# Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference

Advisor : Jia-Ling, Koh

Speaker : Hsiao-Ting Huang

Source : ACL’2021

Date : 2023/01/10

# Outline

- Introduction
- Method
- Experiment
- Conclusion

# Introduction ： Text Classification

- **task ： 判斷po文者有無精神疾病**

| | text | label |
|---|---|---|
| labeled data | 前幾天看身心科，醫生 說我有憂鬱症，要開始吃藥控制了 ... | positive |
| | 我牙齒痛了好幾天，可能是蛀牙了 ... | negative |
| | 但我可以確定我是邊緣性人格 我已經自殺過兩次了 ... | positive |
| | 最近才曉得交往不久的女朋友有憂鬱症。 發掘的過程就不詳述了 ... | negative |
| | ⋮ | ⋮ |
| unlabeled data | 我是鬱症躁症混合 當鬱期一到，我就隨時隨地想自殘 在手腕上不停地的畫 | ? |

# Introduction

**Problem :**

- the cost of annotating data.
- it is common in real-world uses of NLP to have only a small number of labeled examples.
- applying standard supervised learning to small training sets often performs poorly.

**Goal :**

- With the rise of pretrained language models,
- providing task descriptions could successfully be combined with standard supervised learning in few-shot settings

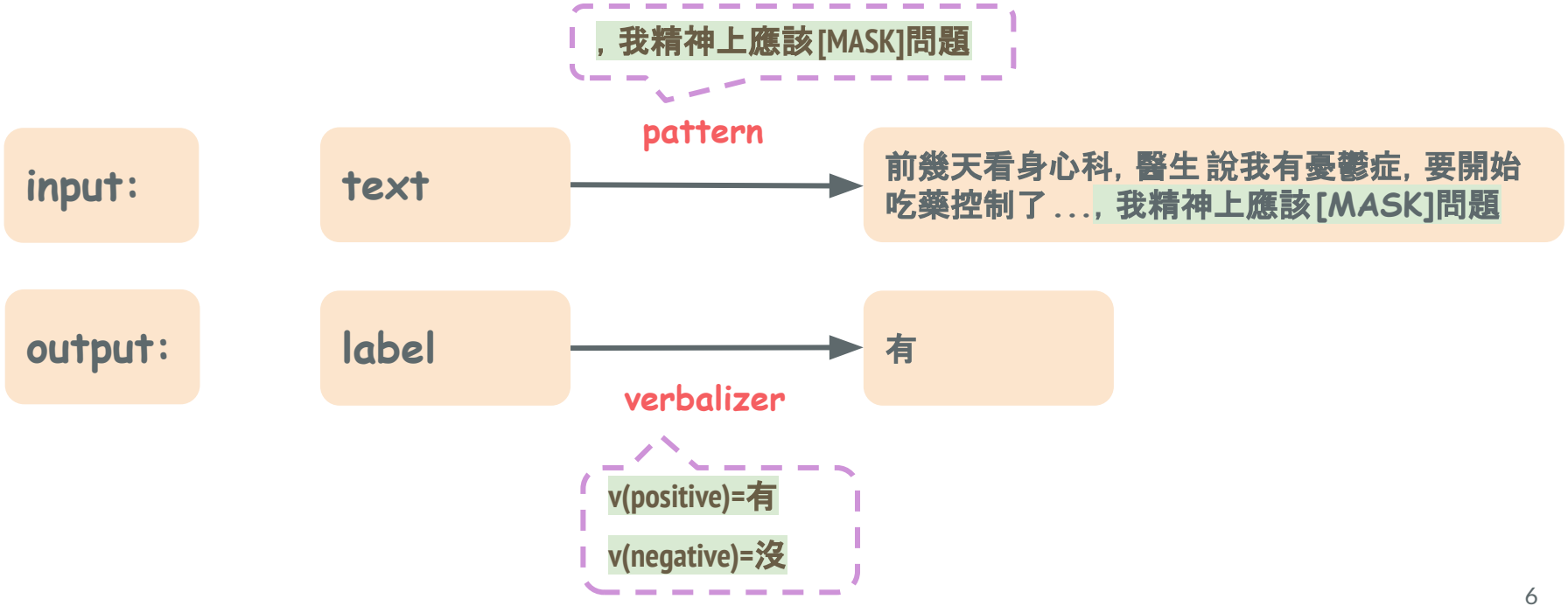
# Introduction : Cloze Questions

- append descriptions in natural language to an input

| | |
|---|---|
| input: | text:<br>前幾天看身心科，醫生 說我有憂鬱症，要開始吃藥控制了 ... |
| output: | label:positive |

