IDPG: An Instance-Dependent Prompt Generation Method

Zhuofeng Wu[1*] Sinong Wang[2] Jiatao Gu[2] Rui Hou[2]
Yuxiao Dong[2,3†] V.G. Vinod Vydiswaran[4,1] Hao Ma[2]

[1]School of Information, University of Michigan
[2]Meta AI [3]Tsinghua University
[4]Department of Learning Health Sciences, University of Michigan

NAACL 2022

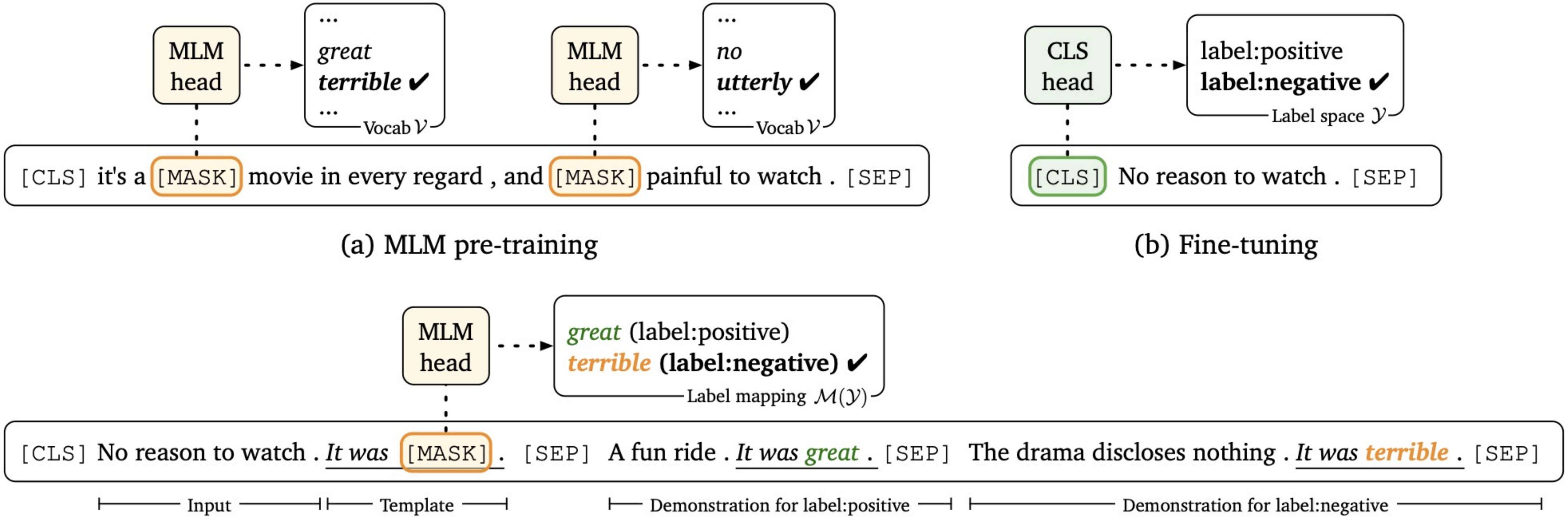
[*] Work partially done while interning at Meta AI. [†] Work done when at Meta AI.

Motivation: fine-tuning has become expansive

- Pre-training a large model and fine-tuning it on different downstream tasks has become the primary transfer learning paradigm.

- However, as the model size proliferates, every time a new task comes, updating and storing the whole model becomes expansive.

Previous Approaches: manual prompt learning



(a) MLM pre-training

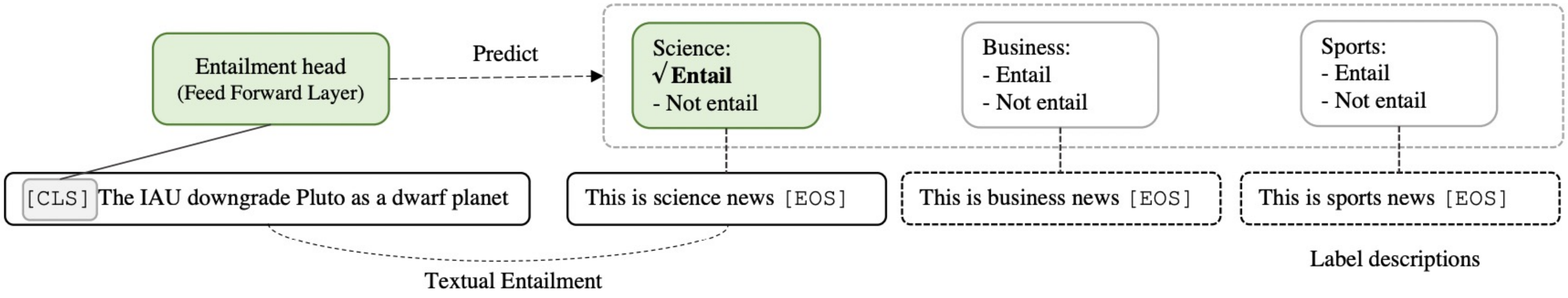
(b) Fine-tuning

(c) Prompt-based fine-tuning with demonstrations (our approach)

LM-BFF appends task-description words “It was [MASK]” to the end of input and predicts the masked tokens.

* Figure from (Gao et al., 2021a).

Previous Approaches: manual prompt learning



EFL reformulates tasks as entailment task to leverage rich NLI supervision data

* Figure from (Wang et al., 2021).

Previous Approaches: drawbacks of manual prompt

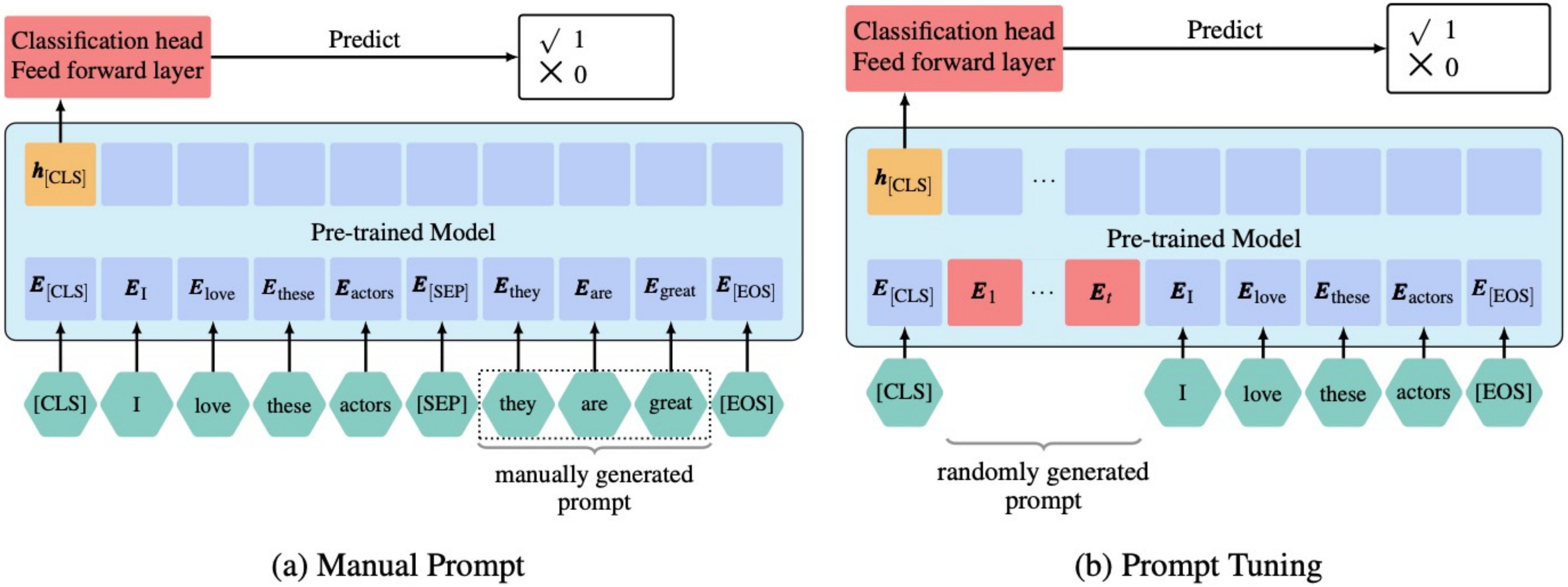
- Manual prompt learning methods request either:
  - manual design -> hard to generalize
  - beam search -> time consuming & limited to few-shot setting

* Figure from (Wang et al., 2021).

Previous Approaches: prompt tuning

- To learn the optimized prompt for each task.
- Prepend **trainable** tokens to input.
- Learn embeddings of these virtual tokens via backpropagation.
- Keep rest of model **fixed**.

