ConEntail: An Entailment-based Framework for Universal Zero and Few Shot Classification with Supervised Contrastive Pretraining

Ranran Haoran Zhang[1], **Aysa Xuemo Fan**[2], **Rui Zhang**[1]
[1] Penn State University
[2] University of Illinois at Urbana-Champaign
`{hzz5361, rmz5227}@psu.edu`
`xuemof2@illinois.edu`

Abstract

A universal classification model aims to generalize to diverse classification tasks in both zero and few shot settings. A promising way toward universal classification is to cast heterogeneous data formats into a dataset-agnostic "meta-task" (e.g., textual entailment, question answering) then pretrain a model on the combined meta dataset. The existing work is either pretrained on specific subsets of classification tasks, or pretrained on both classification and generation data but the model could not fulfill its potential in universality and reliability. These also leave a massive amount of annotated data under-exploited. To fill these gaps, we propose ConEntail, a new framework for universal zero and few shot classification with supervised contrastive pretraining. Our unified meta-task for classification is based on nested entailment. It can be interpreted as "Does sentence a entails [sentence b entails label c]". This formulation enables us to make better use of 57 annotated classification datasets for supervised contrastive pretraining and universal evaluation. In this way, ConEntail helps the model (1) absorb knowledge from different datasets, and (2) gain consistent performance gain with more pretraining data. In experiments, we compare our model with discriminative and generative models pretrained on the same dataset. The results confirm that our framework effectively exploits existing annotated data and outperforms baselines in both zero (9.4% average improvement) and few shot settings (3.5% average improvement). Our code is available at `https://github.com/psunlpgroup/ConEntail`.

1 Introduction

It has been a long-standing effort to solve various text classification tasks by training one universal model (Kumar et al., 2016). With an ideal universal classification model, we can expect extreme generalization with few or zero annotation in new domains/tasks/datasets. To this end, researchers reformulate heterogeneous task definitions into a unified format of a meta-task in natural language (Yin et al., 2020; Khashabi et al., 2020a). Solving the meta-task is equivalent to solving the isolated tasks, thus the meta-task paves the way of supplementing unsupervised pretrained Language Models (PLM) with additional supervised pretraining, to further absorb knowledge from heterogeneous labeled data.

The success of universal classification models hinges on how well a strong PLM understands natural language meta-task. The meta-task format depends on two underlying PLM types: (a) **discriminator** uses Encoder PLMs and treats all classification tasks as binary entailment classification problem (Yin et al., 2019, 2020; Xia et al., 2021; Wang et al., 2021). However, they only pretrain models on Natural Language Inference datasets, whose knowledge is not comprehensive comparing all classification tasks (Ma et al., 2021). (b) **generator** uses Encoder-Decoder PLMs and treats all tasks as text generation problem (Gao et al., 2020; Raffel et al., 2020; Sanh et al., 2021; Aribandi et al., 2021; Ye et al., 2021a; Bragg et al., 2021; Du et al., 2021; Schick and Schütze, 2021a,b). Thus they are compatible with both classification tasks and generation tasks. However, the generator nature implies that the predicted texts may not match any possible labels, thus more likely to fail on classification tasks (Sanh et al., 2021).

Based on our observations and experiments, we argue that the discriminators have more potential in universal classification, and propose a new discriminator framework, ConEntail, that can make better use of existing annotated datasets. Concretely, we reformulate the unified meta-task as a nested entailment: "Does sentence q entails [sentence p entails label h]". Take Fig. 1 as an example, the query "We had a great breakfast at the waffle shop!" entails the same label as the premise "I bought this for myself a short time ago and I love it. An excellent

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1941–1953
May 2-6, 2023 ©2023 Association for Computational Linguistics

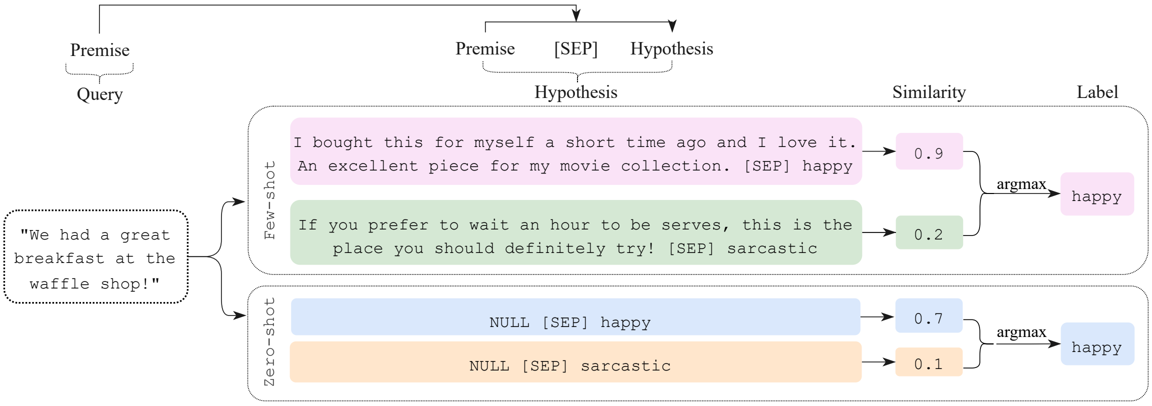

Premise
Query
Premise
[SEP]
Hypothesis
Hypothesis
Similarity
Label
Few-shot
I bought this for myself a short time ago and I love it. An excellent piece for my movie collection. [SEP] happy
If you prefer to wait an hour to be serves, this is the place you should definitely try! [SEP] sarcastic
0.9
0.2
argmax
happy
"We had a great breakfast at the waffle shop!"
Zero-shot
NULL [SEP] happy
NULL [SEP] sarcastic
0.7
0.1
argmax
happy