# Introduction : Cloze Questions

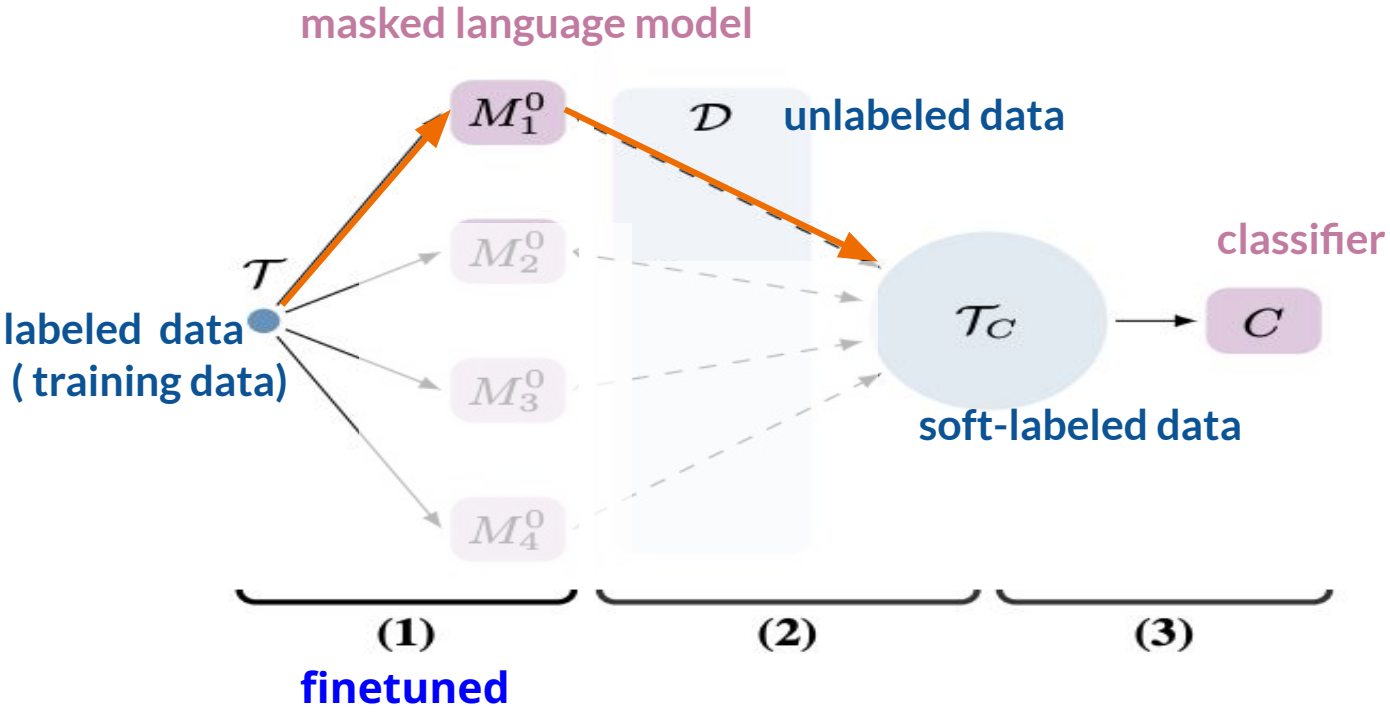
- append descriptions in natural language to an input