Previous Approaches:



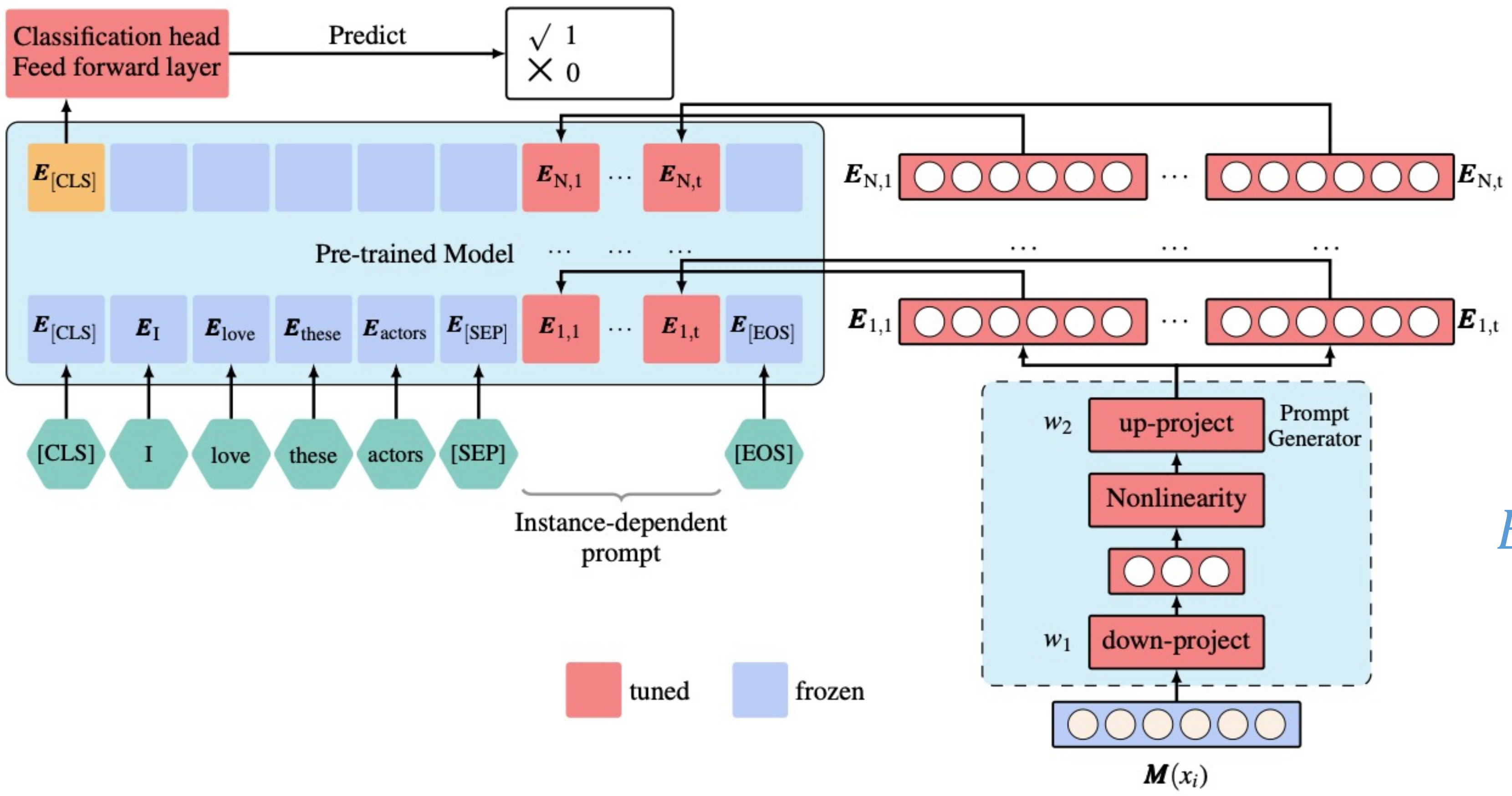
Research gaps

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

- Assuming we already get optimized prompt $E_1^*, \ldots, E_t^*$, every time a new input x_i comes, we prepend $E_1^*, \ldots, E_t^*$ in front of it.
- Indeed, it is unlikely to see many different sentences with the same prefix in the pre-training corpus.
- Thus, a unified prompt may disturb the prediction.

Can we generate input-dependent prompts to smooth the domain difference?

Our method: Instance-Dependent Prompt Generation



$$E_{\{i,j\}} = Wx + b$$

IDPG degenerates into traditional prompt tuning if setting prompt generator as a zero matrix.

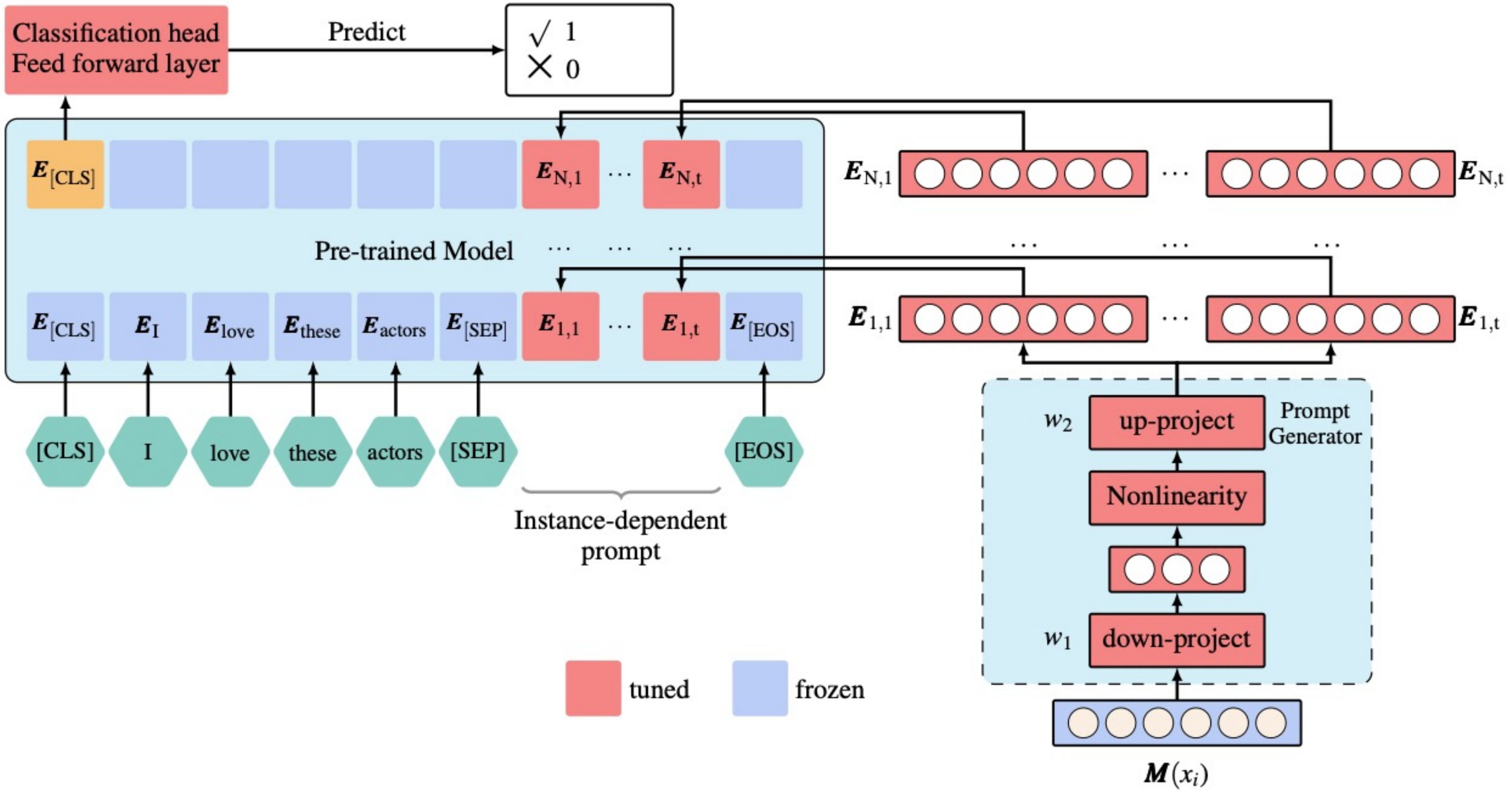
Optimization #1: Parameterized Hypercomplex Multiplication (PHM) Layers

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

- Prompt tuning only prepends several tokens and trains them.
- The bottleneck structure takes millions of parameters.