Figure 1: The overview of the CONENTAIL framework. By casting the classification as a nested entailment task, the model performs classification by telling if a query sentence q entails [premise example p entails hypothesis label h]. In a few-shot setting, the premise is an example sentence; in a zero-shot setting, the premise is a "NULL" placeholder.

piece for my movie collection.", so it yields a high similarity score of 0.9, in this case, it is higher than any other similarities, thus, the prediction would be "happy". For zero-shot generalization, as no annotated sentences are available, we replace the premise p with "NULL" in evaluation. We randomly nullify a small ratio of p in the supervised pretraining for training-evaluation consistency. The supervised contrastive learning framework pulls sentences embeddings with the same label together and pushes those with different labels apart, thus capturing more similarities/dissimilarities from labeled data, and benefiting few/zero-shot learning.

In experiments, we collect 56 classification datasets from Crossfit (Ye et al., 2021a), together with their templates, to formulate a large supervised pretraining dataset. We reproduce EFL (Wang et al., 2021), Unifew (Bragg et al., 2021) and Crossfit (Ye et al., 2021a) in the same setting and control influences of PLM supervised pretraining data, then conduct fair comparison with our proposed CONENTAIL. The experiments show that generators (Unifew and Crossfit) do not fit the classification task well and thus significantly under-perform the random guess in zero-shot evaluation; standard discriminators (EFL) under-exploit supervised pretraining datasets and thus do not gain consistent improvement as pretraining data scale up, while CONENTAIL makes the best use of the supervised pretraining data and keep consistent performances. Our model outperforms baselines in both zero (9.4% average improvement) and few shot settings (3.5% average improvement).

Our contributions are the following:

- We propose a novel universal classification framework based on nested entailment, CONENTAIL, that can be used in both zero and few shot settings. It makes better use of supervised pretraining datasets and consistently improves performances with increases of the pretraining scale.
- We design systematic experiments to compare generative and discriminative models, and more importantly, we give in-depth analysis to reveal their attributes in universal classification task.
- Our model reliably outperforms the baseline models in all kinds of pretraining size, finetuning size, and covers a wide range of tasks.

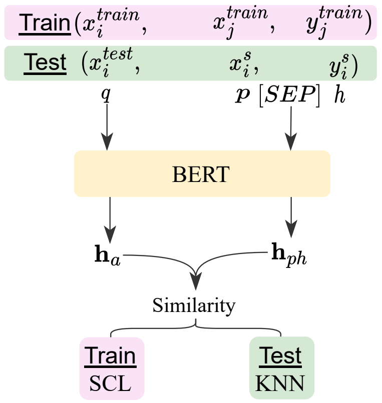
2 Related Work

Universal Meta Task Casting heterogeneous datasets into a unified meta-task allows researchers to train one model to solve all tasks. There are two types of meta-task formats, generation (Schick and Schütze, 2021a,b; Gao et al., 2020; Ye et al., 2021a; Bragg et al., 2021; Khashabi et al., 2020a) and discrimination (Yin et al., 2019, 2020; Xia et al., 2021; Wang et al., 2021). The generators formulate meta-task as a text-to-text generation problem. Although their supervised pretraining usually involves both classification and generation tasks, as the text outputs are open-ended, the model predictions may fall out of all possible labels. The discriminators formulate meta-task as an entailment classification problem, and usually use Natural Language Inference datasets for supervised pretraining. We extend discriminator pretraining to more classification datasets and propose a nested entailment meta-task to enable a more efficient supervised pretraining method.

Supervised Pretraining Supervised pretraining originates from explicit multitask learning (Caruana, 1997) which combines different task knowledge into shared representations. Phang et al. (2018) found that supplementing PLMs with supervised pretraining between unsupervised pretraining and downstream finetuning can significantly boost the performance and few-shot generalization. The discriminator models including UFO-Entail (Yin et al., 2020) and EFL (Wang et al., 2021) are trained on MNLI (Williams et al., 2018) in a supervised fashion, but they do not combine different sources of datasets. Furthermore, T0 (Sanh et al., 2021) and ExT5 (Aribandi et al., 2021) extends T5 (Raffel et al., 2020) by using 107 and 171 datasets for supervised pretraining and conduct zero-shot evaluation. FLEX (Bragg et al., 2021) and Crossfit (Ye et al., 2021a) extends the supervised pretraining evaluation to few-shot learning.

The supervised pretraining strategies from these works vary in pretraining datasets and hyperparameters, but they mostly follow their underlying language model tasks, such as Next Sentence Prediction or Text Generation. We argue that applying the unsupervised pretraining strategy to supervised pretraining is an underuse of the labeled data, and propose a supervised contrastive learning method on PLMs for better zero/few-shot generalization.

Contrastive Learning for NLP Contrastive learning aims to create embeddings such that similar examples are close while dissimilar examples are far away (Chopra et al., 2005). While most works use self-supervised contrastive learning (Shen et al., 2020; Fang et al., 2020; You et al., 2021; Ye et al., 2021b), only a few adopt supervised contrastive learning. CLIP (Radford et al., 2021) uses labeled images and captions as supervision signal. SimCSE (Gao et al., 2021) and SBERT (Reimers and Gurevych, 2019) use labeled sentence pairs from NLI to construct positive and negative examples. However, their contrastive data creations are limited to specific types of data, and thus can be hardly extended to universal classification. We reformulate all NLP classification tasks into a unified contrastive meta-task and use Supervised Contrastive Loss (Khosla et al., 2020) to train on heterogeneous labeled data during supervised pretraining.