# Outline

- Introduction
- Method
- Experiment
- Conclusion

# Method : PET(Pattern-Exploiting Training)

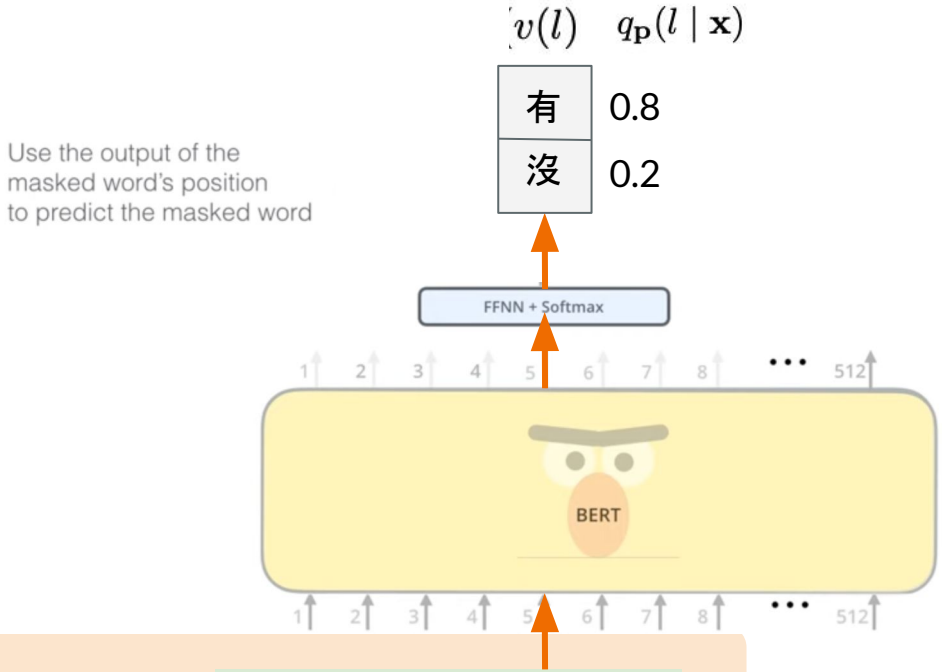


# Method ： PET(Pattern-Exploiting Training)

masked language model:

$$s_{\mathbf{p}}(l \mid \mathbf{x}) = M(v(l) \mid P(\mathbf{x}))$$

$$q_{\mathbf{p}}(l \mid \mathbf{x}) = \frac{e^{s_{\mathbf{p}}(l|\mathbf{x})}}{\sum_{l' \in \mathcal{L}} e^{s_{\mathbf{p}}(l'|\mathbf{x})}}$$




$P(\mathbf{x})$

[CLS]前幾天看身心科，醫生 說我有憂鬱症，要開始吃藥控制了 ...，我精神上應該[MASK]問題

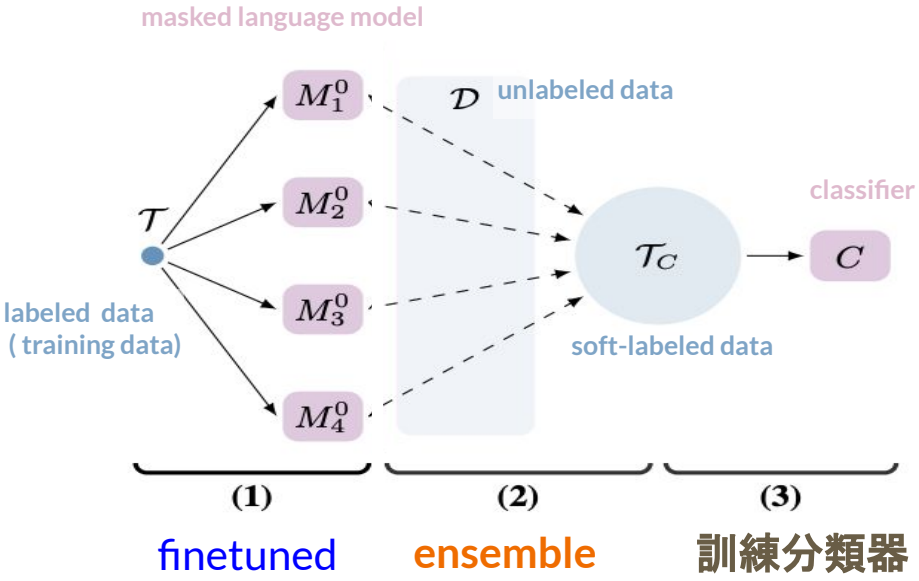
# Method : PET(Pattern-Exploiting Training)

LOSS:

$$L = (1 - \alpha) \cdot L_{\text{CE}} + \alpha \cdot L_{\text{MLM}}$$

$$\alpha = 10^{-4}$$

# Method : PET(Pattern-Exploiting Training)




ensemble:

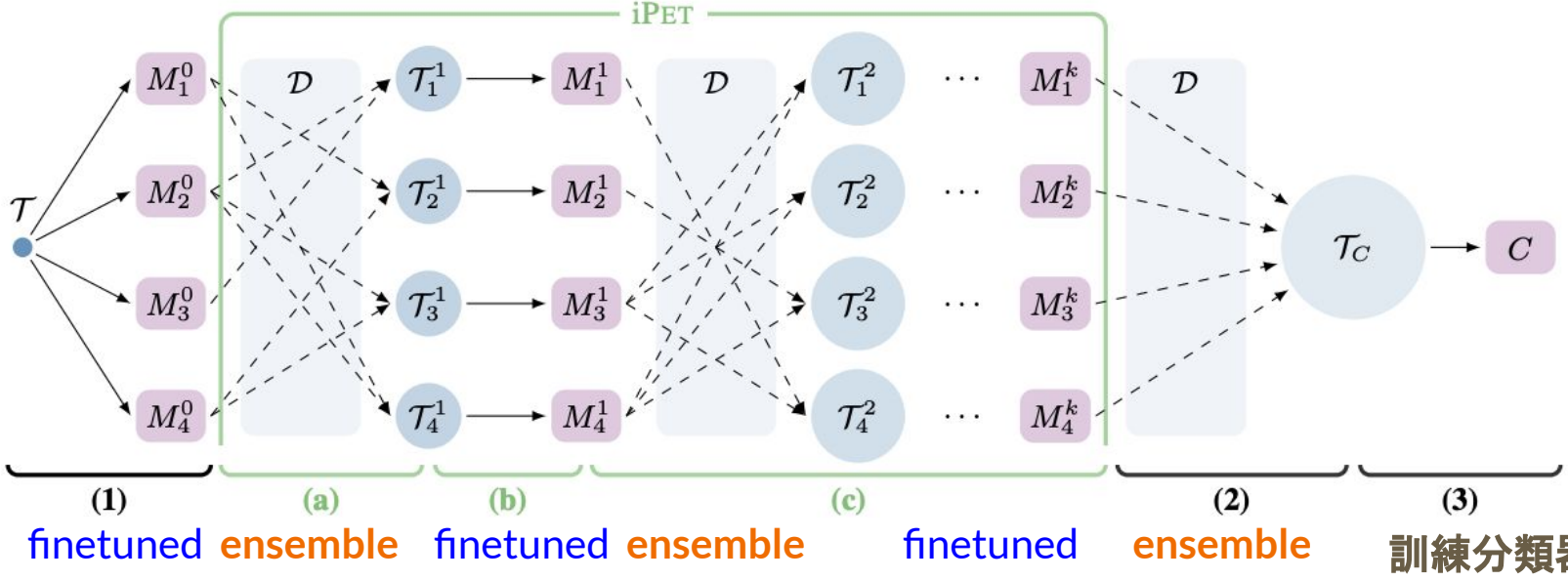
$$s_{\mathcal{M}}(l \mid \mathbf{x}) = \frac{1}{Z} \sum_{\mathbf{p} \in \mathcal{P}} w(\mathbf{p}) \cdot s_{\mathbf{p}}(l \mid \mathbf{x})$$

1. w(p)=1
2. w(p)=acc(train_set_before_training)

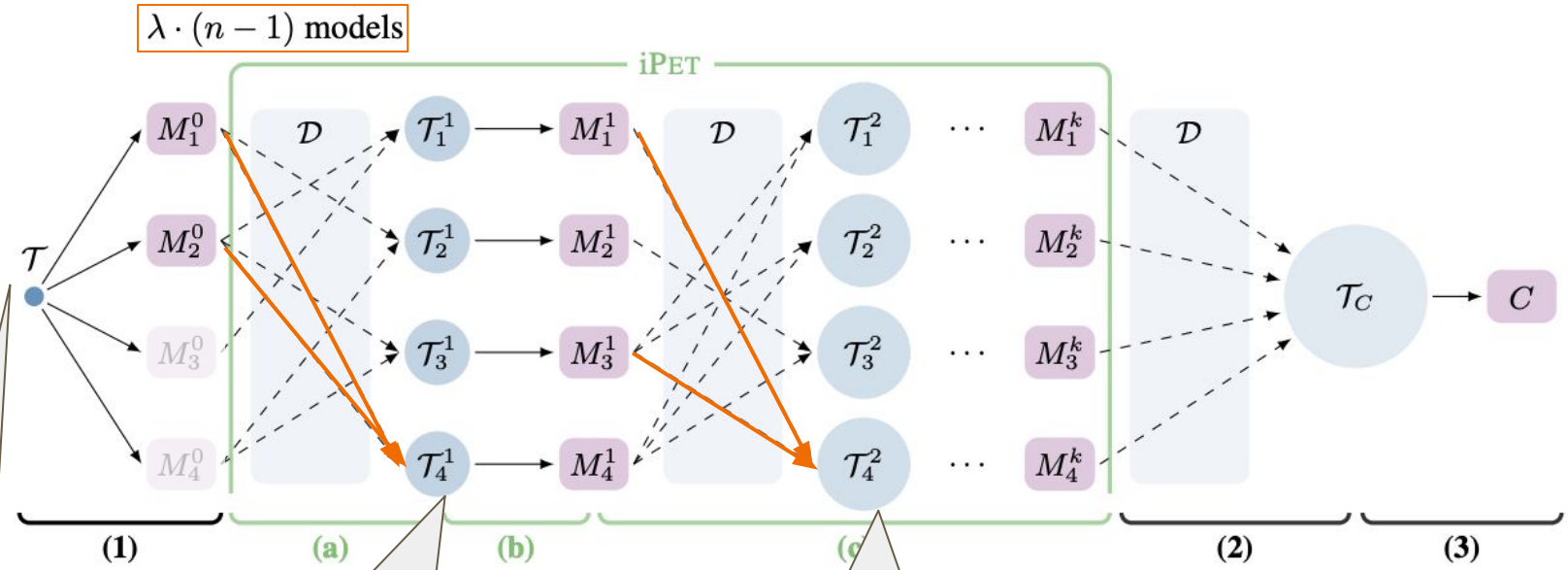
$$q_{\mathbf{p}}(l \mid \mathbf{x}) = \frac{e^{s_{\mathbf{p}}(l|\mathbf{x})}}{\sum_{l' \in \mathcal{L}} e^{s_{\mathbf{p}}(l'|\mathbf{x})}}$$

# iPET(iterative Pattern-Exploiting Training)

The core idea of iPET is to train several generations of models on datasets of increasing size.



# iPET(iterative Pattern-Exploiting Training)




訓練分類器

label(+)=5
label(-) =5

label(+)=5*d(=4)=20
label(-) =5*4=20

label(+)=20*4=80
label(-) =20*4=80

# Outline

- Introduction
- Method
- Experiment
- Conclusion

# Datasets-Yelp

task : estimate the rating that a customer gave to a restaurant on a 1to 5-star scale based on their review's text

- pattern:

$P_1(a) =$ It was ____. $a$ $P_2(a) =$ Just ____! || $a$

$P_3(a) =$ $a$. All in all, it was ____.

$P_4(a) =$ $a$ || In summary, the restaurant is ____.

- verbalizer:

$v(1) =$ terrible $v(2) =$ bad $v(3) =$ okay

$v(4) =$ good $v(5) =$ great

# Datasets-AG's News

task:a headline a and text body b, news have to be classified as belonging to one of the categories.

- pattern:

$P_1(\mathbf{x}) =$ ____: $a$ $b$ $\quad$ $P_2(\mathbf{x}) =$ $a$ ( ____ ) $b$

$P_3(\mathbf{x}) =$ ____ – $a$ $b$ $\quad$ $P_4(\mathbf{x}) =$ $a$ $b$ ( ____ )