Optimization #1: Parameterized Hypercomplex Multiplication (PHM) Layers

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN



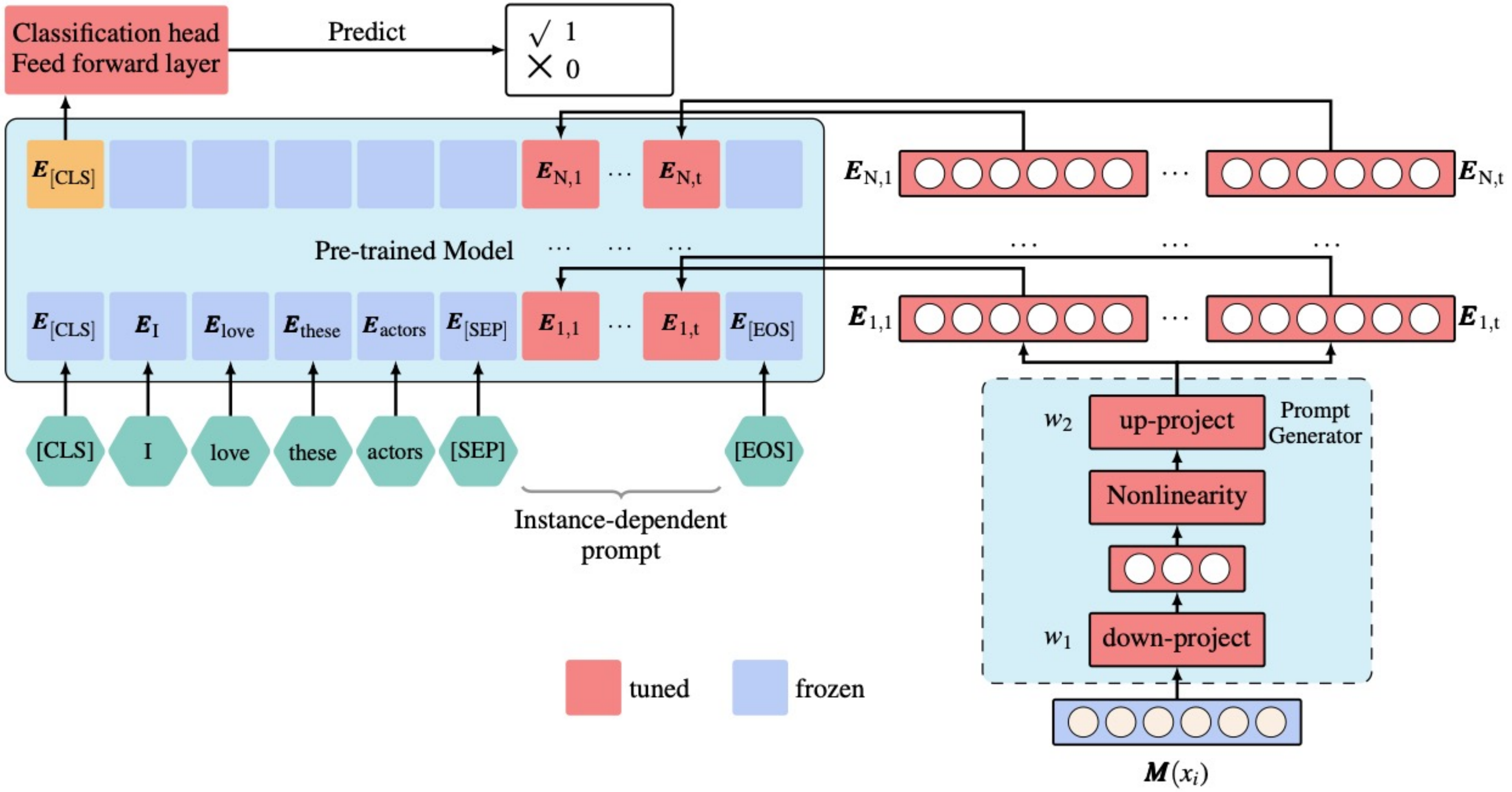
IDPG-DNN:

$$W_1 \in R^{\{m\times d\}}, W_2 \in R^{\{t\times m\times d\}}$$

$$d = 1024$$

$$m \in [16, 32, 64, 128, 256]$$

Optimization #1: Parameterized Hypercomplex Multiplication (PHM) Layers



IDPG-PHM:

$$W = \sum_{\{i=1\}}^{n} A_i \otimes B_i$$

$$A_i \in R^{\{n\times n\}}, B_i \in R^{\left\{\frac{m}{n}\times\frac{d}{n}\right\}}$$

$\otimes$: Kronecker product

param size: $\boldsymbol{n^3 + \frac{m\times d}{n}}$

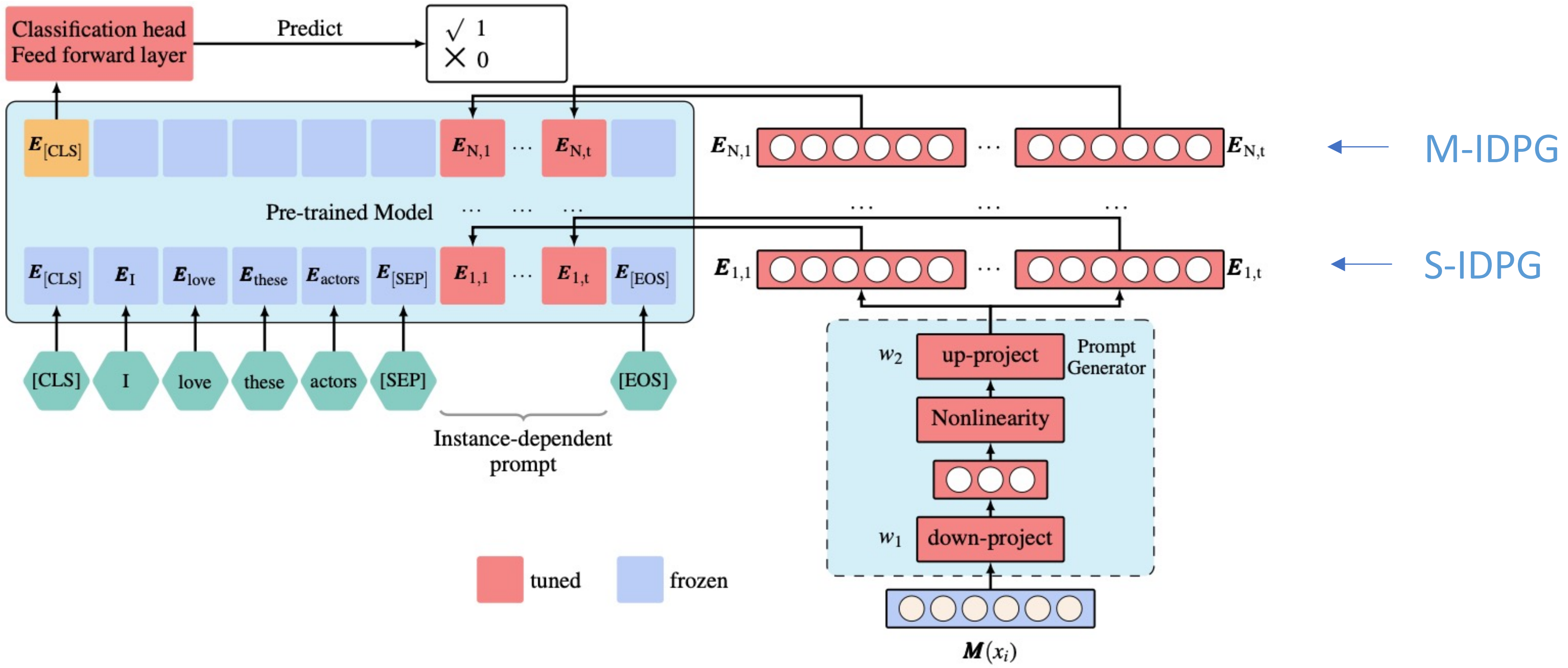
$$\cancel{W_1 \in R^{\{m\times d\}}, W_2 \in R^{\{t\times m\times d\}}}$$

$$\cancel{d = 1024}$$

$$\cancel{m \in [16, 32, 64, 128, 256]}$$

Optimization #2: Multilayer prompts

- We insert the prompt into multi-layer in transformer.



Main results – full data setting

Single sentence tasks   sentence pair tasks

Method	MPQA	Subj	CR	MR	SST-2	QNLI	RTE	MRPC	STS-B	QQP	Avg
Transformer Fine-tuning											
RoBERTa	90.4±0.2	97.1±0.1	90.7±0.7	91.7±0.2	96.4±0.2	**94.7**±0.1	85.7±0.2	91.8±0.4	92.2±0.2	**91.0**±0.1	92.2
EFL	90.3±0.2	97.2±0.1	93.0±0.7	91.7±0.2	**96.5**±0.1	94.4±0.1	85.6±2.4	91.2±0.4	**92.5**±0.1	**91.0**±0.2	92.3
Adapter											
Compacter	91.1±0.2	97.5±0.1	92.7±0.4	**92.6**±0.2	96.0±0.2	94.3±0.2	87.1±1.4	91.6±0.6	91.6±0.1	87.1±0.2	92.2
Adapter	90.8±0.2	97.5±0.1	92.8±0.3	92.5±0.1	96.1±0.1	94.8±0.2	88.1±0.4	91.8±0.6	92.1±0.1	89.9±0.1	**92.6**
Prompting											
Prompt-tuning	90.3±0.2	95.5±0.4	91.2±1.1	91.0±0.2	94.2±0.3	86.0±0.3	87.0±0.4	84.3±0.3	87.2±0.2	81.6±0.1	88.8
Prompt-tuning-134	65.7±19	95.6±0.2	86.7±3.6	89.7±0.5	92.0±0.5	83.0±1.1	87.4±0.5	84.1±0.5	87.6±0.5	82.4±0.3	85.4
Ptuningv2	90.4±0.3	96.5±0.3	92.7±0.3	91.6±0.1	94.4±0.2	92.9±0.1	78.4±4.3	91.4±0.4	89.9±0.2	84.4±0.4	90.3
S-IDPG-PHM	89.6±0.3	94.4±0.3	90.3±0.2	89.3±0.4	94.7±0.2	90.7±0.3	89.2±0.2	84.3±0.8	84.7±0.9	82.5±0.2	89.0
S-IDPG-DNN	89.5±0.7	94.9±0.4	89.9±1.5	90.2±0.6	95.1±0.2	90.5±0.5	**89.4**±0.4	83.0±0.5	85.3±0.7	82.7±0.3	89.1
M-IDPG-PHM-GloVe	90.9±0.2	97.4±0.1	93.3±0.1	**92.6**±0.3	95.4±0.2	94.4±0.2	82.1±0.6	92.1±0.4	91.0±0.4	86.3±0.2	91.6
M-IDPG-PHM	**91.2**±0.2	97.5±0.1	93.2±0.3	**92.6**±0.3	96.0±0.3	94.5±0.1	83.5±0.7	**92.3**±0.2	91.4±0.4	86.2±0.1	91.9
M-IDPG-DNN	**91.2**±0.3	**97.6**±0.2	**93.5**±0.3	**92.6**±0.1	95.9±0.1	94.5±0.2	85.5±0.6	91.8±0.3	91.5±0.2	86.9±0.3	92.1

RoBERTa-large as backbone for all competing methods

Main results – IDPG vs traditional prompting methods

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Single sentence tasks: MPQA, Subj, CR, MR, SST-2. Sentence pair tasks: QNLI, RTE, MRPC, STS-B, QQP.

