesting to find that with a small $\delta$ and $\alpha$ (e.g., $\delta = 10\%$ and $\alpha = 30\%$), $\text{BERT}_{\text{BASE}}$ using IAT can achieve a good performance on GLUE dev set. Therefore, we only consider a limited hyperparameter sweep for each multi-task learning model with $\delta \in \{0.05, 0.1, 0.15\}$ and $\alpha \in \{0.1, 0.2, 0.3\}$.

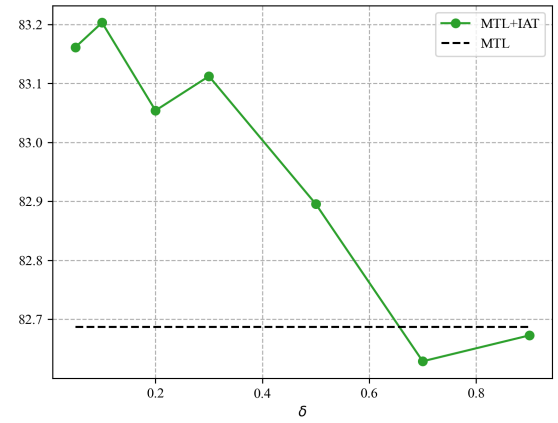


Figure 9: The average performance of $\text{BERT}_{\text{BASE}}$ on GLUE dev set using IAT with different $\delta$ ($\alpha = 50\%$).

Same as the finding in Stickland and Murray (2019), the annealed sampling method is better for multi-task learning of GLUE than the proportional sampling method. The sampling probabilities of task $i$ in annealed sampling are changed with epoch

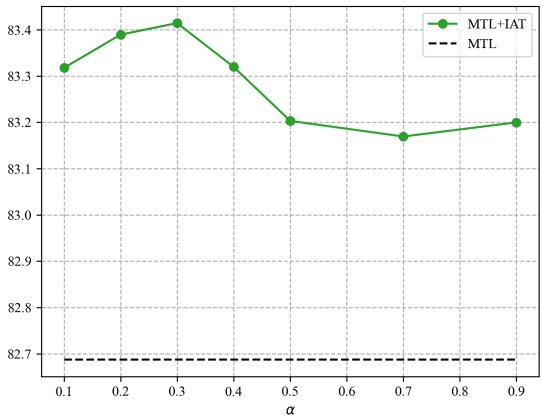


Figure 10: The average performance of $\text{BERT}_{\text{BASE}}$ on GLUE dev set using IAT with different $\alpha$ ($\delta = 10\%$).

$e$, and are calculated as follows:

$$p_i \propto N_i^{\varepsilon} \quad \text{with } \varepsilon = 1 - 0.8\frac{e-1}{E-1} \tag{8}$$

where $N_i$ is the number of samples in task $i$, $E$ is the total number of epochs. In contrast, the $\varepsilon$ in proportional sampling is always equal to 1.

Table 11 shows the results of five multi-task learning models using the proportional sampling method on GLUE test set. We can find that these multi-task learning models with proportional sampling perform better on GLUE test set after using IAT (+0.68% on average), which is in line with the findings in Section 6.3. It further demonstrates the effectiveness of our method.

Additional experimental results on development sets of GLUE for all models tested in this paper are reported in Table 12. In most cases, the standard deviation of average performance on GLUE development set is less than or equal to the baseline after using IAT, which indicates the robustness of our method.

| Model | CoLA Mcc | MNLI-(m/mm) Acc | MRPC F1 | QNLI Acc | QQP F1 | RTE Acc | SST-2 Acc | STS-B $r^s$ | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| TinyBERT | 27.0 | **82.8**/**82.5** | 83.4 | 90.3 | **70.4** | 71.6 | 92.3 | 83.4 | 76.0 |
| +IAT | **33.7** | **82.8**/82.2 | **85.3** | **90.6** | 70.3 | **71.9** | **92.5** | **84.0** | **77.0** |
| $\text{BERT}_{\text{BASE}}$ | 41.8 | 83.6/82.7 | 85.0 | **90.1** | 70.6 | 74.7 | **93.0** | 83.2 | 78.3 |
| +IAT | **45.1** | **84.0**/**83.3** | **85.1** | 89.6 | **70.8** | **76.2** | 92.8 | **83.5** | **78.9** |
| $\text{BERT}_{\text{LARGE}}$ | 53.3 | 85.4/**84.9** | **85.9** | 92.0 | 71.3 | 78.6 | 94.3 | 84.8 | 81.2 |
| +IAT | **56.7** | **85.8**/84.8 | **85.9** | **92.3** | **71.5** | **78.9** | **94.4** | **84.9** | **81.7** |
| $\text{RoBERTa}_{\text{BASE}}$ | 52.5 | **87.5**/**86.8** | 88.4 | 92.3 | 71.9 | 79.1 | 94.8 | **85.8** | 82.1 |
| +IAT | **56.1** | 87.1/86.7 | **88.5** | **92.6** | **72.3** | **79.9** | **95.2** | **85.8** | **82.7** |
| $\text{DeBERTa}_{\text{BASE}}$ | 61.9 | **89.9**/88.9 | 87.6 | 93.7 | 73.8 | 86.0 | 95.8 | 88.4 | 85.1 |
| +IAT | **64.8** | 89.7/**89.0** | **89.2** | **93.8** | **73.9** | **86.7** | **96.1** | **88.8** | **85.8** |

Table 11: GLUE test set results of five multi-task learning models using proportional sampling.

| Model | CoLA Mcc | MNLI-(m/mm) Acc | MRPC F1 | QNLI Acc | QQP F1 | RTE Acc | SST-2 Acc | STS-B $r^s$ | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| *Proportional Sampling* | | | | | | | | | |
| TinyBERT | 31.3 | **83.2**/**83.5** | 83.9 | **90.3** | 86.3 | 71.8 | 91.5 | 85.8 | $78.6_{\pm 0.5}$ |
| +IAT | **37.5** | 83.1/83.1 | **85.5** | 90.5 | **86.7** | **72.1** | **91.6** | **86.6** | $\mathbf{79.6}_{\pm 0.3}$ |
| $\text{BERT}_{\text{BASE}}$ | 47.2 | **83.7**/83.3 | 85.0 | **90.3** | 87.0 | 78.3 | **92.7** | 86.6 | $81.6_{\pm 0.3}$ |
| +IAT | **52.0** | 83.6/**83.6** | **86.5** | 90.1 | **87.2** | **79.4** | 91.9 | **87.2** | $\mathbf{82.4}_{\pm 0.2}$ |
| $\text{BERT}_{\text{LARGE}}$ | 54.1 | 85.7/85.6 | **86.9** | 91.6 | **88.3** | 82.1 | 93.2 | **87.9** | $83.9_{\pm 0.3}$ |
| +IAT | **58.9** | **86.0**/**85.7** | **86.9** | **91.8** | 88.2 | **82.7** | **93.4** | **87.9** | $\mathbf{84.6}_{\pm 0.3}$ |
| $\text{RoBERTa}_{\text{BASE}}$ | 49.7 | **87.6**/**87.1** | 89.7 | 91.9 | 87.6 | 83.0 | **94.5** | **88.4** | $84.4_{\pm 0.3}$ |
| +IAT | **54.2** | 87.2/**87.1** | **89.8** | **92.3** | **87.8** | **84.4** | 94.1 | **88.4** | $\mathbf{85.0}_{\pm 0.1}$ |
| $\text{DeBERTa}_{\text{BASE}}$ | 65.5 | **89.8**/**90.0** | 89.1 | 93.8 | 89.3 | 87.0 | 95.1 | 90.0 | $87.7_{\pm 0.2}$ |
| +IAT | **67.9** | 89.7/**90.0** | **90.1** | **93.9** | **89.4** | **87.6** | **95.5** | **90.1** | $\mathbf{88.2}_{\pm 0.1}$ |
| *Annealed Sampling* | | | | | | | | | |
| TinyBERT | 40.7 | **83.1**/82.9 | 85.0 | **90.4** | 86.2 | 73.6 | 90.6 | 87.5 | $80.0_{\pm 0.3}$ |
| +IAT | **45.8** | 82.8/**83.0** | **85.5** | 90.3 | **86.6** | **74.7** | **91.2** | **87.9** | $\mathbf{80.9}_{\pm 0.4}$ |
| $\text{BERT}_{\text{BASE}}$ | 51.1 | 83.6/83.9 | 87.6 | 90.1 | 87.1 | 79.8 | 92.2 | **88.4** | $82.6_{\pm 0.1}$ |
| +IAT | **53.5** | **83.8**/**84.0** | **89.0** | **90.6** | **87.6** | **80.6** | **93.2** | **88.4** | $\mathbf{83.4}_{\pm 0.2}$ |
| $\text{BERT}_{\text{LARGE}}$ | 58.8 | 85.8/85.8 | 87.4 | **91.8** | 87.9 | 82.0 | 92.8 | 88.6 | $84.5_{\pm 0.2}$ |
| +IAT | **61.6** | **86.0**/**86.0** | **88.7** | 91.5 | **88.0** | **82.2** | **93.7** | **89.1** | $\mathbf{85.2}_{\pm 0.2}$ |
| $\text{RoBERTa}_{\text{BASE}}$ | 54.3 | **87.3**/**87.0** | **92.2** | **92.3** | 86.9 | 84.6 | **94.5** | 89.0 | $85.3_{\pm 0.1}$ |
| +IAT | **59.2** | 87.1/86.8 | **92.2** | 92.2 | **87.1** | **84.7** | 94.3 | **89.1** | $\mathbf{85.9}_{\pm 0.1}$ |
| $\text{DeBERTa}_{\text{BASE}}$ | 65.7 | **90.0**/**90.1** | 90.7 | 93.9 | 89.0 | 88.0 | 95.2 | 90.3 | $88.1_{\pm 0.2}$ |
| +IAT | **68.7** | 89.9/90.0 | **91.4** | **94.0** | **89.1** | **88.5** | **95.3** | **90.7** | $\mathbf{88.6}_{\pm 0.1}$ |

Table 12: GLUE development set results of five multi-task learning models.