$P_5(\mathbf{x}) =$ ____ News: $a$ $b$

$P_6(\mathbf{x}) =$ [ Category: ____ ] $a$ $b$

- verbalizer:
  - v(1) = World
  - v(2) = Sports
  - v(3) = Business
  - v(4) = Tech

agnews

| | headline | textbody |
|---|---|---|
| Business | Wall St. Bears Claw Back Into the Black (Reuters) | Reuters - Short-sellers, Wall Street's dwindling\band of ultra-cynics, are seeing green again.(股票) |
| Business | Carlyle Looks Toward Commercial Aerospace (Reuters) | Reuters - Private investment firm Carlyle Group,\which has a reputation for making well-timed and occasionally\controversial plays in the defense industry, has quietly placed\its bets on another part of the market.(投資) |
| Sports | USC starts at the top | Southern California greeted news of its first preseason No. 1 ranking since 1979 with ambivalence.(南加州季前賽排名第一) |
| World | Seven Georgian soldiers wounded as South Ossetia ceasefire violated (AFP) | AFP - Sporadic gunfire and shelling took place overnight in the disputed Georgian region of South Ossetia in violation of a fragile ceasefire, wounding seven Georgian servicemen.(法新社-南奧塞梯地區一夜之間發生零星的槍擊和砲擊) |

# Datasets-Yahoo

task:Yahoo Questions is a text classification dataset. Given a question a and an answer b, one of ten possible categories has to be assigned

- pattern:a:問題, b:答案

$P_1(\mathbf{x}) =$ ____: $a$ $b$ $P_2(\mathbf{x}) =$ $a$ ( ____ ) $b$

$P_3(\mathbf{x}) =$ ____ – $a$ $b$ $P_4(\mathbf{x}) =$ $a$ $b$ ( ____ )

Question

$P_5(\mathbf{x}) =$ ____ News: $a$ $b$

$P_6(\mathbf{x}) =$ [ Category: ____ ] $a$ $b$

- verbalizer:

v(1) = Society
v(2) = Science
v(3) = Health
v(4) = Education
v(5) = Computer
.
.
.

# Experiment

| Line | Examples | Method | Yelp | AG's | Yahoo | MNLI (m/mm) |
|---|---|---|---|---|---|---|
| 1 | $\|\mathcal{T}\| = 0$ | unsupervised (avg) | 33.8 ±9.6 | 69.5 ±7.2 | 44.0 ±9.1 | 39.1 ±4.3 / 39.8 ±5.1 |
| 2 | | unsupervised (max) | 40.8 ±0.0 | 79.4 ±0.0 | 56.4 ±0.0 | 43.8 ±0.0 / 45.0 ±0.0 |
| 3 | | iPET | **56.7** ±0.2 | **87.5** ±0.1 | **70.7** ±0.1 | **53.6** ±0.1 / **54.2** ±0.1 |
| 4 | $\|\mathcal{T}\| = 10$ | supervised | 21.1 ±1.6 | 25.0 ±0.1 | 10.1 ±0.1 | 34.2 ±2.1 / 34.1 ±2.0 |
| 5 | | PET | 52.9 ±0.1 | 87.5 ±0.0 | 63.8 ±0.2 | 41.8 ±0.1 / 41.5 ±0.2 |
| 6 | | iPET | **57.6** ±0.0 | **89.3** ±0.1 | **70.7** ±0.1 | **43.2** ±0.0 / **45.7** ±0.1 |
| 7 | $\|\mathcal{T}\| = 50$ | supervised | 44.8 ±2.7 | 82.1 ±2.5 | 52.5 ±3.1 | 45.6 ±1.8 / 47.6 ±2.4 |
| 8 | | PET | 60.0 ±0.1 | 86.3 ±0.0 | 66.2 ±0.1 | 63.9 ±0.0 / 64.2 ±0.0 |
| 9 | | iPET | **60.7** ±0.1 | **88.4** ±0.1 | **69.7** ±0.0 | **67.4** ±0.3 / **68.3** ±0.3 |
| 10 | $\|\mathcal{T}\| = 100$ | supervised | 53.0 ±3.1 | 86.0 ±0.7 | 62.9 ±0.9 | 47.9 ±2.8 / 51.2 ±2.6 |
| 11 | | PET | 61.9 ±0.0 | 88.3 ±0.1 | 69.2 ±0.0 | 74.7 ±0.3 / 75.9 ±0.4 |
| 12 | | iPET | **62.9** ±0.0 | **89.6** ±0.1 | **71.2** ±0.1 | **78.4** ±0.7 / **78.6** ±0.5 |
| 13 | $\|\mathcal{T}\| = 1000$ | supervised | 63.0 ±0.5 | **86.9** ±0.4 | 70.5 ±0.3 | 73.1 ±0.2 / 74.8 ±0.3 |
| 14 | | PET | **64.8** ±0.1 | **86.9** ±0.2 | **72.7** ±0.0 | **85.3** ±0.2 / **85.5** ±0.4 |