Method	MPQA	Subj	CR	MR	SST-2	QNLI	RTE	MRPC	STS-B	QQP	Avg	#params
Transformer Fine-tuning												
RoBERTa	$90.4_{\pm0.2}$	$97.1_{\pm0.1}$	$90.7_{\pm0.7}$	$91.7_{\pm0.2}$	$96.4_{\pm0.2}$	$\mathbf{94.7}_{\pm0.1}$	$85.7_{\pm0.2}$	$91.8_{\pm0.4}$	$92.2_{\pm0.2}$	$\mathbf{91.0}_{\pm0.1}$	92.2	
EFL	$90.3_{\pm0.2}$	$97.2_{\pm0.1}$	$93.0_{\pm0.7}$	$91.7_{\pm0.2}$	$\mathbf{96.5}_{\pm0.1}$	$94.4_{\pm0.1}$	$85.6_{\pm2.4}$	$91.2_{\pm0.4}$	$\mathbf{92.5}_{\pm0.1}$	$\mathbf{91.0}_{\pm0.2}$	92.3	
Adapter												
Compacter	$91.1_{\pm0.2}$	$97.5_{\pm0.1}$	$92.7_{\pm0.4}$	$\mathbf{92.6}_{\pm0.2}$	$96.0_{\pm0.2}$	$94.3_{\pm0.2}$	$87.1_{\pm1.4}$	$91.6_{\pm0.6}$	$91.6_{\pm0.1}$	$87.1_{\pm0.2}$	92.2	
Adapter	$90.8_{\pm0.2}$	$97.5_{\pm0.1}$	$92.8_{\pm0.3}$	$92.5_{\pm0.1}$	$96.1_{\pm0.1}$	$94.8_{\pm0.2}$	$88.1_{\pm0.4}$	$91.8_{\pm0.6}$	$92.1_{\pm0.1}$	$89.9_{\pm0.1}$	**92.6**	
Prompting												
Prompt-tuning	$90.3_{\pm0.2}$	$95.5_{\pm0.4}$	$91.2_{\pm1.1}$	$91.0_{\pm0.2}$	$94.2_{\pm0.3}$	$86.0_{\pm0.3}$	$87.0_{\pm0.4}$	$84.3_{\pm0.3}$	$87.2_{\pm0.2}$	$81.6_{\pm0.1}$	88.8	5K
Prompt-tuning-134	$65.7_{\pm19}$	$95.6_{\pm0.2}$	$86.7_{\pm3.6}$	$89.7_{\pm0.5}$	$92.0_{\pm0.5}$	$83.0_{\pm1.1}$	$87.4_{\pm0.5}$	$84.1_{\pm0.5}$	$87.6_{\pm0.5}$	$82.4_{\pm0.3}$	85.4	134K
Ptuningv2	$90.4_{\pm0.3}$	$96.5_{\pm0.3}$	$92.7_{\pm0.3}$	$91.6_{\pm0.1}$	$94.4_{\pm0.2}$	$92.9_{\pm0.1}$	$78.4_{\pm4.3}$	$91.4_{\pm0.4}$	$89.9_{\pm0.2}$	$84.4_{\pm0.4}$	90.3	120K
S-IDPG-PHM	$89.6_{\pm0.3}$	$94.4_{\pm0.3}$	$90.3_{\pm0.2}$	$89.3_{\pm0.4}$	$94.7_{\pm0.2}$	$90.7_{\pm0.3}$	$89.2_{\pm0.2}$	$84.3_{\pm0.8}$	$84.7_{\pm0.9}$	$82.5_{\pm0.2}$	89.0	
S-IDPG-DNN	$89.5_{\pm0.7}$	$94.9_{\pm0.4}$	$89.9_{\pm1.5}$	$90.2_{\pm0.6}$	$95.1_{\pm0.2}$	$90.5_{\pm0.5}$	$\mathbf{89.4}_{\pm0.4}$	$83.0_{\pm0.5}$	$85.3_{\pm0.7}$	$82.7_{\pm0.3}$	89.1	
M-IDPG-PHM-GloVe	$90.9_{\pm0.2}$	$97.4_{\pm0.1}$	$93.3_{\pm0.1}$	$\mathbf{92.6}_{\pm0.3}$	$95.4_{\pm0.2}$	$94.4_{\pm0.2}$	$82.1_{\pm0.6}$	$92.1_{\pm0.4}$	$91.0_{\pm0.4}$	$86.3_{\pm0.2}$	91.6	
M-IDPG-PHM	$\mathbf{91.2}_{\pm0.2}$	$97.5_{\pm0.1}$	$93.2_{\pm0.3}$	$\mathbf{92.6}_{\pm0.3}$	$96.0_{\pm0.3}$	$94.5_{\pm0.1}$	$83.5_{\pm0.7}$	$\mathbf{92.3}_{\pm0.2}$	$91.4_{\pm0.4}$	$86.2_{\pm0.1}$	91.9	134K
M-IDPG-DNN	$\mathbf{91.2}_{\pm0.3}$	$\mathbf{97.6}_{\pm0.2}$	$\mathbf{93.5}_{\pm0.3}$	$\mathbf{92.6}_{\pm0.1}$	$95.9_{\pm0.1}$	$94.5_{\pm0.2}$	$85.5_{\pm0.6}$	$91.8_{\pm0.3}$	$91.5_{\pm0.2}$	$86.9_{\pm0.3}$	92.1	

M-IDPG-PHM consistently outperforms task-specific prompt tuning methods by 1.6–3.1 points

Main results – IDPG vs fully fine-tuning

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

	Single sentence tasks					sentence pair tasks						
Method	**MPQA**	**Subj**	**CR**	**MR**	**SST-2**	**QNLI**	**RTE**	**MRPC**	**STS-B**	**QQP**	Avg	#params
Transformer Fine-tuning												
RoBERTa	90.4$_{\pm0.2}$	97.1$_{\pm0.1}$	90.7$_{\pm0.7}$	91.7$_{\pm0.2}$	96.4$_{\pm0.2}$	**94.7**$_{\pm0.1}$	85.7$_{\pm0.2}$	91.8$_{\pm0.4}$	92.2$_{\pm0.2}$	**91.0**$_{\pm0.1}$	92.2	355M
EFL	90.3$_{\pm0.2}$	97.2$_{\pm0.1}$	93.0$_{\pm0.7}$	91.7$_{\pm0.2}$	**96.5**$_{\pm0.1}$	94.4$_{\pm0.1}$	85.6$_{\pm2.4}$	91.2$_{\pm0.4}$	**92.5**$_{\pm0.1}$	**91.0**$_{\pm0.2}$	92.3	
Adapter												
Compacter	91.1$_{\pm0.2}$	97.5$_{\pm0.1}$	92.7$_{\pm0.4}$	**92.6**$_{\pm0.2}$	96.0$_{\pm0.2}$	94.3$_{\pm0.2}$	87.1$_{\pm1.4}$	91.6$_{\pm0.6}$	91.6$_{\pm0.1}$	87.1$_{\pm0.2}$	92.2	
Adapter	90.8$_{\pm0.2}$	97.5$_{\pm0.1}$	92.8$_{\pm0.3}$	92.5$_{\pm0.1}$	96.1$_{\pm0.1}$	94.8$_{\pm0.2}$	88.1$_{\pm0.4}$	91.8$_{\pm0.6}$	92.1$_{\pm0.1}$	89.9$_{\pm0.1}$	**92.6**	
Prompting												
Prompt-tuning	90.3$_{\pm0.2}$	95.5$_{\pm0.4}$	91.2$_{\pm1.1}$	91.0$_{\pm0.2}$	94.2$_{\pm0.3}$	86.0$_{\pm0.3}$	87.0$_{\pm0.4}$	84.3$_{\pm0.3}$	87.2$_{\pm0.2}$	81.6$_{\pm0.1}$	88.8	
Prompt-tuning-134	65.7$_{\pm19}$	95.6$_{\pm0.2}$	86.7$_{\pm3.6}$	89.7$_{\pm0.5}$	92.0$_{\pm0.5}$	83.0$_{\pm1.1}$	87.4$_{\pm0.5}$	84.1$_{\pm0.5}$	87.6$_{\pm0.5}$	82.4$_{\pm0.3}$	85.4	
Ptuningv2	90.4$_{\pm0.3}$	96.5$_{\pm0.3}$	92.7$_{\pm0.3}$	91.6$_{\pm0.1}$	94.4$_{\pm0.2}$	92.9$_{\pm0.1}$	78.4$_{\pm4.3}$	91.4$_{\pm0.4}$	89.9$_{\pm0.2}$	84.4$_{\pm0.4}$	90.3	
S-IDPG-PHM	89.6$_{\pm0.3}$	94.4$_{\pm0.3}$	90.3$_{\pm0.2}$	89.3$_{\pm0.4}$	94.7$_{\pm0.2}$	90.7$_{\pm0.3}$	89.2$_{\pm0.2}$	84.3$_{\pm0.8}$	84.7$_{\pm0.9}$	82.5$_{\pm0.2}$	89.0	
S-IDPG-DNN	89.5$_{\pm0.7}$	94.9$_{\pm0.4}$	89.9$_{\pm1.5}$	90.2$_{\pm0.6}$	95.1$_{\pm0.2}$	90.5$_{\pm0.5}$	**89.4**$_{\pm0.4}$	83.0$_{\pm0.5}$	85.3$_{\pm0.7}$	82.7$_{\pm0.3}$	89.1	
M-IDPG-PHM-GloVe	90.9$_{\pm0.2}$	97.4$_{\pm0.1}$	93.3$_{\pm0.1}$	**92.6**$_{\pm0.3}$	95.4$_{\pm0.2}$	94.4$_{\pm0.2}$	82.1$_{\pm0.6}$	92.1$_{\pm0.4}$	91.0$_{\pm0.4}$	86.3$_{\pm0.2}$	91.6	
M-IDPG-PHM	**91.2**$_{\pm0.2}$	97.5$_{\pm0.1}$	93.2$_{\pm0.3}$	**92.6**$_{\pm0.3}$	96.0$_{\pm0.3}$	94.5$_{\pm0.1}$	83.5$_{\pm0.7}$	**92.3**$_{\pm0.2}$	91.4$_{\pm0.4}$	86.2$_{\pm0.1}$	91.9	134K, ~0.037%
M-IDPG-DNN	**91.2**$_{\pm0.3}$	**97.6**$_{\pm0.2}$	**93.5**$_{\pm0.3}$	**92.6**$_{\pm0.1}$	95.9$_{\pm0.1}$	94.5$_{\pm0.2}$	85.5$_{\pm0.6}$	91.8$_{\pm0.3}$	91.5$_{\pm0.2}$	86.9$_{\pm0.3}$	92.1	