Train $(x_i^{train},\ x_j^{train},\ y_j^{train})$
Test $(x_i^{test},\ x_i^{s},\ y_i^{s})$
q
p $[SEP]$ h
BERT
$\mathbf{h}_a$
$\mathbf{h}_{ph}$
Similarity
Train
SCL
Test
KNN

Figure 2: During supervised pertaining, the ConEntail model is optimized with pairwise contrastive learning loss SCL. Testing utilizes the K-Nearest Neighbor predictor to rank pairwise similarities between the query and premise-hypothesis pairs for retrieval of the most likely label. Zero-shot training/testing occurs when the premise example is represented by a "NULL" token."

3 Method

3.1 Universal Classification

Universal classification task aims to build a universal predictor that generalize to new domain/task/-dataset based on only a few or zero newly annotated examples. In order for models to understand a new area, any available resources should be considered for learning, including PLMs trained on large-scale unsupervised data and heterogeneous supervised classification datasets in the NLP community. To leverage heterogeneous datasets, the disparate input-output formats need to be reformulated to a unified PLM comprehensible format, i.e., “meta task”, through either human-curated or machine-generated templates. Then a universal model on the combined meta dataset is trained, which applies universal predictors to new areas. Because the meta task format is compatible with every task, we can cast target tasks into the same format, in this way solving the meta task is equivalent to solving tasks in a new area.

3.2 ConEntail: Nested Entailment

In this paper, we introduce a supervised contrastive pretraining paradigm that makes better use of supervised pretraining. The overview is shown in Fig. 2. Our ConEntail model takes 3 inputs:

$$f\colon \mathcal{Q}, \mathcal{P}, \mathcal{H} \to \{0, 1\}$$
$$q, p, h \mapsto b$$

where $q \in \mathcal{Q}$ is the **query** sentence to be classified. $p \in \mathcal{P}$ is the exemplar sentence as a **premise**, $h \in \mathcal{H}$ is the **hypothesis** verbalized from the label of p. The task of ConEntail is to determine **if** q **entails [**p **entails** h**]**.

We follow (Khashabi et al., 2020b; Ye et al., 2021a) and translate sentence and label (x, y) to (q, p, h) in a PLM comprehensible format, e.g.,

- $x \mapsto q$, where q is the input sentence x with multiple-choice, for example, *(1) happy (2) sarcastic (3) sad, sentence: I bought this for myself ...*
- $x \mapsto p$: where p is the input sentence x with premise, for example, *sentence: I bought this for myself ...*
- $y \mapsto h$ where h is the label name, for example, h: *happy*

where we provide q with all possible labels as multiple-choice questions, and concatenate them in a linearized sentence. In supervised pretraining, q and p are two different surface forms of the same x, so that we can construct positive and negative examples for the later contrastive learning. In the test, q is the query sentence to be clarified and p and h are from the support set.

We use BERT_{base} to encode sentences to vector representation $\mathbf{h}$.

$$\mathbf{h}_q = \text{BERT}_{\text{base}}(q) \tag{1}$$

p and h are then concatenated into one sequence to be fed into the encoder:

$$ph = p\,\texttt{[SEP]}\,h \tag{2}$$

$$\mathbf{h}_{ph} = \text{BERT}_{\text{base}}(ph) \tag{3}$$

In the supervised pretraining, the embeddings of each mini-batch are composed by $\left\{\mathbf{h}_q^i, \mathbf{h}_{ph}^i\right\}_{i=1,\ldots,N}$, where N is the batch size. Then we calculate their pairwise cosine similarity $\text{sim}\left(\mathbf{h}_q^i, \mathbf{h}_{ph}^j\right) = \frac{\mathbf{h}_q^i \cdot \mathbf{h}_{ph}^j}{\|\mathbf{h}_q^i\| \cdot \|\mathbf{h}_{ph}^j\|}$ for contrastive training. $s_{ij} \in \{0, 1\}$ is denoted as the groundtruth of the predicted similarity, where $s_{ij} = 1$ is a positive pair when $y_i = y_j$, and vice versa. The positive/negative examples are constructed by all combinations of instances in the batch, note that we did not mine hard examples. We follow the balanced sampling strategy from Meta Classification Learning (Hsu et al., 2019) that each label in a mini-batch has an equal number of input sentences.

In the test phase, we calculate cosine similarities between q and all possible ph and output the most similar h as the prediction result. Thus, we consider our setting as a K-way N-shot learning, where K is determined by the test set, N varies from 0 to 80 in our experiments.

Given the pairwise similarity, we use Supervised Contrastive Loss (Khosla et al., 2020) to train the model:

$$\mathcal{L} = -\sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{p=1}^{N} \mathbb{1}_{y_i=y_p} \mathbb{1}_{i \neq p} \log \frac{\exp\left(\text{sim}\left(\mathbf{h}_q^i, \mathbf{h}_{ph}^p\right)/\tau\right)}{\sum_{a=1}^{N} \mathbb{1}_{i \neq a} \exp\left(\text{sim}\left(\mathbf{h}_q^i, \mathbf{h}_{ph}^a\right)/\tau\right)} \tag{4}$$

where $|P(i)| = \sum_{p=1}^{N} \mathbb{1}_{y_p=y_i}$ is the number of all positive pairs, τ is the temperature hyperparameters. Different from self-supervised contrastive learning losses, such as SimCSE (Gao et al., 2021), the positive pairs in Supervised Contrastive Loss can be more than one.