zero-shot

# Experiment

| Line | Examples | Method | Yelp | AG's | Yahoo | MNLI (m/mm) |
|---|---|---|---|---|---|---|
| 1 | $\lvert\mathcal{T}\rvert = 0$ | unsupervised (avg) | 33.8 ±9.6 | 69.5 ±7.2 | 44.0 ±9.1 | 39.1 ±4.3 / 39.8 ±5.1 |
| 2 | | unsupervised (max) | 40.8 ±0.0 | 79.4 ±0.0 | 56.4 ±0.0 | 43.8 ±0.0 / 45.0 ±0.0 |
| 3 | | iPET | **56.7** ±0.2 | **87.5** ±0.1 | **70.7** ±0.1 | **53.6** ±0.1 / **54.2** ±0.1 |
| 4 | $\lvert\mathcal{T}\rvert = 10$ | supervised | 21.1 ±1.6 | 25.0 ±0.1 | 10.1 ±0.1 | 34.2 ±2.1 / 34.1 ±2.0 |
| 5 | | PET | 52.9 ±0.1 | 87.5 ±0.0 | 63.8 ±0.2 | 41.8 ±0.1 / 41.5 ±0.2 |
| 6 | | iPET | **57.6** ±0.0 | **89.3** ±0.1 | **70.7** ±0.1 | **43.2** ±0.0 / **45.7** ±0.1 |
| 7 | $\lvert\mathcal{T}\rvert = 50$ | supervised | 44.8 ±2.7 | 82.1 ±2.5 | 52.5 ±3.1 | 45.6 ±1.8 / 47.6 ±2.4 |
| 8 | | PET | 60.0 ±0.1 | 86.3 ±0.0 | 66.2 ±0.1 | 63.9 ±0.0 / 64.2 ±0.0 |
| 9 | | iPET | **60.7** ±0.1 | **88.4** ±0.1 | **69.7** ±0.0 | **67.4** ±0.3 / **68.3** ±0.3 |
| 10 | $\lvert\mathcal{T}\rvert = 100$ | supervised | 53.0 ±3.1 | 86.0 ±0.7 | 62.9 ±0.9 | 47.9 ±2.8 / 51.2 ±2.6 |
| 11 | | PET | 61.9 ±0.0 | 88.3 ±0.1 | 69.2 ±0.0 | 74.7 ±0.3 / 75.9 ±0.4 |
| 12 | | iPET | **62.9** ±0.0 | **89.6** ±0.1 | **71.2** ±0.1 | **78.4** ±0.7 / **78.6** ±0.5 |
| 13 | $\lvert\mathcal{T}\rvert = 1000$ | supervised | 63.0 ±0.5 | **86.9** ±0.4 | 70.5 ±0.3 | 73.1 ±0.2 / 74.8 ±0.3 |
| 14 | | PET | **64.8** ±0.1 | **86.9** ±0.2 | **72.7** ±0.0 | **85.3** ±0.2 / **85.5** ±0.4 |

zero-shot

few-shot

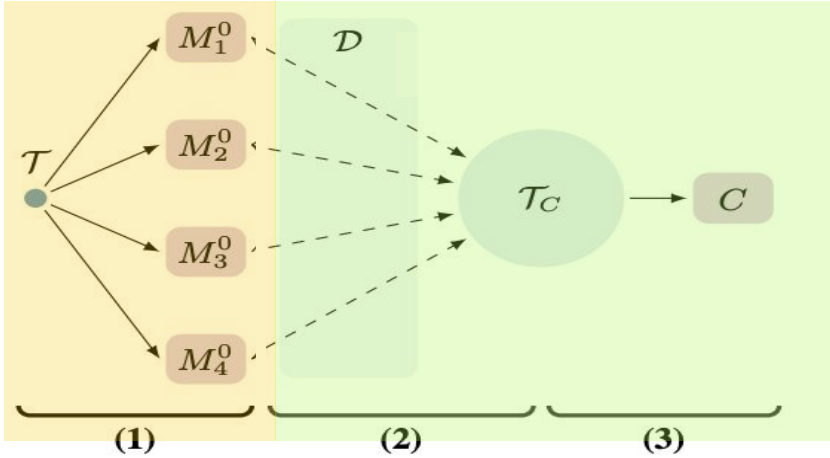
# Experiment

| Line | Examples | Method | Yelp | AG's | Yahoo | MNLI (m/mm) |
|---|---|---|---|---|---|---|
| 1 | $\|\mathcal{T}\| = 0$ | unsupervised (avg) | 33.8 ±9.6 | 69.5 ±7.2 | 44.0 ±9.1 | 39.1 ±4.3 / 39.8 ±5.1 |
| 2 | | unsupervised (max) | 40.8 ±0.0 | 79.4 ±0.0 | 56.4 ±0.0 | 43.8 ±0.0 / 45.0 ±0.0 |
| 3 | | iPET | **56.7** ±0.2 | **87.5** ±0.1 | **70.7** ±0.1 | **53.6** ±0.1 / **54.2** ±0.1 |
| 4 | $\|\mathcal{T}\| = 10$ | supervised | 21.1 ±1.6 | 25.0 ±0.1 | 10.1 ±0.1 | 34.2 ±2.1 / 34.1 ±2.0 |
| 5 | | PET | 52.9 ±0.1 | 87.5 ±0.0 | 63.8 ±0.2 | 41.8 ±0.1 / 41.5 ±0.2 |
| 6 | | iPET | **57.6** ±0.0 | **89.3** ±0.1 | **70.7** ±0.1 | **43.2** ±0.0 / **45.7** ±0.1 |
| 7 | $\|\mathcal{T}\| = 50$ | supervised | 44.8 ±2.7 | 82.1 ±2.5 | 52.5 ±3.1 | 45.6 ±1.8 / 47.6 ±2.4 |
| 8 | | PET | 60.0 ±0.1 | 86.3 ±0.0 | 66.2 ±0.1 | 63.9 ±0.0 / 64.2 ±0.0 |
| 9 | | iPET | **60.7** ±0.1 | **88.4** ±0.1 | **69.7** ±0.0 | **67.4** ±0.3 / **68.3** ±0.3 |
| 10 | $\|\mathcal{T}\| = 100$ | supervised | 53.0 ±3.1 | 86.0 ±0.7 | 62.9 ±0.9 | 47.9 ±2.8 / 51.2 ±2.6 |
| 11 | | PET | 61.9 ±0.0 | 88.3 ±0.1 | 69.2 ±0.0 | 74.7 ±0.3 / 75.9 ±0.4 |
| 12 | | iPET | **62.9** ±0.0 | **89.6** ±0.1 | **71.2** ±0.1 | **78.4** ±0.7 / **78.6** ±0.5 |
| 13 | $\|\mathcal{T}\| = 1000$ | supervised | 63.0 ±0.5 | **86.9** ±0.4 | 70.5 ±0.3 | 73.1 ±0.2 / 74.8 ±0.3 |
| 14 | | PET | **64.8** ±0.1 | **86.9** ±0.2 | **72.7** ±0.0 | **85.3** ±0.2 / **85.5** ±0.4 |

# Experiment

| Method | Yelp | AG's | Yahoo | MNLI |
|---|---|---|---|---|
| min | 39.6 | 82.1 | 50.2 | 36.4 |
| max | 52.4 | 85.0 | 63.6 | 40.2 |
| PET (no distillation) | 51.7 | 87.0 | 62.8 | 40.6 |
| PET uniform | 52.7 | 87.3 | **63.8** | **42.0** |
| PET weighted | **52.9** | **87.5** | **63.8** | 41.8 |

Table 4: Minimum (min) and maximum (max) accuracy of models based on individual PVPs as well as PET with and without knowledge distillation ($|\mathcal{T}| = 10$).



# Conclusion

- given a small to medium number of labeled examples, PET and iPET substantially outperform unsupervised approaches, supervised training and strong semi-supervised baselines.

- With the rise of pretrained language models (PLMs) such as GPT , BERT and RoBERTa , the idea of providing task descriptions has become feasible for neural architectures