Performance: 91.9 vs 92.2
#Params: 134K vs 355M

Main results – IDPG vs Adapter

Single sentence tasks: MPQA, Subj, CR, MR, SST-2 — sentence pair tasks: QNLI, RTE, MRPC, STS-B, QQP

Method	MPQA	Subj	CR	MR	SST-2	QNLI	RTE	MRPC	STS-B	QQP	Avg	#params
Transformer Fine-tuning												
RoBERTa	90.4$_{\pm 0.2}$	97.1$_{\pm 0.1}$	90.7$_{\pm 0.7}$	91.7$_{\pm 0.2}$	96.4$_{\pm 0.2}$	**94.7**$_{\pm 0.1}$	85.7$_{\pm 0.2}$	91.8$_{\pm 0.4}$	92.2$_{\pm 0.2}$	**91.0**$_{\pm 0.1}$	92.2	
EFL	90.3$_{\pm 0.2}$	97.2$_{\pm 0.1}$	93.0$_{\pm 0.7}$	91.7$_{\pm 0.2}$	**96.5**$_{\pm 0.1}$	94.4$_{\pm 0.1}$	85.6$_{\pm 2.4}$	91.2$_{\pm 0.4}$	**92.5**$_{\pm 0.1}$	**91.0**$_{\pm 0.2}$	92.3	
Adapter												
Compacter	91.1$_{\pm 0.2}$	97.5$_{\pm 0.1}$	92.7$_{\pm 0.4}$	**92.6**$_{\pm 0.2}$	96.0$_{\pm 0.2}$	94.3$_{\pm 0.2}$	87.1$_{\pm 1.4}$	91.6$_{\pm 0.6}$	91.6$_{\pm 0.1}$	87.1$_{\pm 0.2}$	92.2	149K
Adapter	90.8$_{\pm 0.2}$	97.5$_{\pm 0.1}$	92.8$_{\pm 0.3}$	92.5$_{\pm 0.1}$	96.1$_{\pm 0.1}$	94.8$_{\pm 0.2}$	88.1$_{\pm 0.4}$	91.8$_{\pm 0.6}$	92.1$_{\pm 0.1}$	89.9$_{\pm 0.1}$	**92.6**	1.55M
Prompting												
Prompt-tuning	90.3$_{\pm 0.2}$	95.5$_{\pm 0.4}$	91.2$_{\pm 1.1}$	91.0$_{\pm 0.2}$	94.2$_{\pm 0.3}$	86.0$_{\pm 0.3}$	87.0$_{\pm 0.4}$	84.3$_{\pm 0.3}$	87.2$_{\pm 0.2}$	81.6$_{\pm 0.1}$	88.8	
Prompt-tuning-134	65.7$_{\pm 19}$	95.6$_{\pm 0.2}$	86.7$_{\pm 3.6}$	89.7$_{\pm 0.5}$	92.0$_{\pm 0.5}$	83.0$_{\pm 1.1}$	87.4$_{\pm 0.5}$	84.1$_{\pm 0.5}$	87.6$_{\pm 0.5}$	82.4$_{\pm 0.3}$	85.4	
Ptuningv2	90.4$_{\pm 0.3}$	96.5$_{\pm 0.3}$	92.7$_{\pm 0.3}$	91.6$_{\pm 0.1}$	94.4$_{\pm 0.2}$	92.9$_{\pm 0.1}$	78.4$_{\pm 4.3}$	91.4$_{\pm 0.4}$	89.9$_{\pm 0.2}$	84.4$_{\pm 0.4}$	90.3	
S-IDPG-PHM	89.6$_{\pm 0.3}$	94.4$_{\pm 0.3}$	90.3$_{\pm 0.2}$	89.3$_{\pm 0.4}$	94.7$_{\pm 0.2}$	90.7$_{\pm 0.3}$	89.2$_{\pm 0.2}$	84.3$_{\pm 0.8}$	84.7$_{\pm 0.9}$	82.5$_{\pm 0.2}$	89.0	
S-IDPG-DNN	89.5$_{\pm 0.7}$	94.9$_{\pm 0.4}$	89.9$_{\pm 1.5}$	90.2$_{\pm 0.6}$	95.1$_{\pm 0.2}$	90.5$_{\pm 0.5}$	**89.4**$_{\pm 0.4}$	83.0$_{\pm 0.5}$	85.3$_{\pm 0.7}$	82.7$_{\pm 0.3}$	89.1	
M-IDPG-PHM-GloVe	90.9$_{\pm 0.2}$	97.4$_{\pm 0.1}$	93.3$_{\pm 0.1}$	**92.6**$_{\pm 0.3}$	95.4$_{\pm 0.2}$	94.4$_{\pm 0.2}$	82.1$_{\pm 0.6}$	92.1$_{\pm 0.4}$	91.0$_{\pm 0.4}$	86.3$_{\pm 0.2}$	91.6	
M-IDPG-PHM	**91.2**$_{\pm 0.2}$	97.5$_{\pm 0.1}$	93.2$_{\pm 0.3}$	**92.6**$_{\pm 0.3}$	96.0$_{\pm 0.3}$	94.5$_{\pm 0.1}$	83.5$_{\pm 0.7}$	**92.3**$_{\pm 0.2}$	91.4$_{\pm 0.4}$	86.2$_{\pm 0.1}$	91.9	134K
M-IDPG-DNN	**91.2**$_{\pm 0.3}$	**97.6**$_{\pm 0.2}$	**93.5**$_{\pm 0.3}$	**92.6**$_{\pm 0.1}$	95.9$_{\pm 0.1}$	94.5$_{\pm 0.2}$	85.5$_{\pm 0.6}$	91.8$_{\pm 0.3}$	91.5$_{\pm 0.2}$	86.9$_{\pm 0.3}$	92.1	

Performance: 91.9 vs 92.2
#Params: 134K vs 149K

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Observation – Multi-layer insertion helps