To enable zero-shot generalization, inspired by BERT masked language model (Devlin et al., 2019), we introduce a dummy premise “NULL” in both supervised pretraining and testing. During supervised pretraining, we randomly replace 5% of the premise p with “NULL” (**if** q **entails [“NULL” entails** h**]**.). During zero-shot test, the support set is empty and the model uses only “NULL” and label names to answer the question.

4 Experiments

In this section, we describe our experiment setups including dataset selection, evaluation, and baseline models.

4.1 Dataset Selection

For universal text classification, we aim to cover the most popular text classification tasks, such as topic classification, sentiment analysis, paraphrase identification, and natural language inference. Therefore, we adopt Crossfit (Ye et al., 2021a) that provides abundant hand-craft templates covering 56 classification tasks as the source of supervised pretraining and testing. We select 47 datasets as supervised pretraining sets and 9 widely accepted datasets as test sets: CoLA (Warstadt et al., 2018), QQP (Iyer et al., 2017), SST-2 (Socher et al., 2013), MRPC (Dolan and Brockett, 2005),

Method	CoLA	QQP	Hate_speech	MRPC	SCITAIL	Amazon	AGNews	Rotten_tomatoes	SST-2	AVG
			Unseen				*Seen*			
Random-guess	50.5	49.8	34.1	50.0	49.8	49.9	24.0	46.8	49.9	44.9
0-*shot*										
Crossfit†	0.0	0.0	0.0	0.0	0.2*	9.9*	0.0	59.9*	33.4*	11.5*
Unifew†	0.0	0.0	0.0	0.0	48.4*	63.7*	8.0*	57.4*	60.6*	26.5*
EFL	**62.6**	**60.5***	12.7	33.1	47.2*	71.9*	**60.8***	72.5*	79.1*	53.8*
CONENTAIL	58.5*	45.3	**78.3***	**58.1***	**68.7***	**89.7***	52.8*	**78.1***	**83.0***	**63.2***
10-*shot fine-tuning*										
Crossfit†	55.3 ±5.0	53.4 ±9.8	42.8 ±14.4	60.0 ±11.1	58.8 ±5.4	87.9 ±6.1	83.7 ±6.6	75.8* ±1.2	81.2 ±8.9	65.3
Unifew†	49.0 ±4.9	**60.4** ±6.0	34.9 ±6.8	57.7 ±6.3	53.4 ±2.4	88.8 ±3.6	**86.5** ±1.8	73.4 ±9.5	71.2 ±11.5	63.9
EFL	**63.7** ±0.2	**60.4** ±0.2	13.8 ±0.6	33.1* ±0.0	47.2* ±0.1	72.0 ±0.0	62.3 ±0.6	72.5* ±0.0	79.5 ±0.2	55.9
CONENTAIL	60.5* ±0.6	55.6 ±3.5	**44.7** ±2.2	**69.9*** ±0.9	**71.0*** ±0.9	**89.4*** ±0.1	70.3* ±2.1	**78.7*** ±0.2	**83.2*** ±0.2	**68.8***

Table 1: The main results of CONENTAIL compared with baselines. † indicates the models are generative models and the others are discriminative models. In the 10-shot evaluation, to offset the high variances from fine-tuning on such a small support set, the models are fine-tuned by 3 different random sampled support sets. After conducting experiments with and without supervised pretraining, we report the mean accuracy scores and the standard deviation of the best versions of models (in **bold**). We split the test sets in two groups, *seen* and *unseen*, which indicates if the test label names have occurred in the supervised pretraining. AVG is the highest average score of the two versions of models. If a model with supervised pretraining is better than that without supervised pretraining, it is indicated with a $*$.

SCITAIL (Khot et al., 2018), Amazon Polarity (Zhang et al., 2015a), AGNews (Zhang et al., 2015b), Rotten_tomatoes (Pang and Lee, 2005), Hate_speech_offensive (Davidson et al., 2017). For the sentence-pair datasets (e.g., QQP, SST-2, MRPC), we adopt the Crossfit method by concatenating the two sentences with [SEP] to form one sequence for either q or p. From the 47 datasets for supervised pretraining, we randomly select 128 annotated examples per label. As the same label name may occur in different datasets, to investigate the effect of label name overlapping, we pick 5 (out of 9) selected test sets with overlapping/seen label names for the supervised pretraining. The detailed dataset list is in Appendix B.

4.2 Evaluation

Supervised Pretraining To investigate the effect of the supervised pretraining, we consider two versions of all the compared models: (1) without supervised pretraining: we apply the original PLMs directly to the reformulated input-output test set. (2) with supervised pretraining: we first perform supervised pretraining on the PLMs and then evaluate the models with the updated parameters.

Zero-shot Evaluation In zero-shot evaluation, the only available resources for the target task are the possible label names and the whole test set will be used to evaluate the model.

Few-shot Evaluation In few-shot evaluation, in addition to the label names, a small support set are available for fine-tuning the universal classification model. The support set for each dataset is composed by k random sampled annotated examples per label, from the training data. With small support sets, the evaluation score may have huge variance, thus we fine-tune and evaluate the model with 3 different support sets and report the mean and standard deviation.

4.3 Baseline Models

We aim to evaluate models in different paradigms in the same universal classification experiment setting. To this end, we compare three baselines that are most representative of the current literature on generators and discriminators.

In this paper, we only consider the differences of the baselines in the meta-task formulation and their generator/discriminator nature while keeping other factors the same, so we reproduce the baselines strictly follow this rule, and use a similar size of pretrained language models as backbones, for a fair comparison. Because our generator/discriminator taxonomy suits many other existing works, with only subtle differences either in the templates or in the backbone PLMs from the baselines mentioned

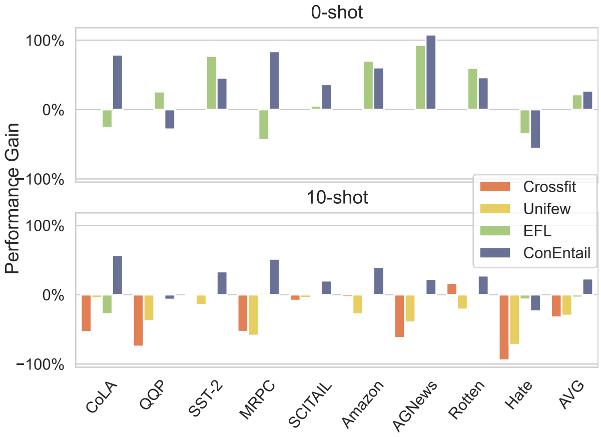


Figure 3: Relative performance gain of supervised pretraining on different datasets and models. The setting is the same with the main experiment. We do not plot zero-shot gains for the generators because most scores are 0 before and after supervised pretraining.

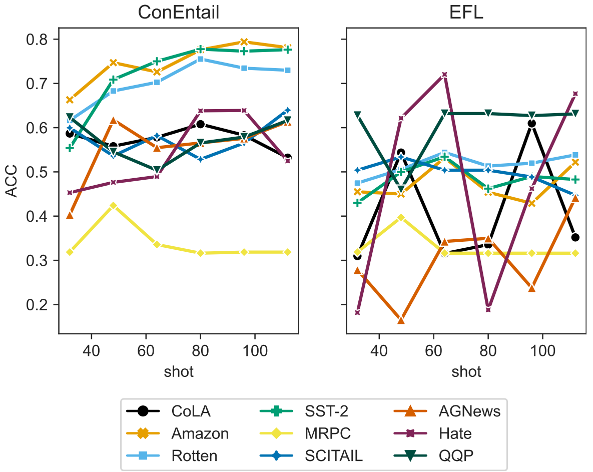


Figure 4: The effect of supervised pretraining data size. We show the zero-shot performance of ConEntail and EFL using different pretraining data size from 32 to 128 annotated sentences per label.

here, we do not add more baselines for comparisons.

Crossfit (Ye et al., 2021a): A generative model uses an encoder-decoder structure. The encoder takes the query sentence, and the decoder generates the label name.

Unifew (Bragg et al., 2021): A generative model concatenates all possible labels to the input sentence as multiple-choice question answering. It uses an encoder-decoder structure and generates the label names as answers.

EFL (Wang et al., 2021): A discriminative model reformulates the tasks as multiple entailment binary classifications. Both the query sentence and the label name are fed into the encoder. The embedding of [CLS] token is used for binary classification. The label with the highest probability is the predicted output. For supervised pretraining, we enumerate all possible labels for input and provide all the ground truths for the binary classification.

5 Results and Analysis

We design the following experiments to demonstrate and analyze the effectiveness of our method. First, we present the best scores of the compared models with or without supervised pretraining as our main result (Section 5.1). Then, we investigate the performance gain or loss of each model brought by the supervised pretraining (Section 5.2). Furthermore, we study the fine-grained impact of more labeled data in supervised pretraining or of more labeled data in support set (Section 5.3). Considering these results, we discuss the difference between discriminators and generators (Section 5.4). Finally, we show a case study of universal classification under a zero-shot scenario (Section 5.5).

5.1 Main Results

We evaluate the models in two scenarios, 0-shot learning and 10-shot learning (Table 1). The average performances of both discriminator models, EFL and ConEntail, significantly outperform random guess and two generation-based models. Particularly, ConEntail, with significantly improved average results, performs the best on 6 out of the 9 datasets in both 0-shot and 10-shot settings.

From the table, we also observe that the seen labels bring most improvements to Unifew in 0-shot setting. The 0-shot performance of Unifew in SST-2, SCITAIL and Amazon is far better than Crossfit. This is because Unifew has included the labels in the query sentences as multiple-choice questions, which provides the model additional familiarities from the supervised pretraining. In other words, although the 0-shot unseen accuracies of the generative models are mostly 0, their performances can be improved quickly with few-shot finetuning. This indicates that generative models are promising few-shot learners but not strong zero-shot learners.

5.2 Performance Gain from Supervised Pretraining

We then quantify the effect of supervised pretraining by Relative Performance Gain introduced (Ye et al., 2021a). Relative Performance Gain is the relative improvement brought by the supervised

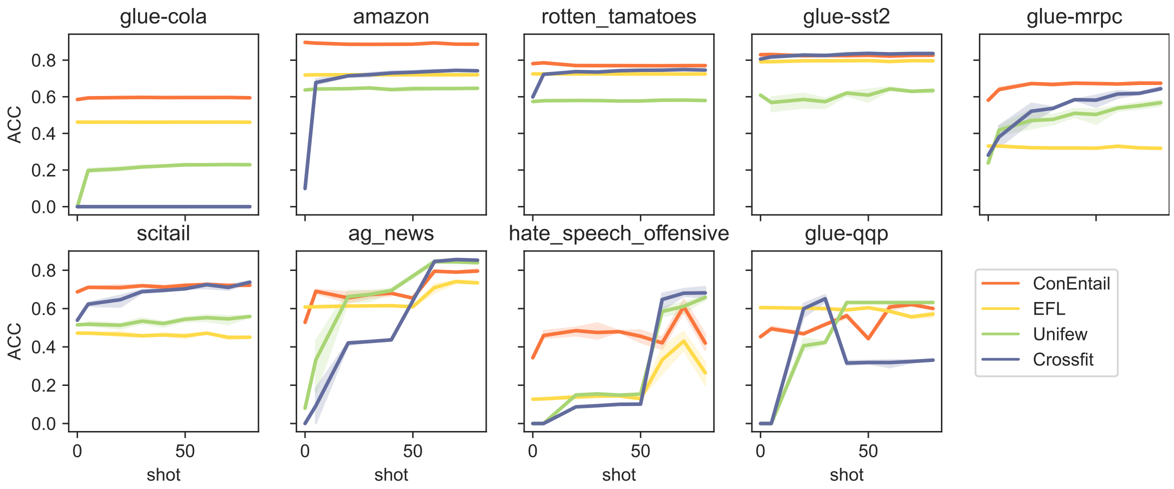


Figure 5: The effect of data size in the support set. We show the accuracy of the compared models fine-tuned with 0 to 80 examples in the support set. For each data size, we randomly sample 3 support sets for fine-tuning and evaluate on the same test set.

pretraining. It is defined as $\frac{\mathrm{Acc}_{\mathrm{w}}-\mathrm{Acc}_{\mathrm{w/o}}}{\mathrm{Acc}_{\mathrm{w/o}}}$, the performance difference between a supervised pretraining model $\mathrm{Acc}_{\mathrm{w}}$ and non-supervised pretraining model $\mathrm{Acc}_{\mathrm{w/o}}$, divided by the latter. The results are shown in Fig. 3.

We observe that supervised pretraining boosts the performance in most datasets in the 0-shot setting. But it lowers the scores in the 10-shot setting, except for CONENTAIL. CONENTAIL's performance rises in 7 out of 9 datasets in both 0-shot and 10-shot setting. This shows the general necessity of supervised pretraining for 0-shot evaluation and the effectiveness of our proposed model in both settings. The baseline models did not benefit from supervised retraining for the 10-shot setting because their conventional fine-tuning strategy is less likely to thoroughly update the parameters than our proposed contrastive learning. Noting that 10-shot evaluation means all the compared models only have 10 labeled examples for finetuning.

5.3 Impact of More Training data

More data in supervised pretraining: we investigate if more labeled data in supervised pretraining can improve zero-shot generalization. As the accuracies of generator models are close to zero in the zero-shot setting, we only consider discriminator models including CONENTAIL and EFL. These two models are supervised pretrained on different-scale datasets (32-128 sentences per label) and evaluated on the 9 test sets. As shown in Fig. 4, the performance of CONENTAIL has fewer fluctuations than the EFL, and the performance improvements of most datasets flat after 80 shots for CONENTAIL. This observation implies that the supervised pretraining has significant and reliable positive effects on CONENTAIL with merely a small amount of supervised dataset.

More data in support set: for models supervised pretrained with 128 annotated sentences per label, we plot the line chart of fine-tuning with 0 to 80 shots. As shown in Fig. 5, adding a few training sentences may not largely boost performance when the universal model is strong enough, but it improves the models significantly if the models have a slow start. Furthermore, though the generator model performances improve fast from 0 to 50 shots, the scores fluctuate largely. But after the first 50 shots, the improvements slow down, and the variances becomes much smaller. This implies that all the compared models are strong few shot learners, so that fine-tuning on large-scaled training data in the downstream tasks is unnecessary.

5.4 Discussion on the Differences Between Discriminator and Generator Models

The ineffectiveness of zero-shot Unifew and Crossfit are rooted in their generation nature. The original motivation of generation-based models is to resolve all kinds of NLP tasks, including both classification and generation. However, the universal classification task (i.e., tasks in this paper) are usually formulated as label picking from limited choices, while generation tasks aim to output human-readable sentences that match the input sentences – the target distributions for these 2 tasks are innately different. In the few-shot setting, fine-tuning with 10 more examples in the target task

I happily donate any covid vaccine dose which may be reserved for me to any person that is stupid enough to get one, or two, three, or four.									
mild	0.59	irony	0.48	happy	0.44	...	...	optimism	0.23
Guys it's OK. Delta says covid is over. IT'S OK NOW.									
mild	0.69	non-irony	0.48	irony	0.47	...	...	hate	0.10
The first patient who died of COVID in Kerala already had BP and cardiac issues, and he was 69. Bottomline : If we take precautions, we are still safe and can ensure others are safe too.									
optimism	0.51	mild	0.51	positive	0.43	...	...	hate	0.08
Could you imagine putting your faith into the narrative, getting jabs, getting sick from side effects (which is now being called the variant) and then being labeled an antivaxxer amidst this lie "only the unnvacinated are getting sick". They will use you up until there's nothing left.									
offensive	0.59	irony	0.58	mild	0.56	...	...	happy	0.25
... I don't see a monetary benefit. I don't see any professional benefit. Ask the people who believe what they are being told for an explanation because I don't see any.									
offensive	0.60	mild	0.54	irony	0.49	...	...	optimism	0.20
I can't do this anymore. I went from a house and 2 beautiful daughters and wife to homeless and left with literally nothing. ... They need to die painfully and even then they will never pay for their sins. All it takes it one moment in history for everything to change. You keep breaking men down to nothing. Those broken men will break you.									
offensive	0.80	negative	0.63	hate	0.59	...	...	optimism	0.10

Table 2: Case study of an unseen task. We use ConEntail in a zero-shot manner to analyze twitter and reddit sentiment during the Covid-Omicron surge. We pick 13 fine-grained sentiment labels and rank the labels by their similarity with the input sentence.

shifts the text generation distribution towards the label distribution, so the generated texts are more likely to be the labels, and this improves model performances. However, as the predictions are still in the large vocabulary space, they are likely to be altered by any disturbances. When using different support sets, the variances of the accuracy are far larger than that of the discriminator models. This also explains why Unifew performs better than Crossfit: the only difference between Unifew and Crossfit is that the input sentences of Unifew are appended with all possible label texts. By providing the generation process label hints, Unifew shifts its generation distribution towards label distribution and outperforms Crossfit. But the accuracy gap between Unifew and Crossfit drops from 15% to merely 0.7% while the number of shots increases from 0 to 10. As we stated before, Unifew performs better in the 0-shot setting because of its extra label hints. However, with an increase of shots, this advantage is diluted, resulting in a smaller performance difference between these two models.

5.5 A Case Study of Universal Classification

Consider a possible application scenario of universal classification: when dealing with new tasks and domains, especially related to newly emerged events, usually people only have the label names in hand. Based on this, we demonstrate a COVID-19 sentiment classification case study to show the universality of the proposed ConEntail model.

We use keywords to collect 50 sentences from Reddit and Twitter during the surge of the Omicron variant, then pick 13 fine-grained sentiment labels for this task: *positive, mild, negative, offensive, happy, anger, sad, hate, irony, non-offensive, non-irony, non-hate, optimism*. For each COVID-related query sentence, ConEntail model retrieves from all 13 possible labels and ranks them by similarity.

From the results Table 2 we observe that the model ranks the labels correctly most of the time. With antonyms paired with each other, such as *hate*/*non-hate* and *happy*/*sad*, our model successfully predicts the labels with only the label names, showing the polarity derived from the pairwise ranking are effective and reliable.

6 Conclusions

In this work, we study the universal classification problem, that leverages heterogeneous labeled datasets to benefit zero/few-shot learning in a new domain/task/dataset. We conduct systematic experiments on mainstream discriminators and generators models, thoroughly evaluate different models, reveal their innate properties of meta-task reformulation and supervised pretraining strategies. The results show that the generators with open-end prediction fail in zero-shot learning and the discriminators with a standard entailment meta-task hardly obtain a performance boost when more pretraining data is available. Our work provides a new angle for future researchers to explore universal NLP, and propose a new nested entailment meta-task and a supervised contrastive learning strategy, CONENTAIL, to make better use of widely available annotated datasets, and adapts to new datasets with limited resources.

Limitations

Although this paper aims to improve the universal generalization in the classification task, there are several limitations: (1) We do not compare with cloze-based models (Schick and Schütze, 2021a,b; Gao et al., 2020), because their templates are more complicated and hard to be reproduced with our current datasets. (2) We do not consider structural classification tasks, such as Named Entity Recognition and Relation Extraction. (3) We only take classification datasets into account because our implementation is restricted by huggingface datasets and human-curated templates. We plan to extend our framework to more datasets in the future. (4) Due to the constraints from the templates and datasets, the class number of each test set is below 10. We plan to extend our framework to more labels in the future work. (5) The compatibility of knowledge in similar tasks is assumed, but this assumption may not hold true due to varying annotation standards across datasets. For instance, MRPC and QQP are both paraphrase identification tasks, but MRPC uses hard example mining techniques, resulting in longer and more sophisticated sentences than QQP. (6) The current study is limited to English datasets and can be extended to multiple languages in the future by using multilingual PLMs and pretraining datasets.

Acknowledgments

We thank Qianying Liu for her valuable discussion.

References

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. *arXiv preprint arXiv:2107.07170*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

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.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. All nlp tasks are generation tasks: A general pretraining framework. *arXiv preprint arXiv:2103.10360*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. First quora dataset release: Question pairs.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020a. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020b. Unifiedqa: Crossing format boundaries with a single qa system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1896–1907.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1378–1387, New York, New York, USA. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tingting Ma, Jin-ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with entailment-based zero-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 786–796.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 115–124, USA. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

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. Association for Computational Linguistics.

Congying Xia, Wenpeng Yin, Yihao Feng, and Philip Yu. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1360, Online. Association for Computational Linguistics.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021a. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *arXiv preprint arXiv:2104.08835*.

Seonghyeon Ye, Jiseon Kim, and Alice Oh. 2021b. Efficient contrastive learning via novel data augmentation and curriculum learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1832–1838, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3905–3914.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021. Self-supervised contrastive cross-modality representation learning for spoken question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 28–39, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

A Hyperparameters and Implementation Details

Unifew and Crossfit, as generative models, use BART$_{\text{base}}$ (Lewis et al., 2020) as the backbone language model. In the supervised pretraining, we use AdamW optimizer (Loshchilov and Hutter, 2017) with learning rate 3e-5, warm-up ratio 0.6% and linear decay. In the meta-testing, we use the same hyperparameters and train 400 epochs for finetuning.

EFL and Entail2, as discriminator models, use BERT$_{\text{base}}$ (Devlin et al., 2019) as the backbone language model. In the supervised pretraining, we use AdamW optimizer (Loshchilov and Hutter, 2017) with learning rate 1e-5, warm-up ratio 6% and linear decay. In the meta-testing, we use the same hyperparameters and train 10 epochs for finetuning.

All the compared models use the same templates (map the input to the text) and the same verbalizers (map the label to the text) from the Crossfit paper (Ye et al., 2021a), as they covered more classification datasets than other frameworks. Note that the choices of template/verbalizer could cause large variance in performance (Zhao et al., 2021), and the effectiveness of Crossfit template/verbalizer had not been fully studied.

We use two NVIDIA A5000 for our experiments. The supervised pretraining takes 3 days and the evaluation takes 1 week for all the compared baselines.

B Details about Task Partition

Datasets	Labels	Test sentences	Citation
glue-cola	2	1043	(Warstadt et al., 2018)
glue-qqp	2	40430	(Iyer et al., 2017)
glue-sst2	2	872	(Socher et al., 2013)
glue-mrpc	2	408	(Dolan and Brockett, 2005)
scitail	2	1304	(Khot et al., 2018)
amazon_polarity	2	1000	(Zhang et al., 2015a)
ag_news	4	7600	(Zhang et al., 2015b)
rotten_tomatoes	2	1066	(Pang and Lee, 2005)
hate_speech_offensive	3	4957	(Davidson et al., 2017)

Table 3: The statistics of the 9 test data.

```
{
    "Suprevised_pretraining": ["tweet_eval-stance_hillary", "ethos-sexual_orientation", "climate_fever
        ", "hate_speech18", "tweet_eval-emotion", "hatexplain", "ethos-race", "emotion", "superglue-
        rte", "discovery", "anli", "wiki_auto", "scicite", "financial_phrasebank", "sms_spam", "
        kilt_fever", "tweet_eval-stance_climate", "medical_questions_pairs", "tweet_eval-
        stance_feminist", "ethos-directed_vs_generalized", "glue-wnli", "health_fact", "liar", "
        yahoo_answers_topics", "ethos-religion", "circa", "ethos-disability", "emo", "tweet_eval-hate
        ", "tweet_eval-sentiment", "superglue-wic", "tweet_eval-emoji", "glue-qnli", "ade_corpus_v2-
        classification", "ethos-national_origin", "dbpedia_14", "poem_sentiment", "yelp_polarity", "
        tweet_eval-stance_atheism", "onestop_english", "glue-rte", "wiki_qa", "ethos-gender", "
        superglue-wsc", "tweet_eval-stance_abortion", "paws", "tweet_eval-offensive"],
    "meta_test": ["glue-cola", "glue-qqp", "glue-sst2", "glue-mrpc", "scitail", "amazon_polarity", "
        ag_news", "rotten_tomatoes", "hate_speech_offensive"]
}
```

C Additional results

Method	**CoLA**	**QQP**	**SST-2**	**MRPC**	**SCITAIL**	**Amazon**	**AGNews**	**rotten_tomatoes**	**hate_speech**	**AVG**
Random-guess	50.5	49.8	49.9	50.0	49.8	49.9	24.0	46.8	34.1	44.9
PLM + 0-*shot*										
Crossfit	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Unifew	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EFL	**62.6**	48.1	44.7	**58.6**	44.8	42.3	**31.5**	45.4	19.6	44.2
CONENTAIL	32.7	**63.2**	**57.0**	31.6	**50.4**	**55.9**	25.4	**53.4**	**78.3**	**49.8**
PLM + *Supervised Pretraining* + 0-*shot*										
Crossfit	0.0	0.0	33.4	0.0	0.2	9.9	0.0	59.9	0.0	11.5
Unifew	0.0	0.0	60.6	0.0	48.4	63.7	8.0	57.4	0.0	26.5
EFL	46.2	**60.5**	79.1	33.1	47.2	71.9	60.8	72.5	12.7	53.8
CONENTAIL	**58.5**	45.3	**83.0**	**58.1**	**68.7**	**89.7**	**52.8**	**78.1**	**34.3**	**63.2**
PLM + 10-*shot fine-tuning*										
Crossfit	55.3 ±5.0	53.4 ±9.8	**81.2** ±8.9	**60.0** ±11.1	58.8 ±5.4	87.9 ±6.1	83.7 ±6.6	65.0 ±23.5	42.8 ±14.4	65.3
Unifew	49.0 ±4.9	**60.4** ±6.0	71.2 ±11.5	57.7 ±6.3	53.4 ±2.4	**88.8** ±3.6	**86.5** ±1.8	**73.4** ±9.5	34.9 ±6.8	**63.9**
EFL	**63.7** ±0.2	**60.4** ±0.2	79.5 ±0.2	32.3 ±0.4	46.7 ±1.1	72.0 ±0.0	62.3 ±0.6	72.4 ±0.2	13.8 ±0.6	55.9
CONENTAIL	38.6 ±4.4	55.6 ±3.5	62.4 ±3.4	46.1 ±2.4	**59.1** ±2.7	64.0 ±1.6	57.4 ±2.3	61.8 ±2.0	**58.6** ±11.0	55.9
PLM + *Supervised Pretraining* + 10-*shot fine-tuning*										
Crossfit	25.7 ±25.1	13.6 ±18.2	80.6 ±4.0	28.1 ±28.7	53.9 ±11.7	85.2 ±6.5	31.7 ±17.7	75.8 ±1.2	2.4 ±3.2	44.1
Unifew	46.6 ±12.4	37.5 ±30.6	60.9 ±1.4	23.8 ±24.9	51.0 ±1.8	63.9 ±2.9	52.2 ±16.1	57.9 ±1.5	9.8 ±6.5	44.8
EFL	46.1 ±0.0	**60.4** ±0.0	79.1 ±0.1	33.1 ±0.0	47.2 ±0.1	72.0 ±0.0	60.9 ±0.0	72.5 ±0.0	12.9 ±0.1	53.8
CONENTAIL	**60.5** ±0.6	51.8 ±1.9	**83.2** ±0.2	**69.9** ±0.9	**71.0** ±0.9	**89.4** ±0.1	**70.3** ±2.1	**78.7** ±0.2	**44.7** ±2.2	**68.8**

Table 4: The complete table of the main result.