dongguk UNIVERSITY

# Learning to few-shot learning across the text classification tasks

Daeung Kim

2020.06.12.

# INDEX

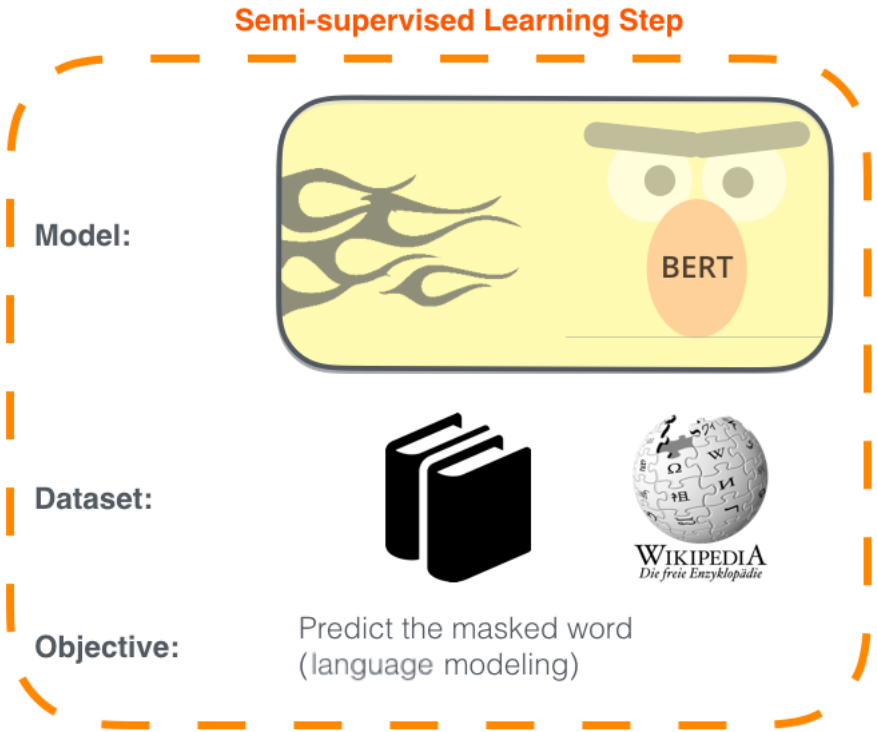
# 1. Prerequisites

# 1. Prerequisites

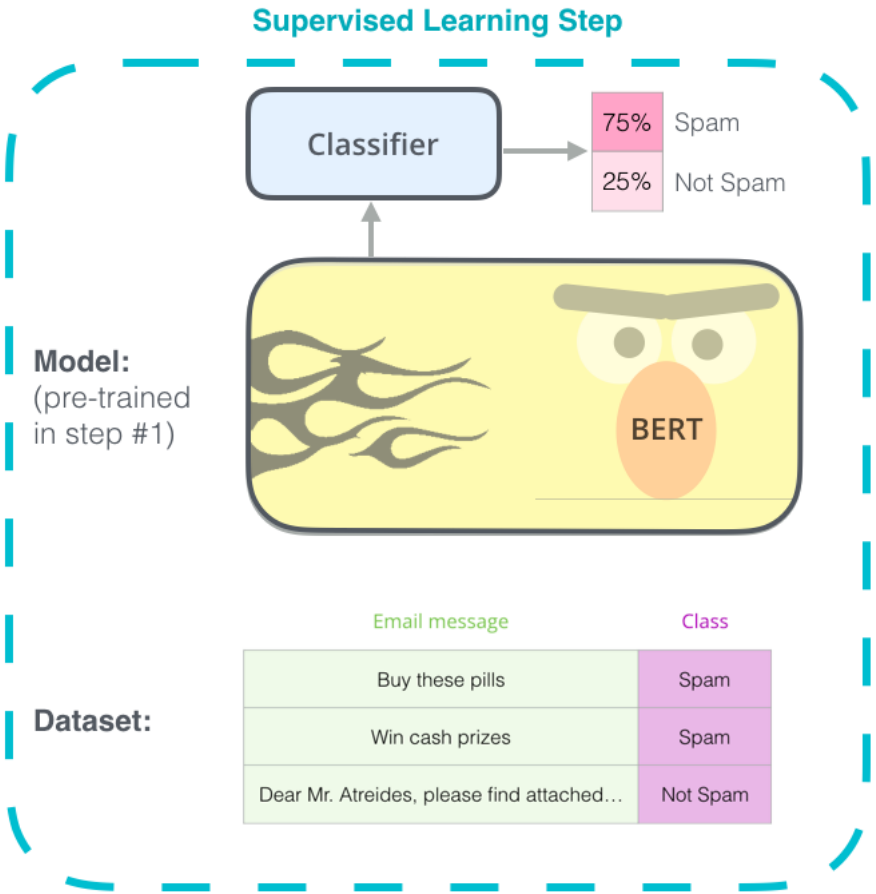
## Pre-training Language Model

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.




2 - Supervised training on a specific task with a labeled dataset.




https://medium.com/huggingface/introducing-fastbert-a-simple-deep-learning-library-for-bert-models-89ff763ad384

# 1. Prerequisites

## GLUE

- A Multi-Task Benchmark And Analysis Platform For Natural Language Understanding


| Corpus | \|Train\| | \|Test\| | Task | Metrics | Domain |
|---|---|---|---|---|---|
| | | | Single-Sentence Tasks | | |
| CoLA | 8.5k | **1k** | acceptability | Matthews corr. | misc. |
| SST-2 | 67k | 1.8k | sentiment | acc. | movie reviews |
| | | | Similarity and Paraphrase Tasks | | |
| MRPC | 3.7k | 1.7k | paraphrase | acc./F1 | news |
| STS-B | 7k | 1.4k | sentence similarity | Pearson/Spearman corr. | misc. |
| QQP | 364k | **391k** | paraphrase | acc./F1 | social QA questions |
| | | | Inference Tasks | | |
| MNLI | 393k | **20k** | NLI | matched acc./mismatched acc. | misc. |
| QNLI | 105k | 5.4k | QA/NLI | acc. | Wikipedia |
| RTE | 2.5k | 3k | NLI | acc. | news, Wikipedia |
| WNLI | 634 | **146** | coreference/NLI | acc. | fiction books |

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

https://gluebenchmark.com/

# 1. Prerequisites

## GLUE

- A Multi-Task Benchmark And Analysis Platform For Natural Language Understanding

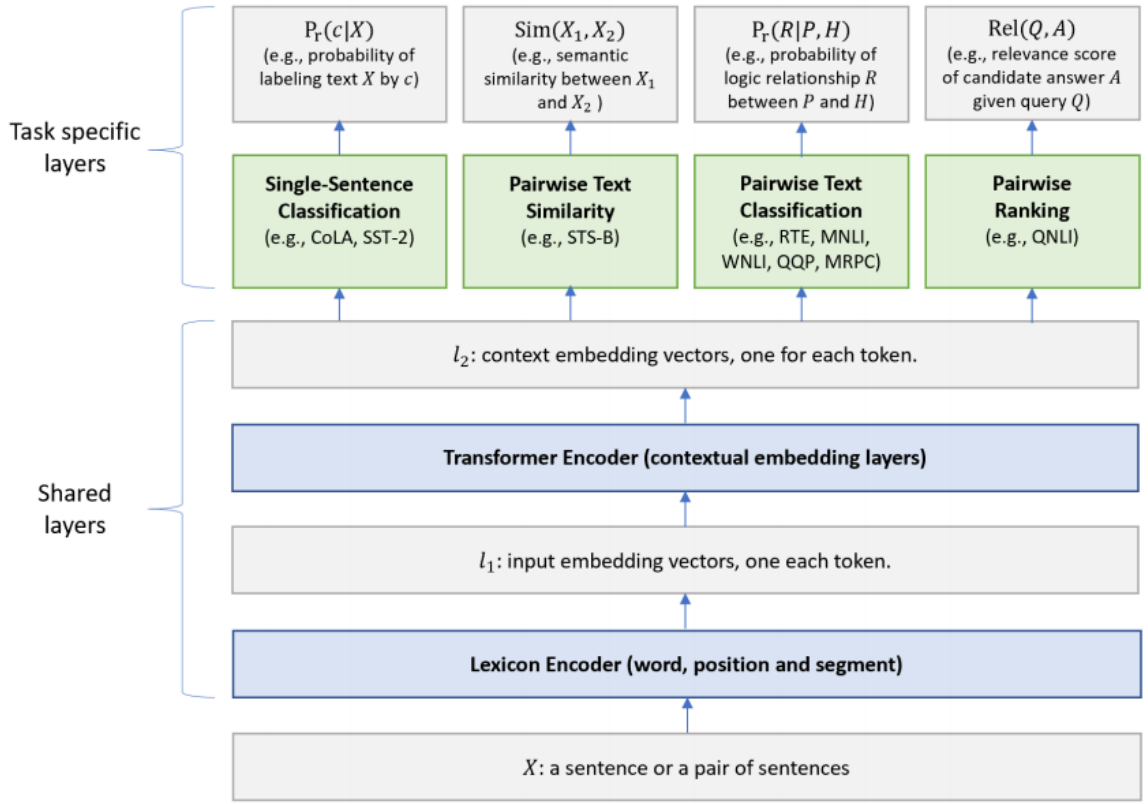
| Dataset | Description | Data example | Metric |
|---|---|---|---|
| CoLA | Is the sentence grammatical or ungrammatical? | "This building is than that one."<br>= **Ungrammatical** | Matthews |
| SST-2 | Is the movie review positive, negative, or neutral? | "The movie is funny , smart , visually inventive , and most of all , alive ."<br>= **.93056 (Very Positive)** | Accuracy |
| MRPC | Is the sentence B a paraphrase of sentence A? | A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ."<br>B) "The island reported another 35 probable cases yesterday , taking its total to 418 ."<br>= **A Paraphrase** | Accuracy / F1 |
| STS-B | How similar are sentences A and B? | A) "Elephants are walking down a trail."<br>B) "A herd of elephants are walking along a trail."<br>= **4.6 (Very Similar)** | Pearson / Spearman |
| QQP | Are the two questions similar? | A) "How can I increase the speed of my internet connection while using a VPN?"<br>B) "How can Internet speed be increased by hacking through DNS?"<br>= **Not Similar** | Accuracy / F1 |
| MNLI-mm | Does sentence A entail or contradict sentence B? | A) "Tourist Information offices can be very helpful."<br>B) "Tourist Information offices are never of any help."<br>= **Contradiction** | Accuracy |
| QNLI | Does sentence B contain the answer to the question in sentence A? | A) "What is essential for the mating of the elements that create radio waves?"<br>B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field."<br>= **Answerable** | Accuracy |
| RTE | Does sentence A entail sentence B? | A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members."<br>B) "Yunus supported more than 50,000 Struggling Members."<br>= **Entailed** | Accuracy |
| WNLI | Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun? | A) "Lily spoke to Donna, breaking her concentration."<br>B) "Lily spoke to Donna, breaking Lily's concentration."<br>= **Incorrect Referent** | Accuracy |