	Single sentence tasks					sentence pair tasks						
Method	**MPQA**	**Subj**	**CR**	**MR**	**SST-2**	**QNLI**	**RTE**	**MRPC**	**STS-B**	**QQP**	Avg	#params
Transformer Fine-tuning												
RoBERTa	90.4$_{\pm0.2}$	97.1$_{\pm0.1}$	90.7$_{\pm0.7}$	91.7$_{\pm0.2}$	96.4$_{\pm0.2}$	**94.7**$_{\pm0.1}$	85.7$_{\pm0.2}$	91.8$_{\pm0.4}$	92.2$_{\pm0.2}$	**91.0**$_{\pm0.1}$	92.2	
EFL	90.3$_{\pm0.2}$	97.2$_{\pm0.1}$	93.0$_{\pm0.7}$	91.7$_{\pm0.2}$	**96.5**$_{\pm0.1}$	94.4$_{\pm0.1}$	85.6$_{\pm2.4}$	91.2$_{\pm0.4}$	**92.5**$_{\pm0.1}$	**91.0**$_{\pm0.2}$	92.3	
Adapter												
Compacter	91.1$_{\pm0.2}$	97.5$_{\pm0.1}$	92.7$_{\pm0.4}$	**92.6**$_{\pm0.2}$	96.0$_{\pm0.2}$	94.3$_{\pm0.2}$	87.1$_{\pm1.4}$	91.6$_{\pm0.6}$	91.6$_{\pm0.1}$	87.1$_{\pm0.2}$	92.2	
Adapter	90.8$_{\pm0.2}$	97.5$_{\pm0.1}$	92.8$_{\pm0.3}$	92.5$_{\pm0.1}$	96.1$_{\pm0.1}$	94.8$_{\pm0.2}$	88.1$_{\pm0.4}$	91.8$_{\pm0.6}$	92.1$_{\pm0.1}$	89.9$_{\pm0.1}$	**92.6**	
Prompting												
Prompt-tuning	90.3$_{\pm0.2}$	95.5$_{\pm0.4}$	91.2$_{\pm1.1}$	91.0$_{\pm0.2}$	94.2$_{\pm0.3}$	86.0$_{\pm0.3}$	87.0$_{\pm0.4}$	84.3$_{\pm0.3}$	87.2$_{\pm0.2}$	81.6$_{\pm0.1}$	88.8	
Prompt-tuning-134	65.7$_{\pm19}$	95.6$_{\pm0.2}$	86.7$_{\pm3.6}$	89.7$_{\pm0.5}$	92.0$_{\pm0.5}$	83.0$_{\pm1.1}$	87.4$_{\pm0.5}$	84.1$_{\pm0.5}$	87.6$_{\pm0.5}$	82.4$_{\pm0.3}$	85.4	
Ptuningv2	90.4$_{\pm0.3}$	96.5$_{\pm0.3}$	92.7$_{\pm0.3}$	91.6$_{\pm0.1}$	94.4$_{\pm0.2}$	92.9$_{\pm0.1}$	78.4$_{\pm4.3}$	91.4$_{\pm0.4}$	89.9$_{\pm0.2}$	84.4$_{\pm0.4}$	90.3	
S-IDPG-PHM	89.6$_{\pm0.3}$	94.4$_{\pm0.3}$	90.3$_{\pm0.2}$	89.3$_{\pm0.4}$	94.7$_{\pm0.2}$	90.7$_{\pm0.3}$	89.2$_{\pm0.2}$	84.3$_{\pm0.8}$	84.7$_{\pm0.9}$	82.5$_{\pm0.2}$	89.0	105K
S-IDPG-DNN	89.5$_{\pm0.7}$	94.9$_{\pm0.4}$	89.9$_{\pm1.5}$	90.2$_{\pm0.6}$	95.1$_{\pm0.2}$	90.5$_{\pm0.5}$	**89.4**$_{\pm0.4}$	83.0$_{\pm0.5}$	85.3$_{\pm0.7}$	82.7$_{\pm0.3}$	89.1	
M-IDPG-PHM-GloVe	90.9$_{\pm0.2}$	97.4$_{\pm0.1}$	93.3$_{\pm0.1}$	**92.6**$_{\pm0.3}$	95.4$_{\pm0.2}$	94.4$_{\pm0.2}$	82.1$_{\pm0.6}$	92.1$_{\pm0.4}$	91.0$_{\pm0.4}$	86.3$_{\pm0.2}$	91.6	
M-IDPG-PHM	**91.2**$_{\pm0.2}$	97.5$_{\pm0.1}$	93.2$_{\pm0.3}$	**92.6**$_{\pm0.3}$	96.0$_{\pm0.3}$	94.5$_{\pm0.1}$	83.5$_{\pm0.7}$	**92.3**$_{\pm0.2}$	91.4$_{\pm0.4}$	86.2$_{\pm0.1}$	91.9	134K
M-IDPG-DNN	**91.2**$_{\pm0.3}$	**97.6**$_{\pm0.2}$	**93.5**$_{\pm0.3}$	**92.6**$_{\pm0.1}$	95.9$_{\pm0.1}$	94.5$_{\pm0.2}$	85.5$_{\pm0.6}$	91.8$_{\pm0.3}$	91.5$_{\pm0.2}$	86.9$_{\pm0.3}$	92.1	

M-IDPG-PHM vs S-IDPG-PHM: +2.9 points while +29K params

Observation – PHM performs slightly worse while saving params

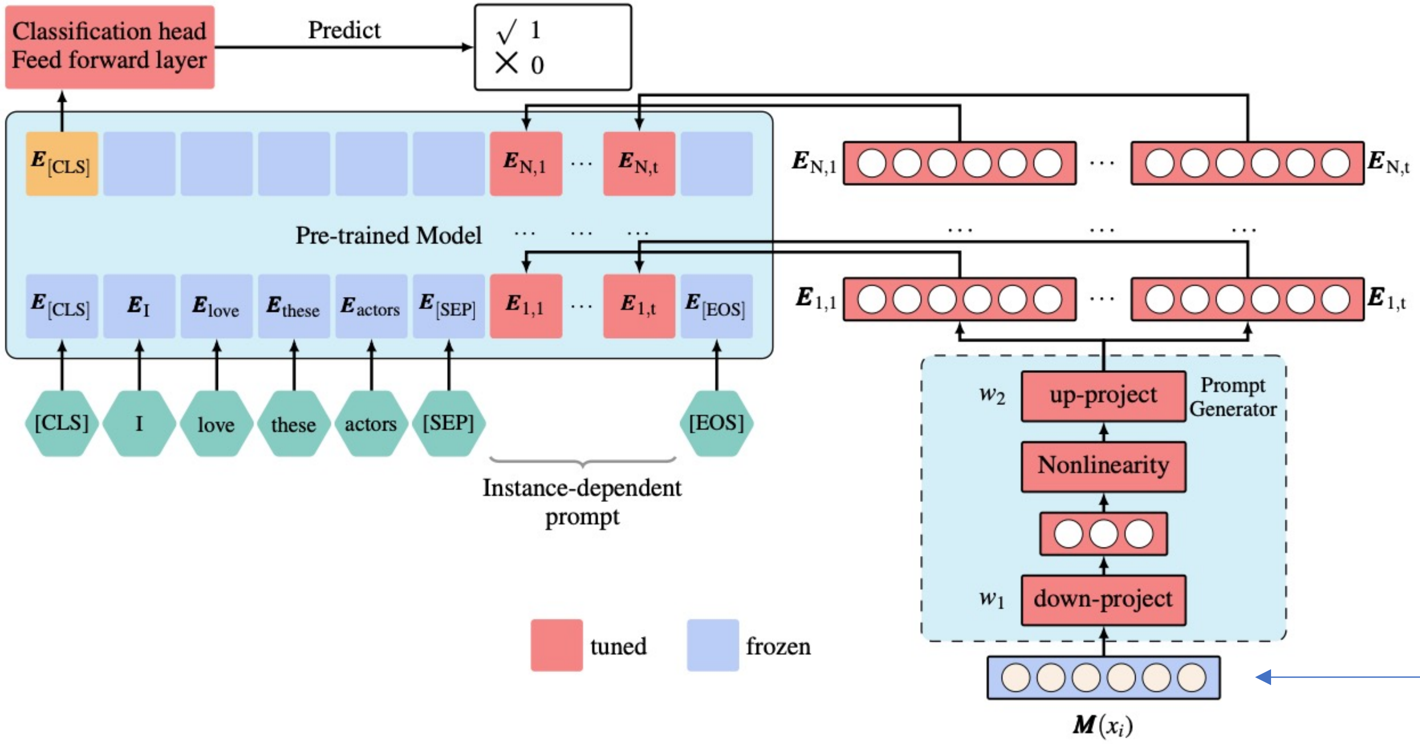
Single sentence tasks: MPQA, Subj, CR, MR, SST-2; sentence pair tasks: QNLI, RTE, MRPC, STS-B, QQP

Method	MPQA	Subj	CR	MR	SST-2	QNLI	RTE	MRPC	STS-B	QQP	Avg	#params
Transformer Fine-tuning												
RoBERTa	90.4±0.2	97.1±0.1	90.7±0.7	91.7±0.2	96.4±0.2	**94.7**±0.1	85.7±0.2	91.8±0.4	92.2±0.2	**91.0**±0.1	92.2	
EFL	90.3±0.2	97.2±0.1	93.0±0.7	91.7±0.2	**96.5**±0.1	94.4±0.1	85.6±2.4	91.2±0.4	**92.5**±0.1	**91.0**±0.2	92.3	
Adapter												
Compacter	91.1±0.2	97.5±0.1	92.7±0.4	**92.6**±0.2	96.0±0.2	94.3±0.2	87.1±1.4	91.6±0.6	91.6±0.1	87.1±0.2	92.2	
Adapter	90.8±0.2	97.5±0.1	92.8±0.3	92.5±0.1	96.1±0.1	94.8±0.2	88.1±0.4	91.8±0.6	92.1±0.1	89.9±0.1	**92.6**	
Prompting												
Prompt-tuning	90.3±0.2	95.5±0.4	91.2±1.1	91.0±0.2	94.2±0.3	86.0±0.3	87.0±0.4	84.3±0.3	87.2±0.2	81.6±0.1	88.8	
Prompt-tuning-134	65.7±19	95.6±0.2	86.7±3.6	89.7±0.5	92.0±0.5	83.0±1.1	87.4±0.5	84.1±0.5	87.6±0.5	82.4±0.3	85.4	
Ptuningv2	90.4±0.3	96.5±0.3	92.7±0.3	91.6±0.1	94.4±0.2	92.9±0.1	78.4±4.3	91.4±0.4	89.9±0.2	84.4±0.4	90.3	
S-IDPG-PHM	89.6±0.3	94.4±0.3	90.3±0.2	89.3±0.4	94.7±0.2	90.7±0.3	89.2±0.2	84.3±0.8	84.7±0.9	82.5±0.2	89.0	
S-IDPG-DNN	89.5±0.7	94.9±0.4	89.9±1.5	90.2±0.6	95.1±0.2	90.5±0.5	**89.4**±0.4	83.0±0.5	85.3±0.7	82.7±0.3	89.1	
M-IDPG-PHM-GloVe	90.9±0.2	97.4±0.1	93.3±0.1	**92.6**±0.3	95.4±0.2	94.4±0.2	82.1±0.6	92.1±0.4	91.0±0.4	86.3±0.2	91.6	
M-IDPG-PHM	$\boxed{\textbf{91.2}\pm0.2}$	97.5±0.1	93.2±0.3	$\boxed{\textbf{92.6}\pm0.3}$	$\boxed{96.0\pm0.3}$	$\boxed{94.5\pm0.1}$	83.5±0.7	$\boxed{\textbf{92.3}\pm0.2}$	91.4±0.4	86.2±0.1	91.9	134K
M-IDPG-DNN	**91.2**±0.3	$\boxed{\textbf{97.6}\pm0.2}$	$\boxed{\textbf{93.5}\pm0.3}$	$\boxed{\textbf{92.6}\pm0.1}$	95.9±0.1	$\boxed{94.5\pm0.2}$	$\boxed{85.5\pm0.6}$	91.8±0.3	$\boxed{91.5\pm0.2}$	$\boxed{86.9\pm0.3}$	92.1	216K

M-IDPG-PHM vs M-IDPG-DNN: -0.3 points while -82K params

Intrinsic Study – what is the efficient way to get the input of prompt generator?

- Using a pre-trained model requires twice of the FLOPS than traditional prompt tuning.



Contextual encoder

Non-contextual encoder

Observation – our method doesn't benefit a lot from a strong contextual LM

Single sentence tasks: MPQA, Subj, CR, MR, SST-2 — sentence pair tasks: QNLI, RTE, MRPC, STS-B, QQP

Method	MPQA	Subj	CR	MR	SST-2	QNLI	RTE	MRPC	STS-B	QQP	Avg	#params
Transformer Fine-tuning												
RoBERTa	90.4±0.2	97.1±0.1	90.7±0.7	91.7±0.2	96.4±0.2	**94.7**±0.1	85.7±0.2	91.8±0.4	92.2±0.2	**91.0**±0.1	92.2	
EFL	90.3±0.2	97.2±0.1	93.0±0.7	91.7±0.2	**96.5**±0.1	94.4±0.1	85.6±2.4	91.2±0.4	**92.5**±0.1	**91.0**±0.2	92.3	
Adapter												
Compacter	91.1±0.2	97.5±0.1	92.7±0.4	**92.6**±0.2	96.0±0.2	94.3±0.2	87.1±1.4	91.6±0.6	91.6±0.1	87.1±0.2	92.2	
Adapter	90.8±0.2	97.5±0.1	92.8±0.3	92.5±0.1	96.1±0.1	94.8±0.2	88.1±0.4	91.8±0.6	92.1±0.1	89.9±0.1	**92.6**	
Prompting												
Prompt-tuning	90.3±0.2	95.5±0.4	91.2±1.1	91.0±0.2	94.2±0.3	86.0±0.3	87.0±0.4	84.3±0.3	87.2±0.2	81.6±0.1	88.8	
Prompt-tuning-134	65.7±19	95.6±0.2	86.7±3.6	89.7±0.5	92.0±0.5	83.0±1.1	87.4±0.5	84.1±0.5	87.6±0.5	82.4±0.3	85.4	
Ptuningv2	90.4±0.3	96.5±0.3	92.7±0.3	91.6±0.1	94.4±0.2	92.9±0.1	78.4±4.3	91.4±0.4	89.9±0.2	84.4±0.4	90.3	
S-IDPG-PHM	89.6±0.3	94.4±0.3	90.3±0.2	89.3±0.4	94.7±0.2	90.7±0.3	89.2±0.2	84.3±0.8	84.7±0.9	82.5±0.2	89.0	
S-IDPG-DNN	89.5±0.7	94.9±0.4	89.9±1.5	90.2±0.6	95.1±0.2	90.5±0.5	**89.4**±0.4	83.0±0.5	85.3±0.7	82.7±0.3	89.1	
M-IDPG-PHM-GloVe	90.9±0.2	97.4±0.1	93.3±0.1	**92.6**±0.3	95.4±0.2	94.4±0.2	82.1±0.6	92.1±0.4	91.0±0.4	86.3±0.2	91.6	141K
M-IDPG-PHM	**91.2**±0.2	97.5±0.1	93.2±0.3	**92.6**±0.3	96.0±0.3	94.5±0.1	83.5±0.7	**92.3**±0.2	91.4±0.4	86.2±0.1	91.9	134K
M-IDPG-DNN	**91.2**±0.3	**97.6**±0.2	**93.5**±0.3	**92.6**±0.1	95.9±0.1	94.5±0.2	85.5±0.6	91.8±0.3	91.5±0.2	86.9±0.3	92.1	