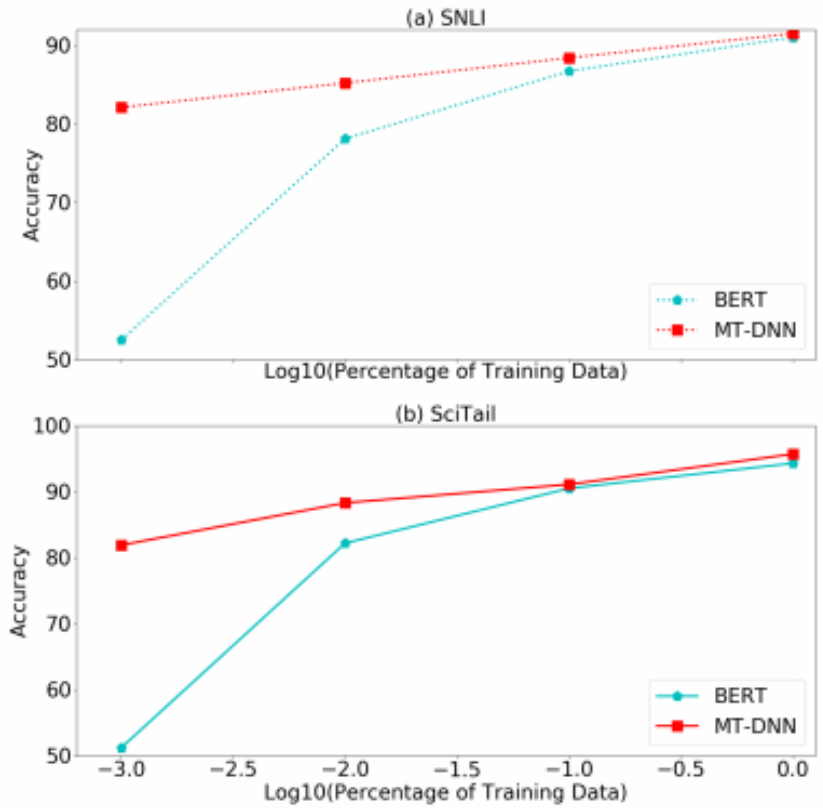
https://docs.google.com/spreadsheets/d/1BrOdjJgky7FfeiwC_VDURZuRPUFUAz_jfczPPT35P00/edit#gid=0

# 1. Prerequisites

## MT-DNN (Liu et al., 2019)

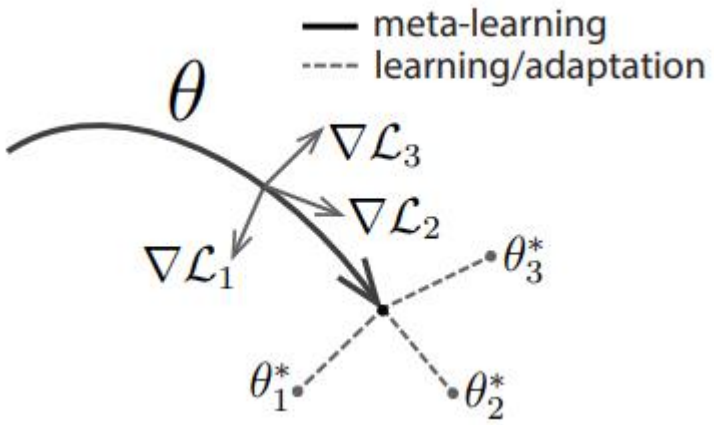
- Multi-task Deep Neural Networks for Natural Language Understanding
  - It Improves the performance by multi-task learning



# 1. Prerequisites

## Modal-Agnostic Meta-Learning(MAML) (Finn et al., 2017)




**Algorithm 1** Model-Agnostic Meta-Learning

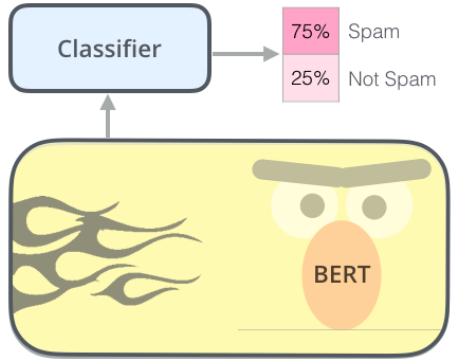
**Require:** $p(\mathcal{T})$: distribution over tasks
**Require:** $\alpha$, $\beta$: step size hyperparameters
1: randomly initialize $\theta$
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ with respect to $K$ examples
6: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
7: **end for**
8: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
9: **end while**

# 2. LEOPARD

# 2. LEOPARD

Learning to gEnerate sOftmax Parameters foR Diverse classification

- **Problem Definition**
  - Fine-tuning on a new task still requires **large amount of task-specific labelled data** to achieve good performance
  - Consider this problem of learning to **generalize to new tasks with few samples** as a **meta-learning** problem




$\tau_i \sim P(T)$

$\tau_1$

| Email message | Class |
|---|---|
| Buy these pills | Spam |
| Win cash prizes | Spam |
| Dear Mr. Atreides, please find attached… | Not Spam |

$\tau_2$

| Email message | Class |
|---|---|
| Buy these pills | Spam |
| Win cash prizes | Spam |
| Dear Mr. Atreides, please find attached… | Not Spam |

…

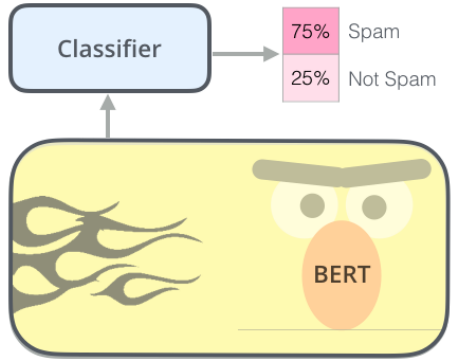
$\tau_i$

| Email message | Class |
|---|---|
| Buy these pills | Spam |
| Win cash prizes | Spam |
| Dear Mr. Atreides, please find attached… | Not Spam |

# 2. LEOPARD

## Learning to gEnerate sOftmax Parameters foR Diverse classification

- **Problem Definition**
  - But, in the NLP field, meta-learning's application still limited to simulated problems or problems with limited diversity across tasks.
    - E.g. FewRel(Han et al., 2018), Amazon Review Sentiment Classification(Blitzer et al., 2007), 20 Newsgroups(Lang, 1995)
  - LEOPARD enables optimization-based meta-learning across tasks **with different number of classes. So there is no limitation on the type of task.**




$\tau_i \sim P(T)$

$\tau_1$

| Email message | Class |
|---|---|
| Buy these pills | Spam |
| Win cash prizes | Spam |
| Dear Mr. Atreides, please find attached… | Not Spam |

$\tau_2$

| Email message | Class |
|---|---|
| Buy these pills | Spam |
| Win cash prizes | Spam |
| Dear Mr. Atreides, please find attached… | Not Spam |

…

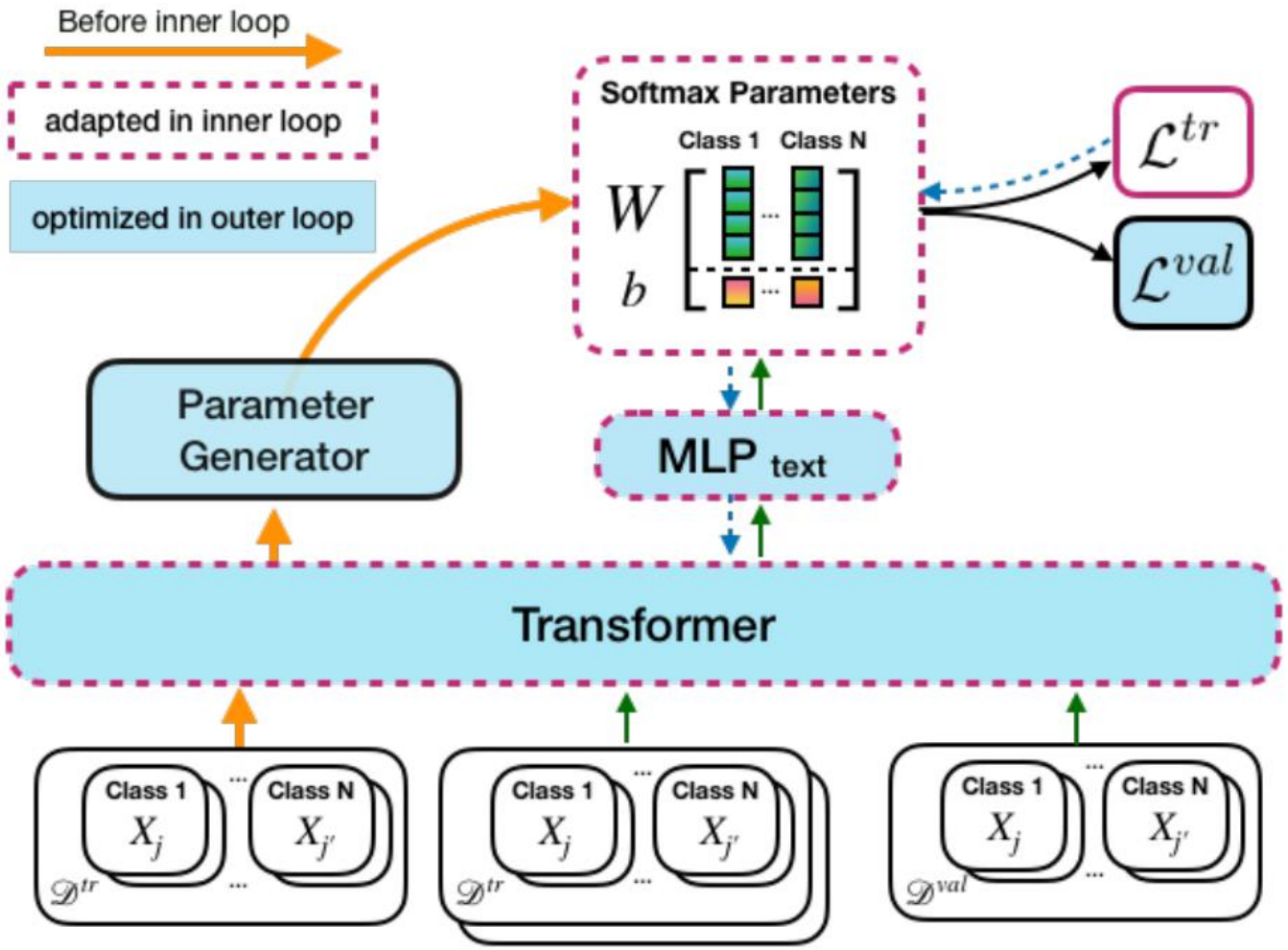
$\tau_i$

| Email message | Class |
|---|---|
| Buy these pills | Spam |
| Win cash prizes | Spam |
| Dear Mr. Atreides, please find attached… | Not Spam |

# 2. LEOPARD

## Model

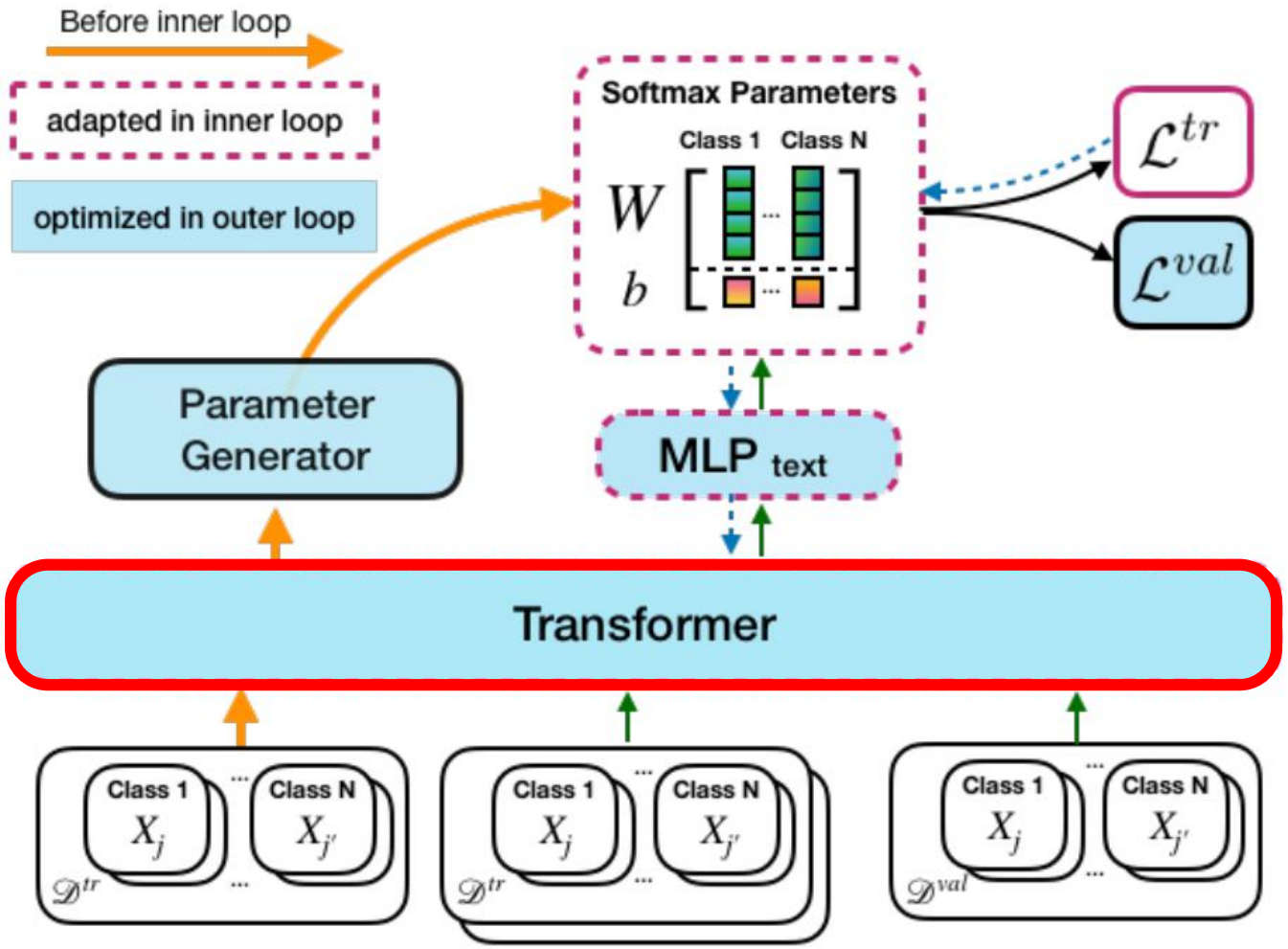
1. A Shared Neural Input Encoder
2. A Softmax Parameter Generator
3. A MAML-based Adaptation Method



# 2. LEOPARD

## Model

1. A Shared Neural Input Encoder
   - BERT-base Model(Devlin et al., 2018)
   - Generate feature representations useful across tasks

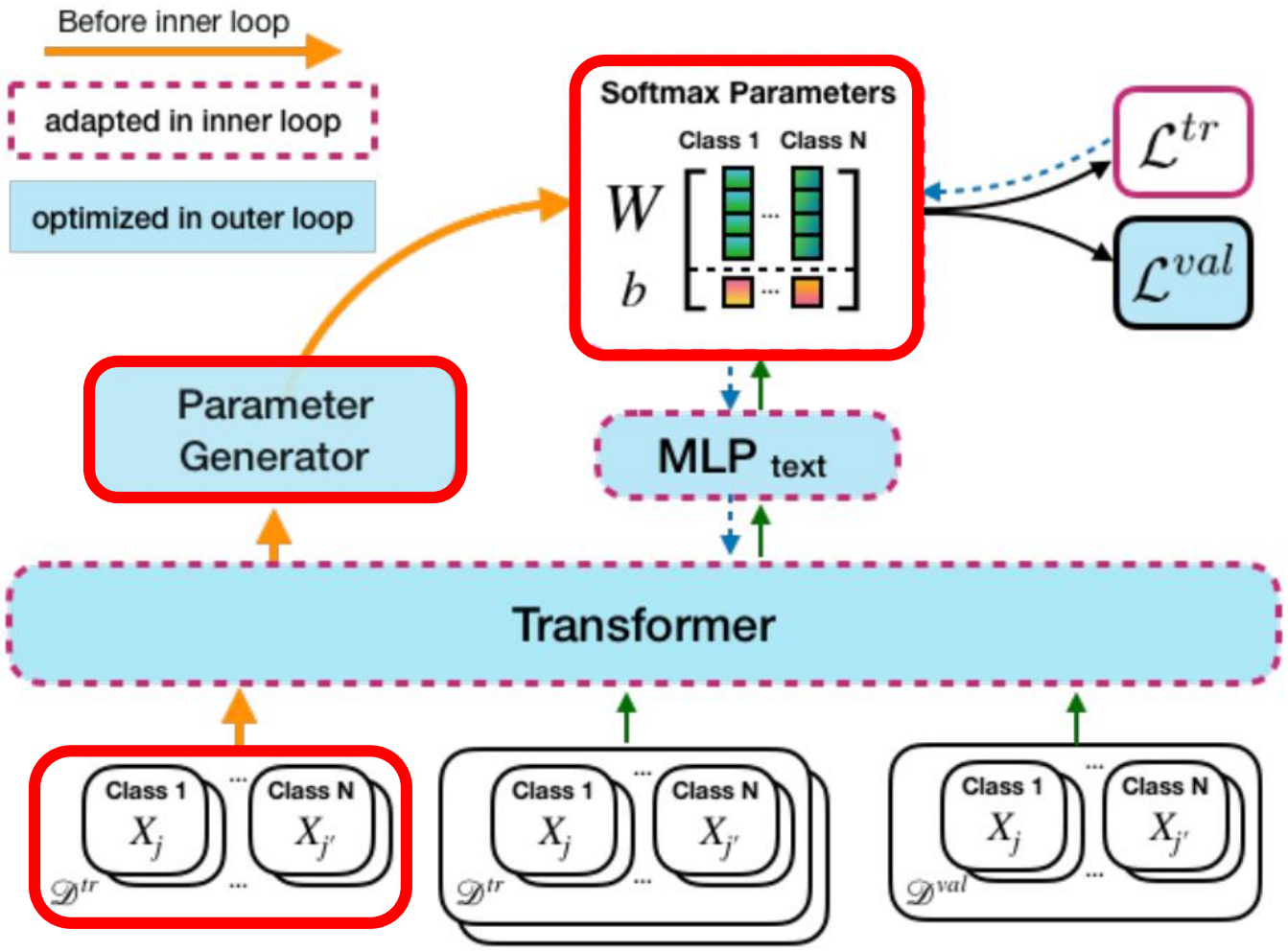


# 2. LEOPARD

## Model

### 2. A Softmax Parameter Generator

- Conditioned on the meta-training dataset for an $N$-way task, which generates the softmax parameters for the task
- MAML can't deal with a variable $N$ *(i.e. the number of classes)*
- Like Prototypical Networks, it builds a linear layer to compute the score for a certain class. It takes a subset of an $N$-way task, which are belong to a certain class and outputs the similarity score to certain class

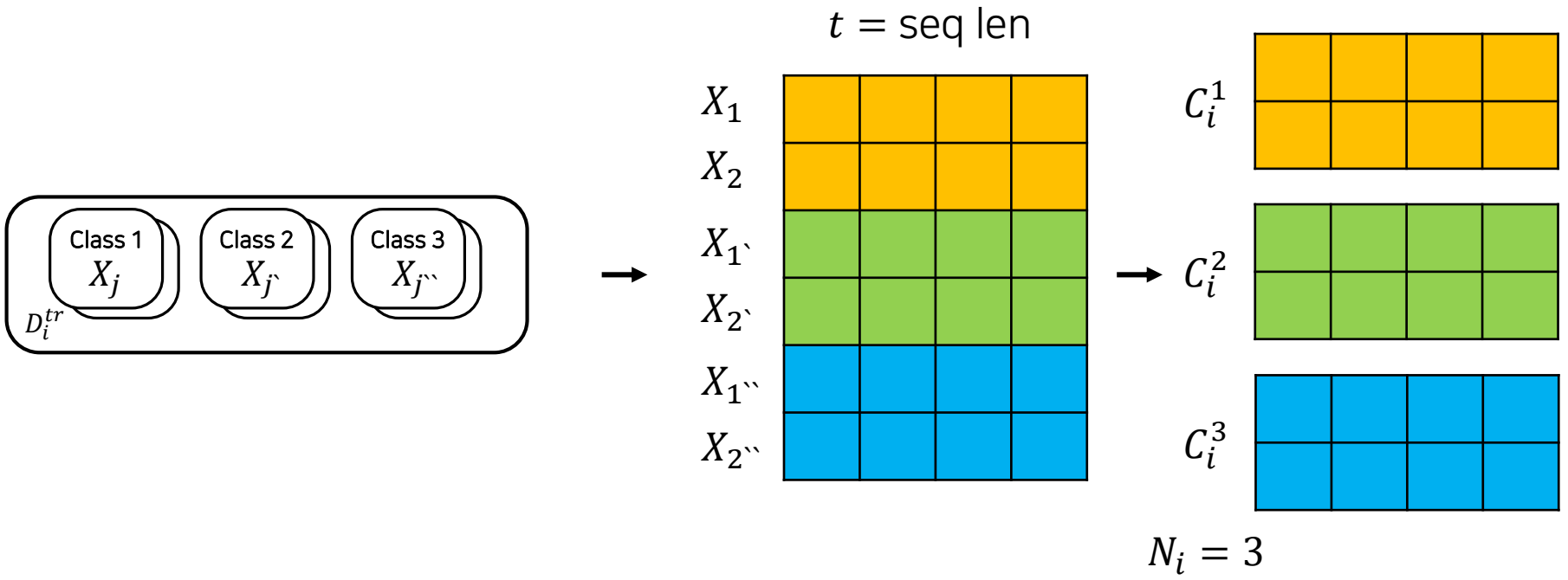


# 2. LEOPARD

## Model

### 2. A Softmax Parameter Generator

1. Given the training data $D_i^{tr} = \{(x_j, y_j)\}$, for a task $T_i$ in episode, the input is partitioned into the $N_i$ number of classes for the task

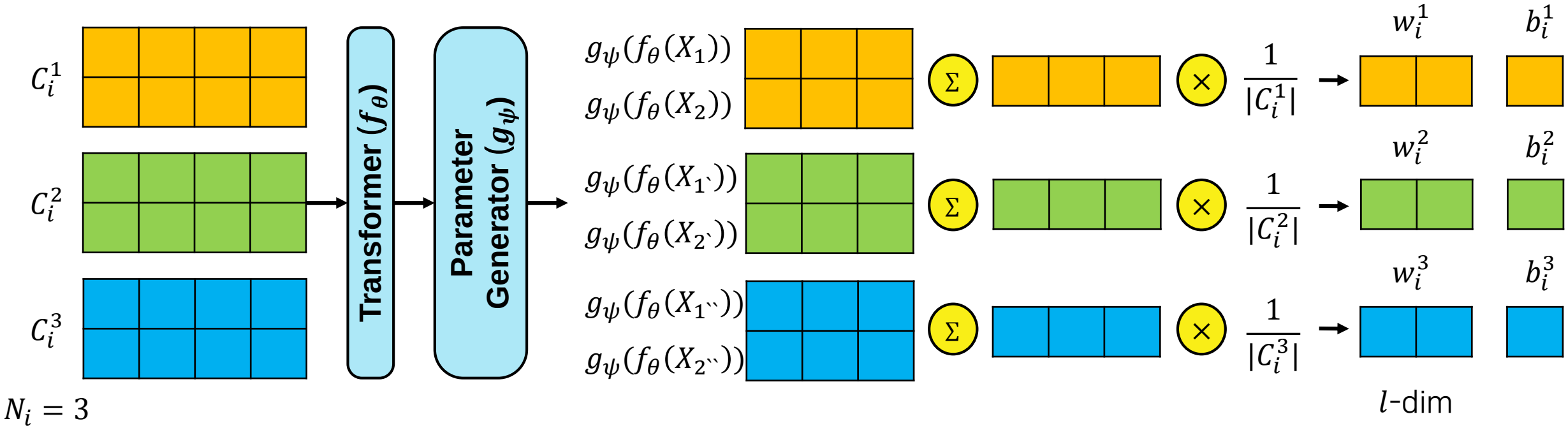


# 2. LEOPARD

## Model

### 2. A Softmax Parameter Generator

2. Text encoder ($\boldsymbol{f_\theta}$) encodes the each of the inputs into the representation ($X_j$), and Parameter generator ($\boldsymbol{g_\psi}$) obtains a set representation for the class $n$ as $w_i^n, b_i^n$

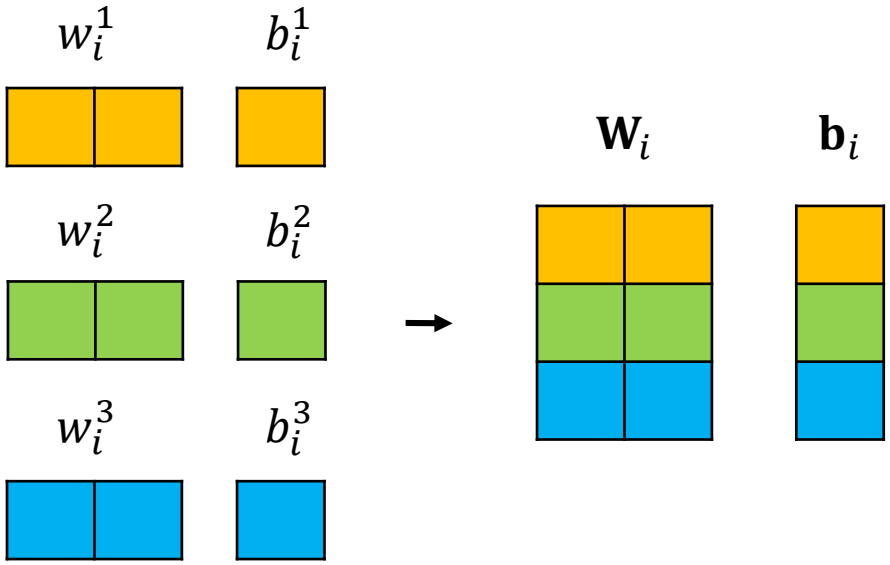


# 2. LEOPARD

## Model

### 2. A Softmax Parameter Generator

3. The softmax classification weights $\mathbf{W}_i \in R_i^{N_i \times l}$ and bias $\mathbf{b}_i \in R_i^{N_i}$ for task $T_i$ are obtained by row-wise concatenation of the per-class weight



# 2. LEOPARD

## Model

### 2. A Softmax Parameter Generator

4. Given the softmax parameters, the prediction for a new data-point $\mathbf{x}^*$ is given as:

$$p(y|\mathrm{x}^*) = \mathrm{softmax}\{\mathbf{W_i} h_{\emptyset}\left(f_{\theta}(\mathrm{x}^*)\right) + \mathbf{b_i}\}$$

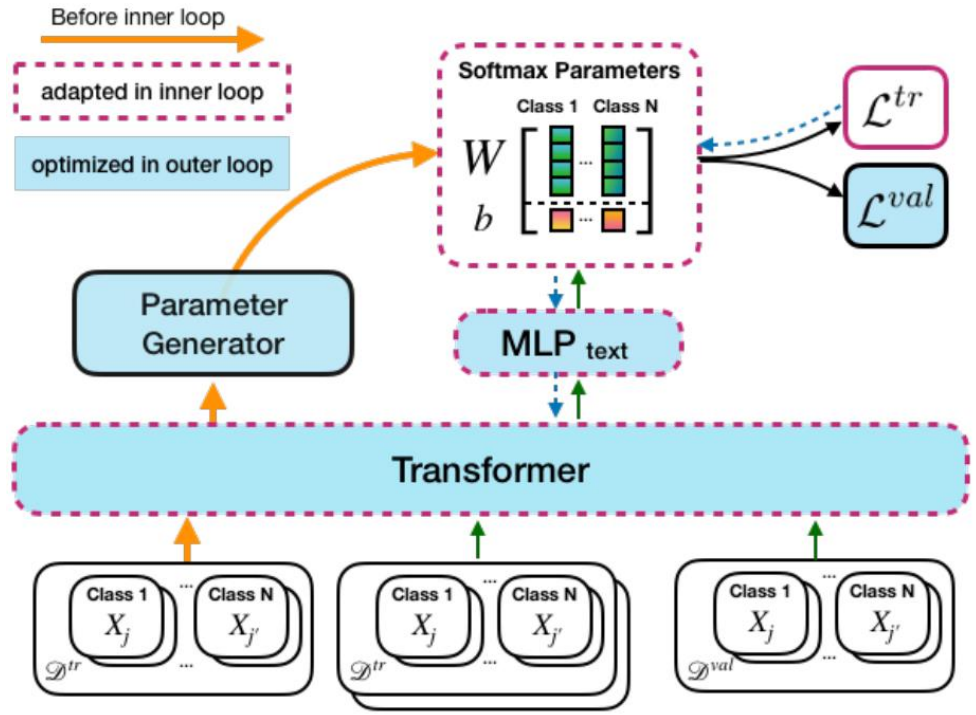
where $h_{\emptyset}(\cdot)$ is another MLP with parameters $\emptyset$ and output dimension $l$ , and the softmax is over the set of classes $N_i$ for the task

# 2. LEOPARD

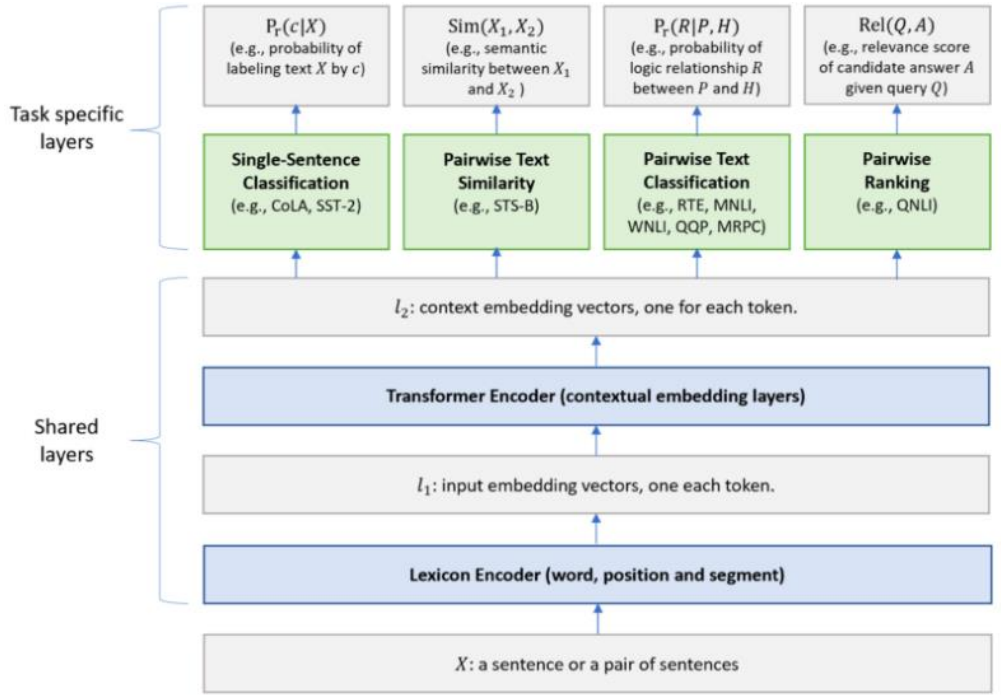
## Model

### 2. A Softmax Parameter Generator

- None of task-specific parameters are introduced, instead the parameter are used to generate a good initial point for softmax parameters across tasks which can be adapted using SGD




LEOPARD




Multi-task BERT

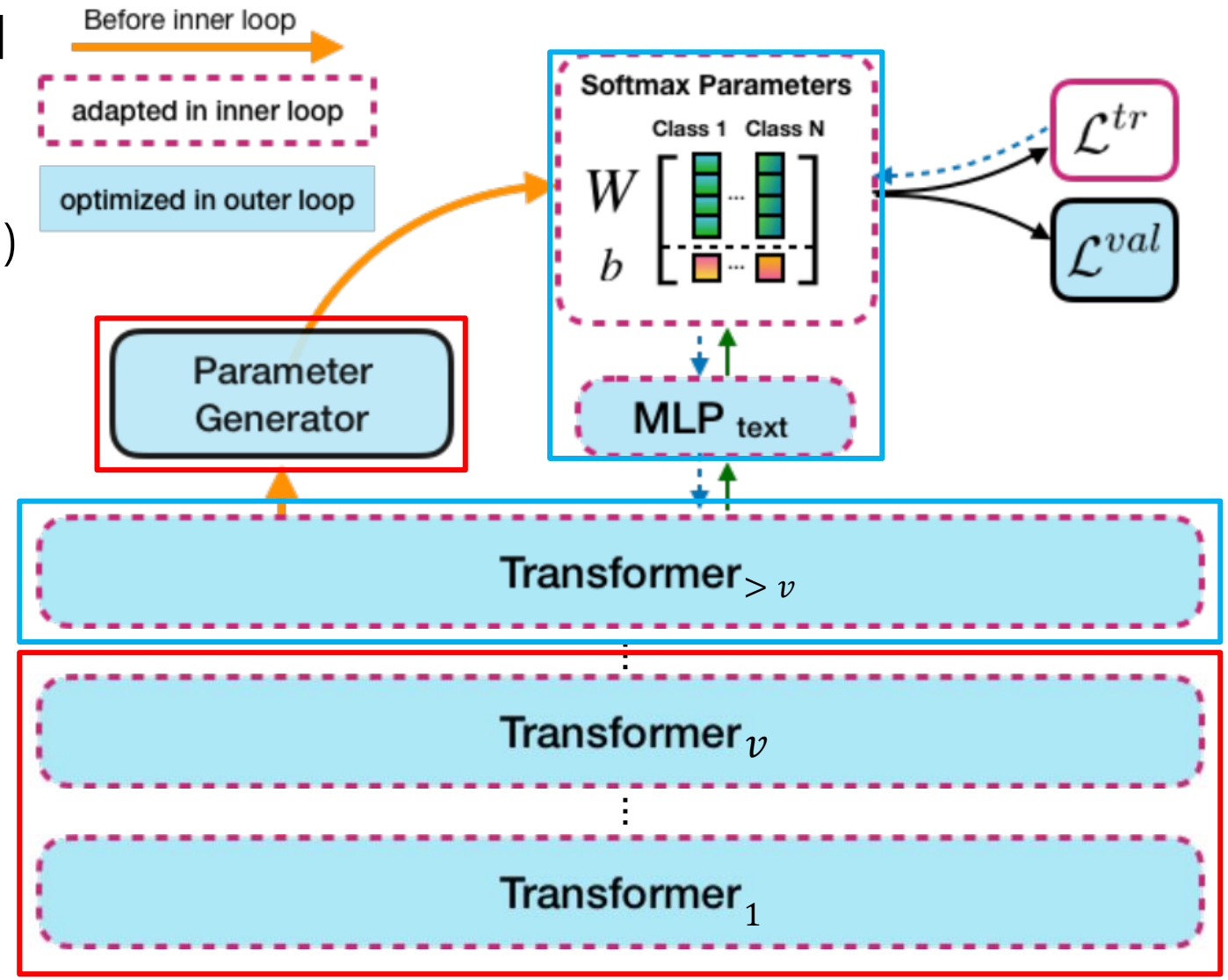
# 2. LEOPARD

## Model

### 3. A MAML-based Adaptation Method

- Task-specific parameters
  - Updated per tasks(inner loop)
  - The higher layers of Transformer($\theta_{>v}$)
  - MLP_text(Text Encoder) ($\emptyset$)
  - Softmax Parameters ($\mathbf{W}_i, \mathbf{b}_i$)

- Task-agnostic parameters
  - Updated per episodes(outer loop)
  - Shared across tasks
  - The lower layers of Transformer($\theta_{\leq v}$)
  - Parameter Generator($\psi$)



# 2. LEOPARD

## Model

**Algorithm 1** LEOPARD

**Require:** set of $M$ training tasks and losses $\{(T_1, L_1), \ldots, (T_M, L_M)\}$, model parameters $\Theta = \{\theta, \psi, \alpha\}$, hyper-parameters $\nu, G, \beta$

Initialize $\theta$ with pre-trained BERT-base;

**while** not converged **do**
  *# sample batch of tasks*
  **for all** $T_i \in T$ **do**
    $\mathcal{D}_i^{tr} \sim T_i$ *# sample a batch of train data*
    $C_i^n \leftarrow \{x_j | y_j = n\}$ *# partition data according to class labels*
    $w_i^n, b_i^n \leftarrow \frac{1}{|C_i^n|} \sum_{x_j \in C_i^n} g_\psi(f_\theta(\mathcal{D}_i^{tr}))$ *# generate softmax parameters*
    $\mathbf{W}_i \leftarrow [w_i^1; \ldots; w_i^{N_i}]$; $\mathbf{b}_i \leftarrow [b_i^1; \ldots; b_i^{N_i}]$
    $\Phi_i^{(0)} \leftarrow \theta_{>\nu} \cup \{\phi, \mathbf{W}_i, \mathbf{b}_i\}$ *# task-specific parameters*
    **for** $s := 0 \ldots G-1$ **do**
      $\mathcal{D}_i^{tr} \sim T_i$ *# sample a batch of train data*
      $\Phi_i^{(s+1)} \leftarrow \Phi_i^{(s)} - \alpha_s \nabla_\Phi \mathcal{L}_i(\{\Theta, \Phi_i\}, \mathcal{D}_i^{tr})$ *# adapt task-specific parameters*
    **end for**
    $\mathcal{D}_i^{val} \sim T_i$ *# sample a batch of validation data*
    $g_i \leftarrow \nabla_\Theta \mathcal{L}_i(\{\Theta, \Phi_i^{(G)}\}, \mathcal{D}_i^{val})$ *# gradient of task-agnostic parameters on validation*
  **end for**
  $\Theta \leftarrow \Theta - \beta \cdot \sum_i g_i$ *# optimize task-agnostic parameters*
**end while**

# 3. Experiments

# 3. Experiments

1. Training Tasks

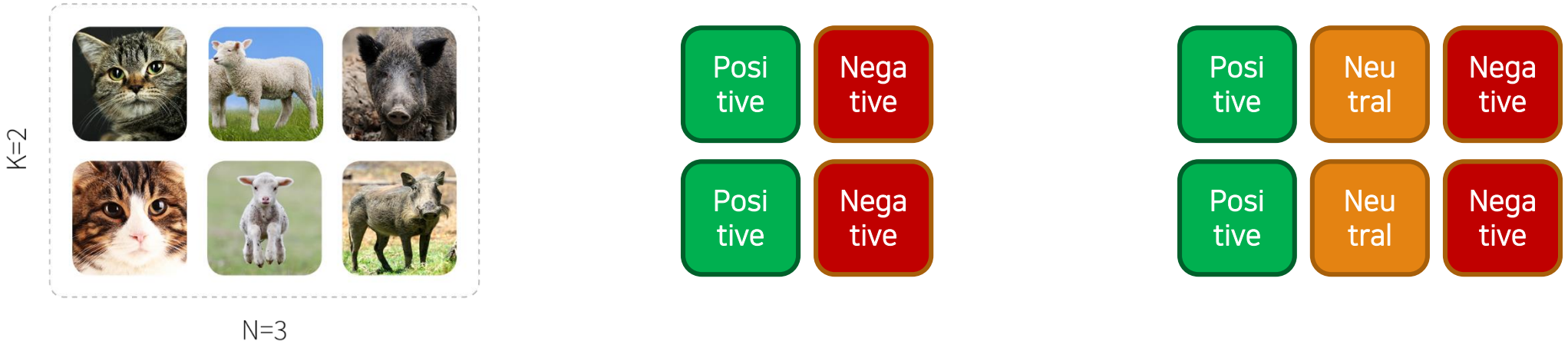
2. Evaluation and Baselines

3. Results
    1. Generalization Beyond Training Tasks
    2. Few-Shot Domain Transfer
    3. Ablation Study

# 3. Experiments

## Training Tasks

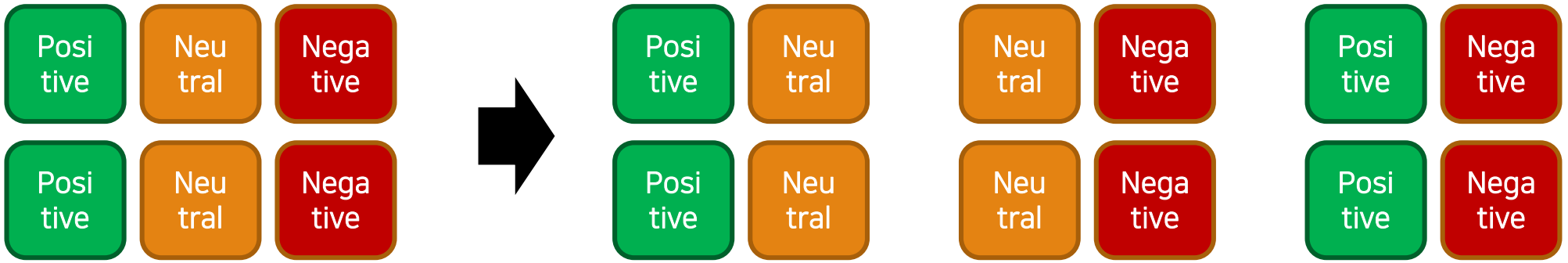
- GLUE Benchmark tasks(Wang et al., 2018)
  - MNLI(m/mm), SST-2, QNLI, QQP, MRPC, RTE, SNLI
  - WNLI, STS-B datasets are excluded
    - WNLI : it's training data is small
    - STS-B : it is a regression task



# 3. Experiments

## Training Tasks

- Data Augmentation
  - For tasks with more than 2 labels,
    they classify between every pair of labels



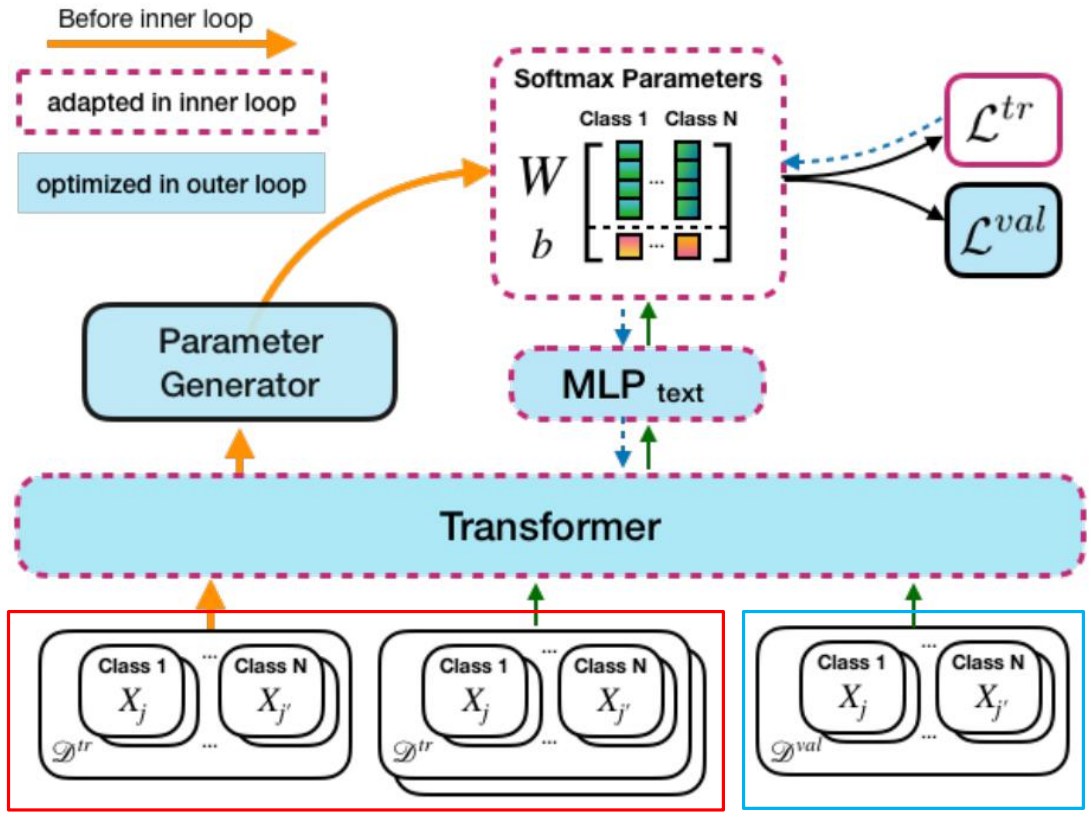
# 3. Experiments

## Evaluation and Baselines

- Training and Evaluation Process

1. The models are trained on the set of training tasks

   The hyper-parameters are tuned with the set of validation tasks



# 3. Experiments

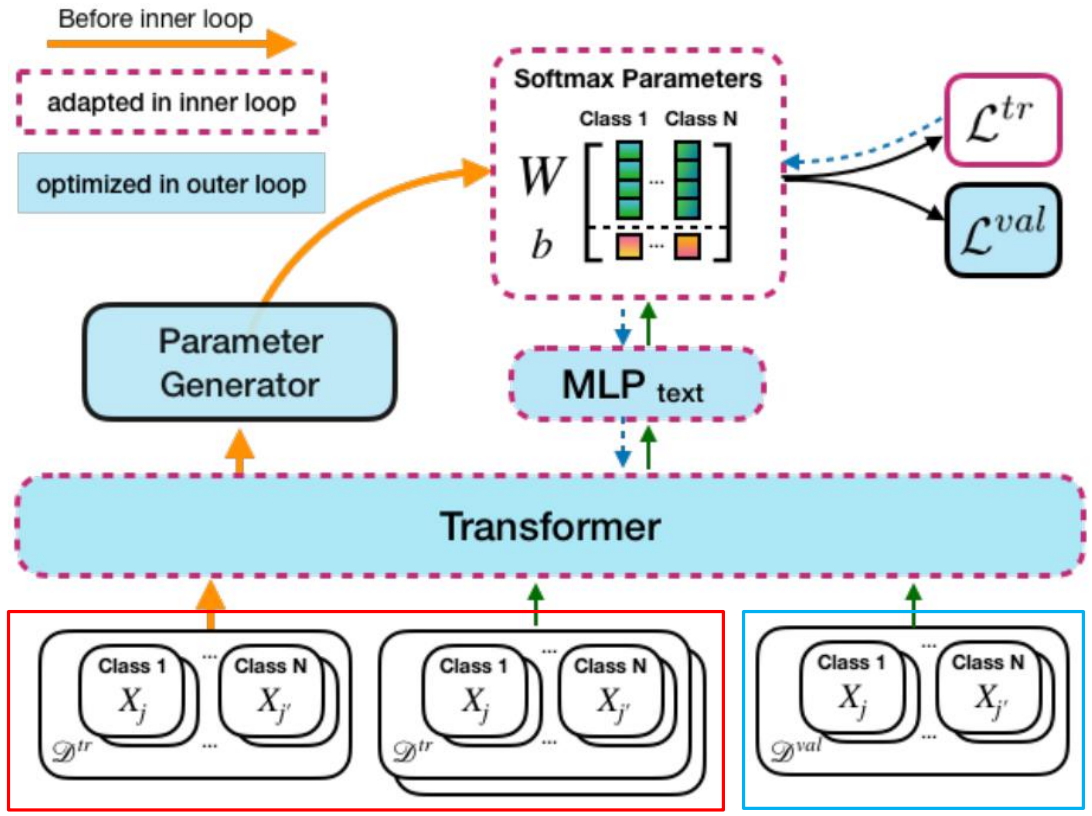
## Evaluation and Baselines

- Training and Evaluation Process

2. The models are fine-tuned with $k$ training examples per label for a target test task ($k \in \{4, 8, 16\}$)

   - For the fine-turning step, **tuning the hyper-parameters** for **all baselines** on a held out validation task

   - SciTail, a scientific NLI tasks, and electronics domain of Amazon sentiment classification task



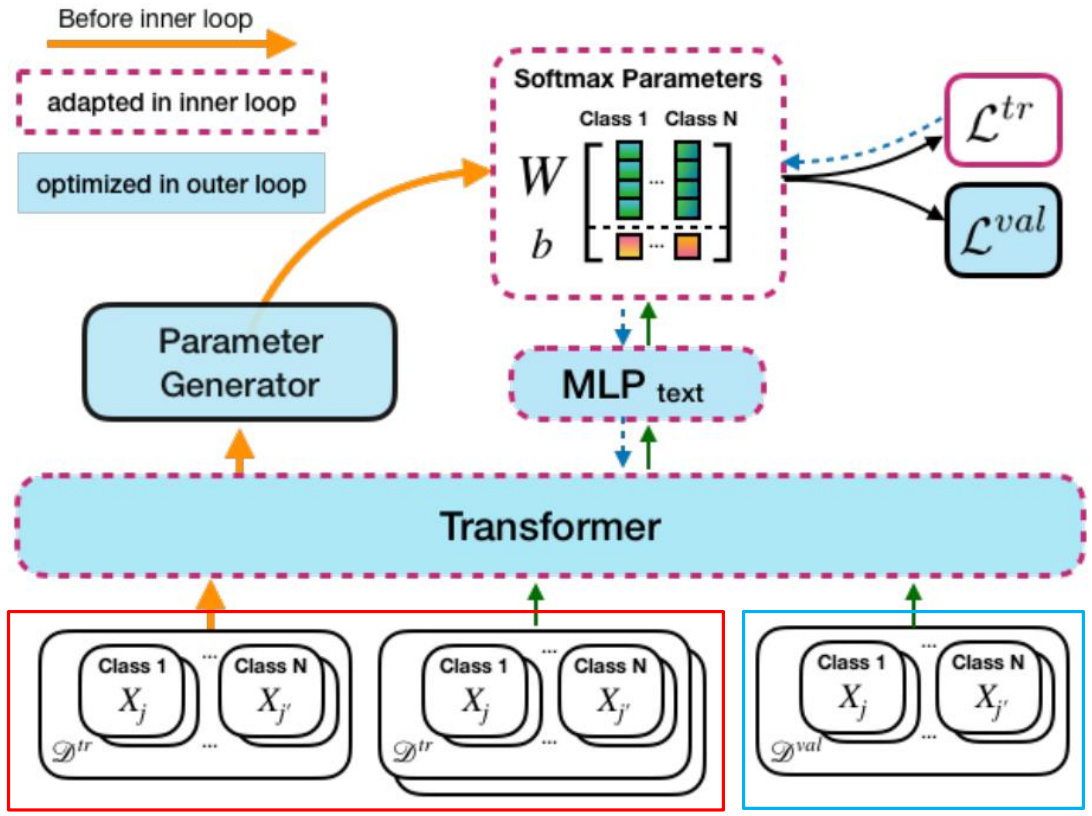
# 3. Experiments

## Evaluation and Baselines

- Training and Evaluation Process

2. The models are fine-tuned with $k$ training examples per label for a target test task ($k \in \{4, 8, 16\}$)

   - For the fine-turning step, **only tuning the number of epochs** for **LEOPARD** on a held out validation task

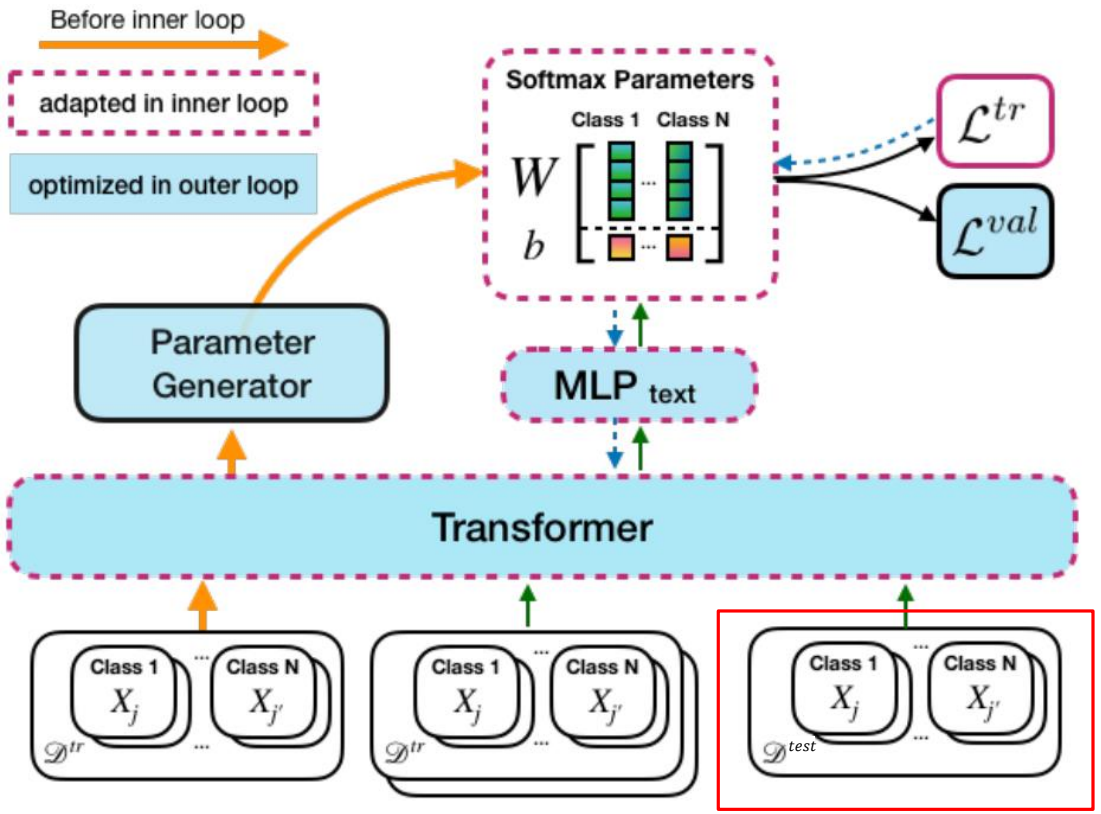


# 3. Experiments

## Evaluation and Baselines

- Training and Evaluation Process

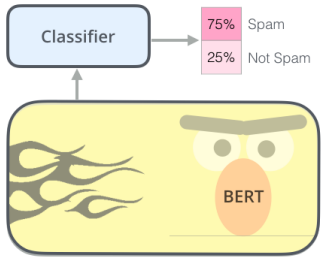
3. The fine-tuned models are evaluated on the *entire test-set* for the task.



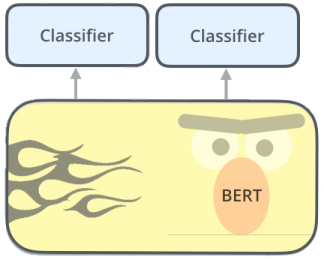
# 3. Experiments

## Evaluation and Baselines

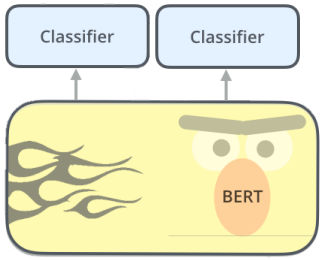
- Baselines
  - Transfer learning baselines
    - $\text{BERT}_{base}$
    - Multi-task BERT(MT-BERT)
    - $\text{MT-BERT}_{softmax}$
  - Meta-learning baselines
    - Prototypical BERT(Proto-BERT)



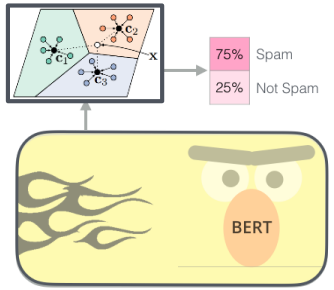

$\text{BERT}_{\text{base}}$




MT-BERT




$\text{MT-BERT}_{\text{softmax}}$




Proto-BERT

# 3. Experiments

## Results

- **Generalization Beyond Training Tasks**
  - Performance on **new tasks** not seen at training time
  - Datasets

# 3. Experiments

## Results

- Generalization Beyond Training Tasks
  - Robust to varying number of labels across tasks and across different text domains
  - It adapts quicker to new text domains than MT-BERT
    - Relative gain in accuracy
      - 14.45%, 10.75%, 10.9% $k = 4, 8, 16$ respectively

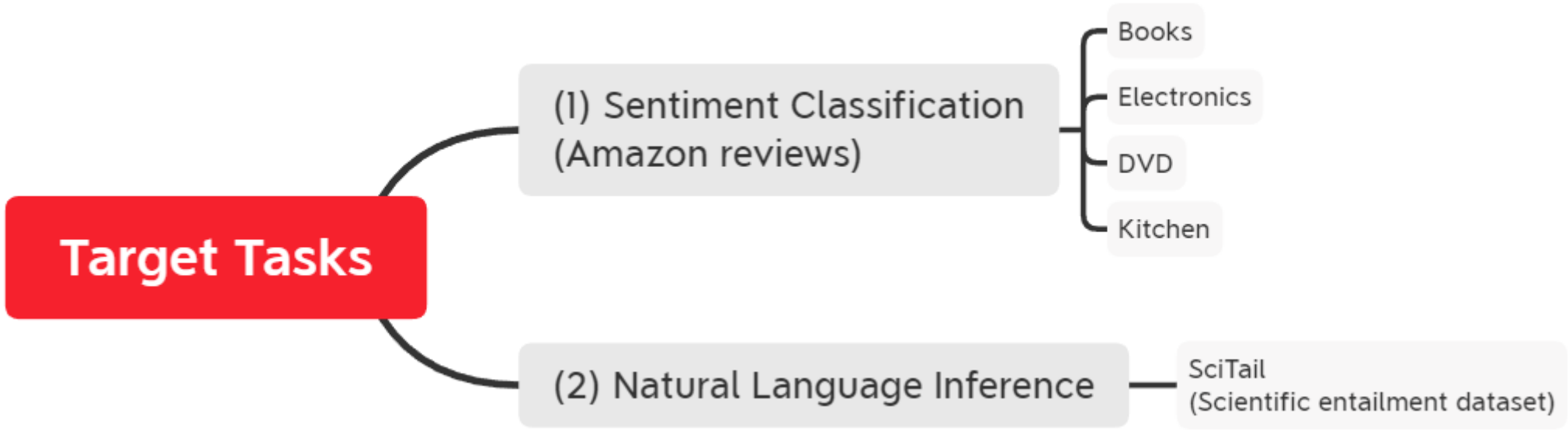
| | $N$ | $k$ | $\text{BERT}_{\text{base}}$ | $\text{MT-BERT}_{\text{softmax}}$ | MT-BERT | Proto-BERT | LEOPARD |
|---|---|---|---|---|---|---|---|
| **Entity Typing** | | | | | | | |
| CoNLL | 4 | 4 | 50.44 ± 08.57 | 52.28 ± 4.06 | **55.63** ± 4.99 | 32.23 ± 5.10 | 54.16 ± 6.32 |
| | | 8 | 50.06 ± 11.30 | 65.34 ± 7.12 | 58.32 ± 3.77 | 34.49 ± 5.15 | **67.38** ± 4.33 |
| | | 16 | 74.47 ± 03.10 | 71.67 ± 3.03 | 71.29 ± 3.30 | 33.75 ± 6.05 | **76.37** ± 3.08 |
| MITR | 8 | 4 | 49.37 ± 4.28 | 45.52 ± 5.90 | **50.49** ± 4.40 | 17.36 ± 2.75 | 49.84 ± 3.31 |
| | | 8 | 49.38 ± 7.76 | 58.19 ± 2.65 | 58.01 ± 3.54 | 18.70 ± 2.38 | **62.99** ± 3.28 |
| | | 16 | 69.24 ± 3.68 | 66.09 ± 2.24 | 66.16 ± 3.46 | 16.41 ± 1.87 | **70.44** ± 2.89 |
| **Text Classification** | | | | | | | |
| Airline | 3 | 4 | 42.76 ± 13.50 | 43.73 ± 7.86 | 46.29 ± 12.26 | 40.27 ± 8.19 | **54.95** ± 11.81 |
| | | 8 | 38.00 ± 17.06 | 52.39 ± 3.97 | 49.81 ± 10.86 | 51.16 ± 7.60 | **61.44** ± 03.90 |
| | | 16 | 58.01 ± 08.23 | 58.79 ± 2.97 | 57.25 ± 09.90 | 48.73 ± 6.79 | **62.15** ± 05.56 |
| Disaster | 2 | 4 | **55.73** ± 10.29 | 52.87 ± 6.16 | 50.61 ± 8.33 | 50.87 ± 1.12 | 51.45 ± 4.25 |
| | | 8 | **56.31** ± 09.57 | 56.08 ± 7.48 | 54.93 ± 7.88 | 51.30 ± 2.30 | 55.96 ± 3.58 |
| | | 16 | 64.52 ± 08.93 | **65.83** ± 4.19 | 60.70 ± 6.05 | 52.76 ± 2.92 | 61.32 ± 2.83 |
| Emotion | 13 | 4 | 09.20 ± 3.22 | 09.41 ± 2.10 | 09.84 ± 2.14 | 09.18 ± 3.14 | **11.71** ± 2.16 |
| | | 8 | 08.21 ± 2.12 | 11.61 ± 2.34 | 11.21 ± 2.11 | 11.18 ± 2.95 | **12.90** ± 1.63 |
| | | 16 | 13.43 ± 2.51 | **13.82** ± 2.02 | 12.75 ± 2.04 | 12.32 ± 3.73 | 13.38 ± 2.20 |
| Political Bias | 2 | 4 | 54.57 ± 5.02 | 54.32 ± 3.90 | 54.66 ± 3.74 | 56.33 ± 4.37 | **60.49** ± 6.66 |
| | | 8 | 56.15 ± 3.75 | 57.36 ± 4.32 | 54.79 ± 4.19 | 58.87 ± 3.79 | **61.74** ± 6.73 |
| | | 16 | 60.96 ± 4.25 | 59.24 ± 4.25 | 60.30 ± 3.26 | 57.01 ± 4.44 | **65.08** ± 2.14 |
| Political Audience | 2 | 4 | **51.02** ± 1.23 | 50.45 ± 1.01 | 50.96 ± 1.72 | 49.55 ± 1.98 | 50.84 ± 1.33 |
| | | 8 | 50.87 ± 1.88 | 51.63 ± 1.81 | 50.36 ± 1.53 | 50.62 ± 1.35 | **51.74** ± 1.37 |
| | | 16 | **53.09** ± 1.93 | 52.41 ± 1.25 | 51.24 ± 2.18 | 50.92 ± 1.56 | 51.90 ± 1.43 |
| Political Message | 9 | 4 | 15.64 ± 2.73 | 13.71 ± 1.10 | 14.49 ± 1.75 | 14.22 ± 1.25 | **15.69** ± 1.57 |
| | | 8 | 13.38 ± 1.74 | 14.33 ± 1.32 | 15.24 ± 2.81 | 15.67 ± 1.96 | **18.02** ± 2.32 |
| | | 16 | **20.67** ± 3.89 | 18.11 ± 1.48 | 19.20 ± 2.20 | 16.49 ± 1.96 | 18.07 ± 2.41 |
| Rating Books | 3 | 4 | 39.42 ± 07.22 | 44.82 ± 9.00 | 38.97 ± 13.27 | 48.44 ± 7.43 | **54.92** ± 6.18 |
| | | 8 | 39.55 ± 10.01 | 51.14 ± 6.78 | 46.77 ± 14.12 | 52.13 ± 4.79 | **59.16** ± 4.13 |
| | | 16 | 43.08 ± 11.78 | 54.61 ± 6.79 | 51.68 ± 11.27 | 57.28 ± 4.57 | **61.02** ± 4.19 |
| Rating DVD | 3 | 4 | 32.22 ± 08.72 | 45.94 ± 7.48 | 41.23 ± 10.98 | 47.73 ± 6.20 | **49.76** ± 9.80 |
| | | 8 | 36.35 ± 12.50 | 46.23 ± 6.03 | 45.24 ± 9.76 | 47.11 ± 4.00 | **53.28** ± 4.66 |
| | | 16 | 42.79 ± 10.18 | 49.23 ± 6.68 | 45.19 ± 11.56 | 48.39 ± 3.74 | **53.52** ± 4.77 |
| Rating Electronics | 3 | 4 | 39.27 ± 10.15 | 39.89 ± 5.83 | 41.20 ± 10.69 | 37.40 ± 3.72 | **51.71** ± 7.20 |
| | | 8 | 28.74 ± 08.22 | 46.53 ± 5.44 | 45.41 ± 09.49 | 43.64 ± 7.31 | **54.78** ± 6.48 |
| | | 16 | 45.48 ± 06.13 | 48.71 ± 6.16 | 47.29 ± 10.55 | 44.83 ± 5.96 | **58.69** ± 2.41 |
| Rating Kitchen | 3 | 4 | 34.76 ± 11.20 | 40.41 ± 5.33 | 36.77 ± 10.62 | 44.72 ± 9.13 | **50.21** ± 09.63 |
| | | 8 | 34.49 ± 08.72 | 48.35 ± 7.87 | 47.98 ± 09.73 | 46.03 ± 8.57 | **53.72** ± 10.31 |
| | | 16 | 47.94 ± 08.28 | 52.94 ± 7.14 | 53.79 ± 09.47 | 49.85 ± 9.31 | **57.00** ± 08.69 |
| Overall Average | | 4 | 38.06 | 40.04 | 40.05 | 36.13 | **45.84** |
| | | 8 | 36.83 | 45.73 | 43.92 | 39.05 | **50.65** |
| | | 16 | 48.10 | 49.60 | 48.74 | 39.63 | **55.02** |

# 3. Experiments

## Results

- Few-Shot Domain Transfer
  - Performance on **new domains** of tasks seen at training time

  - Datasets



# 3. Experiments

## Results

- Few-Shot Domain Transfer
  - Perform better than the baselines on all domains sentiment classification
  - On SciTail, MT-BERT perform better, potentially because training consisted of many related NLI datasets

| | $k$ | $\text{BERT}_{\text{base}}$ | $\text{MT-BERT}_{\text{softmax}}$ | MT-BERT | $\text{MT-BERT}_{\text{reuse}}$ | Proto-BERT | LEOPARD |
|---|---|---|---|---|---|---|---|
| **Natural Language Inference** | | | | | | | |
| Scitail | 4 | 58.53 ± 09.74 | 74.35 ± 5.86 | 63.97 ± 14.36 | **76.65 ± 2.45** | 76.27 ± 4.26 | 69.50 ± 9.56 |
| | 8 | 57.93 ± 10.70 | **79.11 ± 3.11** | 68.24 ± 10.33 | 76.86 ± 2.09 | 78.27 ± 0.98 | 75.00 ± 2.42 |
| | 16 | 65.66 ± 06.82 | **79.60 ± 2.31** | 75.35 ± 04.80 | 79.53 ± 2.17 | 78.59 ± 0.48 | 77.03 ± 1.82 |
| **Amazon Review Sentiment Classification** | | | | | | | |
| Books | 4 | 54.81 ± 3.75 | 68.69 ± 5.21 | 64.93 ± 8.65 | 74.79 ± 6.91 | 73.15 ± 5.85 | **82.54 ± 1.33** |
| | 8 | 53.54 ± 5.17 | 74.86 ± 2.17 | 67.38 ± 9.78 | 78.21 ± 3.49 | 75.46 ± 6.87 | **83.03 ± 1.28** |
| | 16 | 65.56 ± 4.12 | 74.88 ± 4.34 | 69.65 ± 8.94 | 78.87 ± 3.32 | 77.26 ± 3.27 | **83.33 ± 0.79** |
| Kitchen | 4 | 56.93 ± 7.10 | 63.07 ± 7.80 | 60.53 ± 9.25 | 75.40 ± 6.27 | 62.71 ± 9.53 | **78.35 ± 18.36** |
| | 8 | 57.13 ± 6.60 | 68.38 ± 4.47 | 69.66 ± 8.05 | 75.13 ± 7.22 | 70.19 ± 6.42 | **84.88 ± 01.12** |
| | 16 | 68.88 ± 3.39 | 75.17 ± 4.57 | 77.37 ± 6.74 | 80.88 ± 1.60 | 71.83 ± 5.94 | **85.27 ± 01.31** |

Table 2: Domain transfer evaluation (accuracy) on NLI and Sentiment classification datasets.

# 3. Experiments

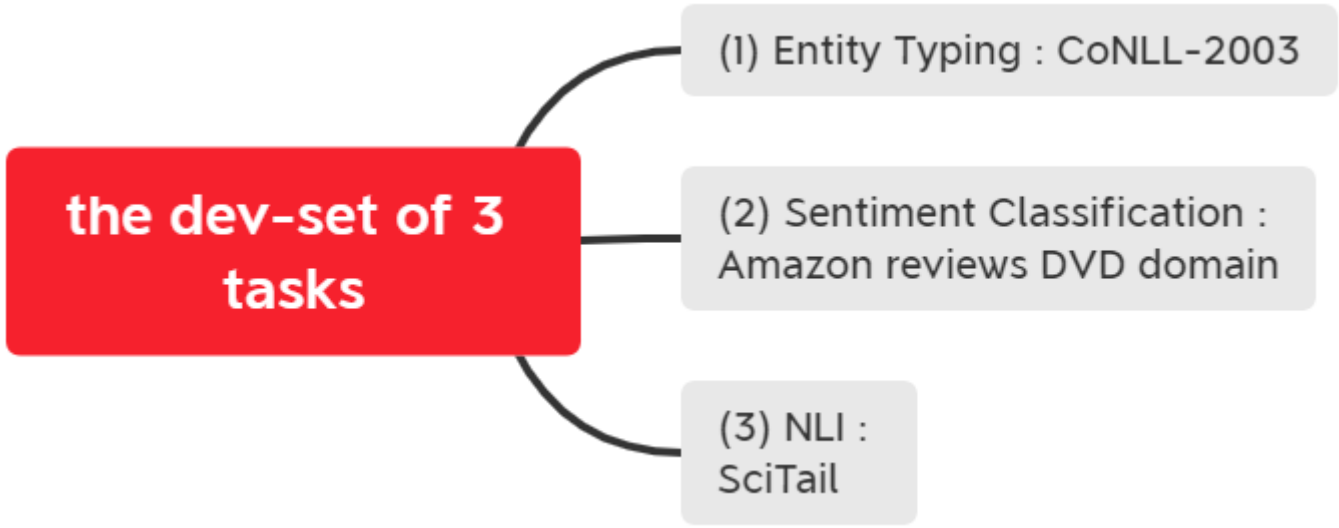
## Results

- Ablation Study
  1. Importance of softmax parameters
  2. Parameter efficiency
  3. Importance of training tasks

# 3. Experiments

## Results

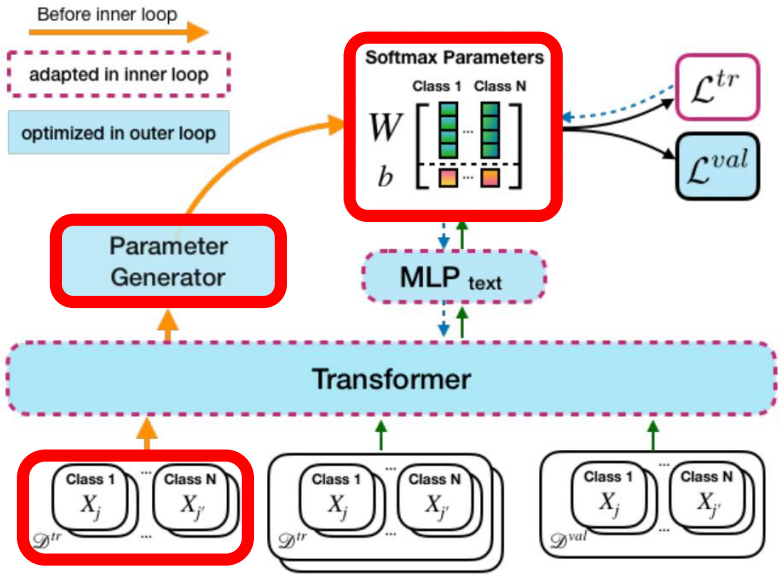
- Ablation Study
  - Datasets



# 3. Experiments

## Results

- Ablation Study
  1. Importance of softmax parameters
     - To study how the softmax generator works, it is replaced with softmax weight and bias with zero initialization for each task
       → LEOPARD-ZERO
     - It performs worse on new tasks(Entity Typing)




| $k$ | Model | Entity Typing | Sentiment Classification | NLI |
|---|---|---|---|---|
| 16 | LEOPARD $_{10}$ | 37.62 $\pm$ 7.37 | 58.10 $\pm$ 5.40 | 78.53 $\pm$ 1.55 |
| | LEOPARD $_{5}$ | 62.49 $\pm$ 4.23 | 71.50 $\pm$ 5.93 | 73.27 $\pm$ 2.63 |
| | LEOPARD | 69.00 $\pm$ 4.76 | 76.65 $\pm$ 2.47 | 76.10 $\pm$ 2.21 |
| | LEOPARD-ZERO | 44.79 $\pm$ 9.34 | 74.45 $\pm$ 3.34 | 74.36 $\pm$ 6.67 |

# 3. Experiments

## Results

- Ablation Study
  2. Parameter efficiency
     - 3 variants of LEOPARD with parameter efficient training
     - LEOPARD$_v$ : It does not adapt layers 0 to $v$ in the inner-loop of meta-training

       *NOTE : Even for $v \neq 0$, the parameters are still optimized in the outer-loop

| $k$ | Model | Entity Typing | Sentiment Classification | NLI |
|---|---|---|---|---|
| 16 | LEOPARD $_{10}$ | 37.62 $\pm$ 7.37 | 58.10 $\pm$ 5.40 | 78.53 $\pm$ 1.55 |
| | LEOPARD $_5$ | 62.49 $\pm$ 4.23 | 71.50 $\pm$ 5.93 | 73.27 $\pm$ 2.63 |
| | LEOPARD | 69.00 $\pm$ 4.76 | 76.65 $\pm$ 2.47 | 76.10 $\pm$ 2.21 |
| | LEOPARD-ZERO | 44.79 $\pm$ 9.34 | 74.45 $\pm$ 3.34 | 74.36 $\pm$ 6.67 |

# 3. Experiments

## Results

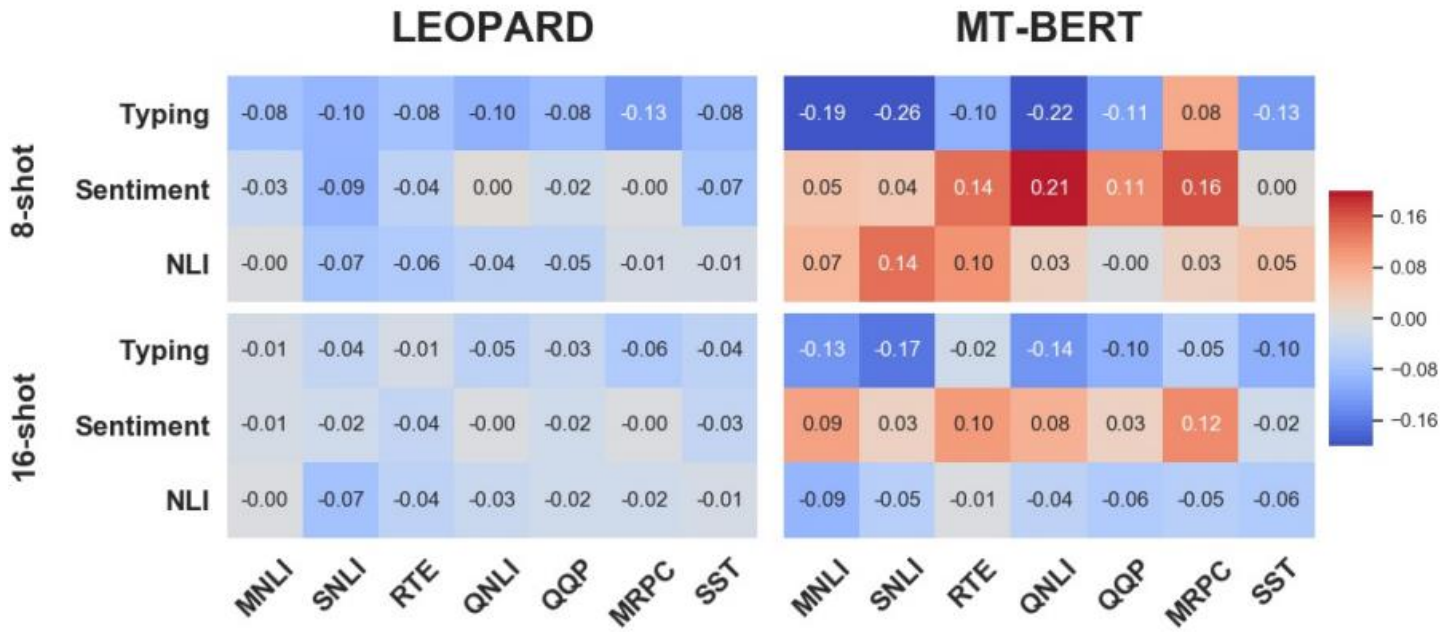
- Ablation Study
  2. Parameter efficiency
     - For all tasks (except NLI) adapting all parameter is better
     - On SciTail (NLI) adapting fewer parameters is better for small k

| $k$ | Model | Entity Typing | Sentiment Classification | NLI |
|---|---|---|---|---|
| 16 | LEOPARD $_{10}$ | 37.62 $\pm$ 7.37 | 58.10 $\pm$ 5.40 | 78.53 $\pm$ 1.55 |
| | LEOPARD $_{5}$ | 62.49 $\pm$ 4.23 | 71.50 $\pm$ 5.93 | 73.27 $\pm$ 2.63 |
| | LEOPARD | 69.00 $\pm$ 4.76 | 76.65 $\pm$ 2.47 | 76.10 $\pm$ 2.21 |
| | LEOPARD-ZERO | 44.79 $\pm$ 9.34 | 74.45 $\pm$ 3.34 | 74.36 $\pm$ 6.67 |

# 3. Experiments

## Results

- Ablation Study
  3. Importance of training tasks
     - How target-task performance of MT-BERT and LEOPARD is dependent on tasks used for training
     - LEOPARDS's performance is more consistent



# 4. Conclusions

# 4. Conclusions

- Learning general linguistic intelligence has been a long-term goal of NLP
- LEOPARD learns more general purpose parameters that better prime the model to solve completely new tasks with few examples
- But performance with few-examples sill lags behind human-level performance

dongguk UNIVERSITY

# Q & A