M-IDPG-PHM-GloVe vs M-IDPG-PHM: -0.3 points while reducing half FLOPS

Main results – few-shot setting

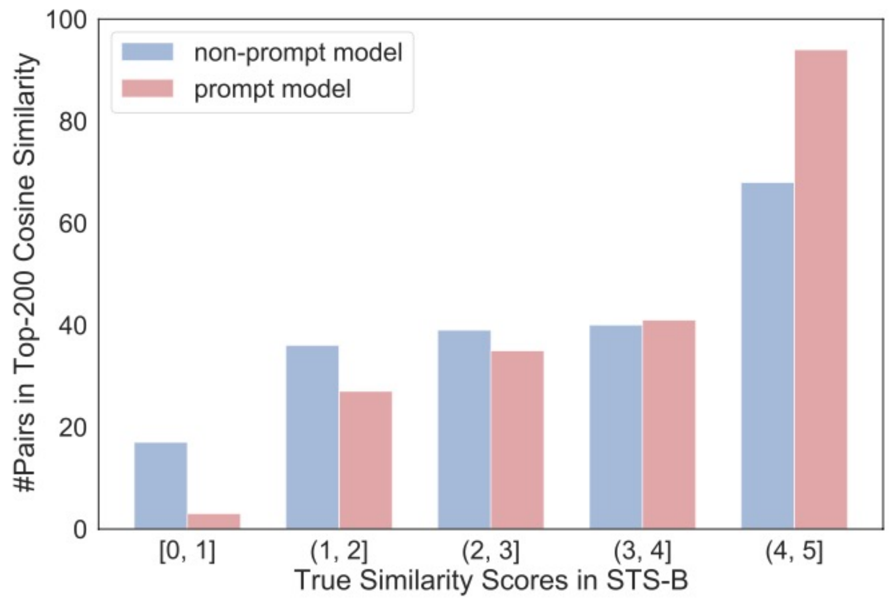
Method	MPQA	Subj	CR	MR	SST-2	QNLI	RTE	MRPC	STS-B	QQP	Avg	
$K = 100$												
Fine-tuning (EFL)	$\mathbf{86.2}_{\pm 0.4}$	$88.4_{\pm 0.8}$	$83.7_{\pm 2.4}$	$81.4_{\pm 1.0}$	$86.2_{\pm 1.3}$	$77.7_{\pm 1.5}$	$84.2_{\pm 1.2}$	$72.6_{\pm 3.7}$	$\mathbf{84.1}_{\pm 1.6}$	$\mathbf{78.1}_{\pm 0.4}$	**82.2**	
Adapter-tuning (Compacter)	$81.0_{\pm 2.9}$	$88.7_{\pm 0.8}$	$\mathbf{84.7}_{\pm 2.1}$	$\mathbf{83.7}_{\pm 0.7}$	$85.7_{\pm 0.9}$	$75.6_{\pm 0.8}$	$84.7_{\pm 0.6}$	$80.0_{\pm 0.9}$	$78.1_{\pm 1.4}$	$77.1_{\pm 0.6}$	81.9	
prompt tuning	$75.9_{\pm 1.6}$	$86.8_{\pm 0.8}$	$72.9_{\pm 1.4}$	$74.1_{\pm 1.4}$	$82.9_{\pm 2.0}$	$\mathbf{82.7}_{\pm 0.2}$	$86.5_{\pm 0.6}$	$80.0_{\pm 1.3}$	$70.2_{\pm 3.1}$	$76.5_{\pm 0.4}$	78.9	
P-Tuningv2	$74.3_{\pm 2.9}$	$89.7_{\pm 0.8}$	$80.1_{\pm 1.0}$	$82.5_{\pm 1.1}$	$85.1_{\pm 1.6}$	$78.2_{\pm 0.5}$	$83.6_{\pm 0.7}$	$\mathbf{80.1}_{\pm 0.6}$	$78.8_{\pm 3.0}$	$76.8_{\pm 0.5}$	80.9	
S-IDPG-PHM	$79.0_{\pm 3.7}$	$87.6_{\pm 1.1}$	$75.0_{\pm 1.6}$	$76.2_{\pm 1.3}$	$87.6_{\pm 1.3}$	$80.4_{\pm 1.2}$	$86.3_{\pm 0.5}$	$79.3_{\pm 0.4}$	$70.9_{\pm 2.5}$	$76.1_{\pm 0.6}$	79.8	
S-IDPG-DNN	$78.0_{\pm 2.1}$	$84.2_{\pm 1.6}$	$76.3_{\pm 4.5}$	$77.4_{\pm 0.5}$	$\mathbf{89.6}_{\pm 1.2}$	$81.1_{\pm 0.8}$	$\mathbf{87.4}_{\pm 0.8}$	$78.8_{\pm 1.3}$	$70.6_{\pm 2.8}$	$74.1_{\pm 0.9}$	79.8	
M-IDPG-PHM-GloVe	$76.6_{\pm 2.0}$	$\mathbf{90.7}_{\pm 0.4}$	$80.6_{\pm 2.6}$	$83.0_{\pm 1.5}$	$85.6_{\pm 0.8}$	$77.9_{\pm 1.3}$	$84.4_{\pm 0.9}$	$79.6_{\pm 0.9}$	$77.8_{\pm 1.6}$	$76.1_{\pm 0.7}$	81.2	
M-IDPG-PHM	$75.5_{\pm 4.6}$	$90.5_{\pm 0.6}$	$80.2_{\pm 1.5}$	$82.5_{\pm 1.1}$	$85.9_{\pm 1.2}$	$78.8_{\pm 1.6}$	$84.0_{\pm 0.4}$	$79.9_{\pm 0.8}$	$79.3_{\pm 0.4}$	$77.1_{\pm 0.2}$	81.4	-0.8
$K = 500$												
Fine-tuning (EFL)	$85.1_{\pm 1.7}$	$94.1_{\pm 0.4}$	$\mathbf{90.9}_{\pm 0.6}$	$87.6_{\pm 0.5}$	$\mathbf{92.5}_{\pm 0.6}$	$85.7_{\pm 0.6}$	$57.5_{\pm 1.0}$	$82.3_{\pm 0.6}$	$\mathbf{88.8}_{\pm 0.5}$	$\mathbf{79.0}_{\pm 0.3}$	84.3	
Adapter-tuning (Compacter)	$\mathbf{86.0}_{\pm 0.8}$	$94.9_{\pm 0.2}$	$89.5_{\pm 1.0}$	$88.5_{\pm 0.2}$	$91.9_{\pm 0.9}$	$82.2_{\pm 0.6}$	$83.9_{\pm 0.8}$	$82.7_{\pm 0.5}$	$86.6_{\pm 0.5}$	$78.9_{\pm 0.3}$	**86.5**	
prompt tuning	$82.4_{\pm 1.3}$	$91.2_{\pm 0.1}$	$86.8_{\pm 0.4}$	$84.6_{\pm 0.8}$	$88.6_{\pm 1.0}$	$\mathbf{86.3}_{\pm 0.4}$	$86.5_{\pm 0.4}$	$80.0_{\pm 0.4}$	$77.4_{\pm 1.9}$	$77.8_{\pm 0.3}$	84.2	
P-Tuningv2	$84.0_{\pm 1.3}$	$94.6_{\pm 0.3}$	$89.0_{\pm 1.8}$	$88.1_{\pm 0.5}$	$91.3_{\pm 0.7}$	$84.6_{\pm 0.8}$	$84.2_{\pm 1.5}$	$\mathbf{83.2}_{\pm 0.7}$	$83.8_{\pm 0.5}$	$78.6_{\pm 0.3}$	86.1	
S-IDPG-PHM	$81.6_{\pm 2.7}$	$91.4_{\pm 0.7}$	$85.8_{\pm 2.0}$	$85.8_{\pm 0.5}$	$88.5_{\pm 1.3}$	$85.0_{\pm 0.4}$	$86.3_{\pm 1.3}$	$81.9_{\pm 0.8}$	$78.3_{\pm 1.5}$	$78.1_{\pm 0.3}$	84.3	
S-IDPG-DNN	$84.8_{\pm 0.7}$	$90.8_{\pm 0.6}$	$89.7_{\pm 1.0}$	$86.1_{\pm 2.8}$	$90.4_{\pm 1.6}$	$84.8_{\pm 0.3}$	$\mathbf{87.7}_{\pm 0.7}$	$82.0_{\pm 1.4}$	$79.1_{\pm 2.3}$	$77.1_{\pm 0.4}$	85.3	
M-IDPG-PHM-GloVe	$84.0_{\pm 1.7}$	$\mathbf{95.0}_{\pm 0.2}$	$89.0_{\pm 1.1}$	$88.1_{\pm 0.5}$	$90.4_{\pm 1.3}$	$85.1_{\pm 0.1}$	$84.0_{\pm 1.0}$	$82.3_{\pm 0.5}$	$84.1_{\pm 0.8}$	$78.2_{\pm 0.8}$	86.0	
M-IDPG-PHM	$85.2_{\pm 1.1}$	$94.6_{\pm 0.0}$	$89.1_{\pm 1.6}$	$\mathbf{88.8}_{\pm 0.4}$	$91.6_{\pm 1.1}$	$84.9_{\pm 0.9}$	$83.9_{\pm 0.7}$	$82.5_{\pm 0.5}$	$84.2_{\pm 0.5}$	$78.6_{\pm 0.3}$	86.3	+2.0
$K = 1000$												
Fine-tuning (EFL)	$87.7_{\pm 0.7}$	$95.1_{\pm 0.2}$	$89.8_{\pm 1.2}$	$89.2_{\pm 0.5}$	$93.6_{\pm 0.4}$	$\mathbf{88.0}_{\pm 0.7}$	$87.3_{\pm 1.3}$	$\mathbf{87.9}_{\pm 0.9}$	$\mathbf{90.8}_{\pm 0.2}$	$79.8_{\pm 0.3}$	**88.9**	
Adapter-tuning (Compacter)	$\mathbf{88.2}_{\pm 0.6}$	$95.6_{\pm 0.3}$	$\mathbf{89.9}_{\pm 1.4}$	$\mathbf{90.0}_{\pm 0.3}$	$92.9_{\pm 0.2}$	$85.2_{\pm 0.7}$	$86.8_{\pm 0.7}$	$86.1_{\pm 0.6}$	$89.6_{\pm 0.5}$	$\mathbf{79.9}_{\pm 0.3}$	88.4	
prompt tuning	$83.9_{\pm 2.0}$	$92.6_{\pm 0.4}$	$87.2_{\pm 1.4}$	$86.7_{\pm 0.3}$	$89.9_{\pm 1.0}$	$86.9_{\pm 0.1}$	$86.4_{\pm 0.7}$	$82.5_{\pm 0.3}$	$82.9_{\pm 1.3}$	$78.6_{\pm 0.3}$	85.8	
P-Tuningv2	$87.0_{\pm 0.9}$	$\mathbf{95.9}_{\pm 0.4}$	$88.3_{\pm 1.5}$	$89.5_{\pm 0.3}$	$93.2_{\pm 0.5}$	$87.4_{\pm 0.4}$	$85.1_{\pm 1.1}$	$82.6_{\pm 1.1}$	$87.8_{\pm 0.3}$	$79.3_{\pm 0.4}$	87.6	
S-IDPG-PHM	$83.4_{\pm 1.7}$	$93.4_{\pm 0.9}$	$89.2_{\pm 0.8}$	$88.0_{\pm 0.9}$	$90.2_{\pm 1.0}$	$85.5_{\pm 0.6}$	$86.9_{\pm 0.6}$	$83.1_{\pm 0.4}$	$83.9_{\pm 0.8}$	$78.9_{\pm 0.4}$	86.3	
S-IDPG-DNN	$85.9_{\pm 0.8}$	$93.3_{\pm 1.2}$	$\mathbf{89.9}_{\pm 0.8}$	$89.6_{\pm 1.1}$	$92.2_{\pm 0.8}$	$85.2_{\pm 1.3}$	$\mathbf{87.7}_{\pm 0.8}$	$82.5_{\pm 0.9}$	$84.7_{\pm 0.9}$	$78.0_{\pm 0.8}$	86.9	
M-IDPG-PHM-GloVe	$86.5_{\pm 0.7}$	$95.5_{\pm 0.3}$	$87.7_{\pm 1.3}$	$89.3_{\pm 0.4}$	$93.4_{\pm 0.3}$	$87.5_{\pm 0.3}$	$84.9_{\pm 0.9}$	$82.7_{\pm 0.7}$	$87.6_{\pm 0.3}$	$79.1_{\pm 0.7}$	87.4	
M-IDPG-PHM	$87.7_{\pm 0.5}$	$95.6_{\pm 0.2}$	$89.2_{\pm 1.2}$	$89.8_{\pm 0.4}$	$\mathbf{93.7}_{\pm 0.6}$	$87.2_{\pm 0.5}$	$85.6_{\pm 0.6}$	$82.5_{\pm 0.9}$	$87.8_{\pm 0.8}$	$79.1_{\pm 0.4}$	87.8	-1.1

Intrinsic study – how prompt helps

- Compare a vanilla model w/o prompts with M-IDPG-PHM

- Sort all sentence pairs in STS-B dev set in descending order by the cosine similarity scores

- Compute top-200 score distribution



Take-aways

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

- We **first** factor in an **instance-dependent prompt**, which is robust to data variance.
- **Parameterized Hypercomplex Multiplication** (PHM) is applied to shrink the training parameters.
- Despite adding few parameters on traditional prompt tuning, IDPG shows **consistent improvement**.
- It is also **on par with** the lightweight adapter tuning methods such as **Compacter** while using a similar amount of trainable parameters.

Thanks

July 26, 2022

Q & A

July 26, 2022

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN