# Contrastively-Trained Cross-Attention Improves Zero-Shot Natural Language Understanding

**Anonymous ACL submission**


## Abstract

Developing a general purpose model that can tackle many different Natural Language Understanding (NLU) tasks without requiring manually annotated data has become an ambitious yet desirable goal for the NLP research community. A simple and prominent approach for zero-shot text classification is to train a model on a generic language understanding task such as Natural Language Inference (NLI), and perform inference on NLU classification tasks using instructions or candidate templates. Those methods jointly encode the input document and the instruction into a single sequence leveraging self-attention layers and the next-sentence-prediction (NSP) pre-training task.

We hypothesize that this joint encoding limits the capabilities of large pre-trained encoders while being sub-optimal in many practical applications. To tackle those issues, we propose a novel approach that separates the encoding of the input document and use it as a ground reference to enhance the encoding of the instruction through cross-attention using an encoder-decoder architecture. We further propose a simple transformation on traditional NLI datasets that focuses on the learning of these Cross-Attention layers using contrasted data. Finally, we show that this approach do not need a full-sized decoder for best performance. Our experiments show that the proposed approach outperforms similar approaches by a large margin and sometimes achieves comparable results to fully fine-tuned methods.


## 1 Introduction

Natural language understanding (NLU) is a major research topic in natural language processing that has various practical applications. NLU is a broad task, with the goal of comprehending and determining the meaning behind a given text. Many NLU tasks, such as sentiment analysis, emotion recognition, or topic detection, involve assigning a semantic label (e.g. sentiment, emotion, or topic) to an input sentence. The conventional approach for building classification models is to use supervised learning with a large quantity of annotated training data. However, the construction of such dataset requires much time for collecting, curating, and annotation. Pre-trained language models provide us a partial solution to this problem, however, the training process still takes much time and requires large amount of resources (Vaswani et al., 2017; Devlin et al., 2018; Liu et al., 2019). In addition to that, the resulting model can only handle a single task. Therefore, we need separate models for each task, increasing the overall cost. As a result, it is desirable to create unified classification models that can perform multiple NLU classification tasks without requiring specific training datasets for each task.

As a solution for the above problem, several studies proposed to fine-tune large pre-trained model on generic classification tasks, such as Natural Language Inference. Natural language inference (NLI) is the task of determining whether a *hypothesis* is true (`ENTAILMENT`), false (`CONTRADICTION`), or undetermined (`NEUTRAL`) given a *premise*. We can see that by treating the input text of NLU tasks as the *premise* and the class labels as the *hypothesis*, we can use models trained on NLI to perform Zero-Shot NLU classification tasks. Yin et al. (2019) investigated the utilization of NLI datasets as the source training task of Zero-Shot models and showed promising results on 3 closed-set classification tasks. However, the majority of current studies consider the input document and the instruction text as a single sequence which is unpractical for real-world applications.

In this work, we propose to leverage cross-attention for zero-shot NLU classification tasks using contrasted NLI with instruction training. The proposed method uses an encoder-decoder architecture to process the instruction text separately from

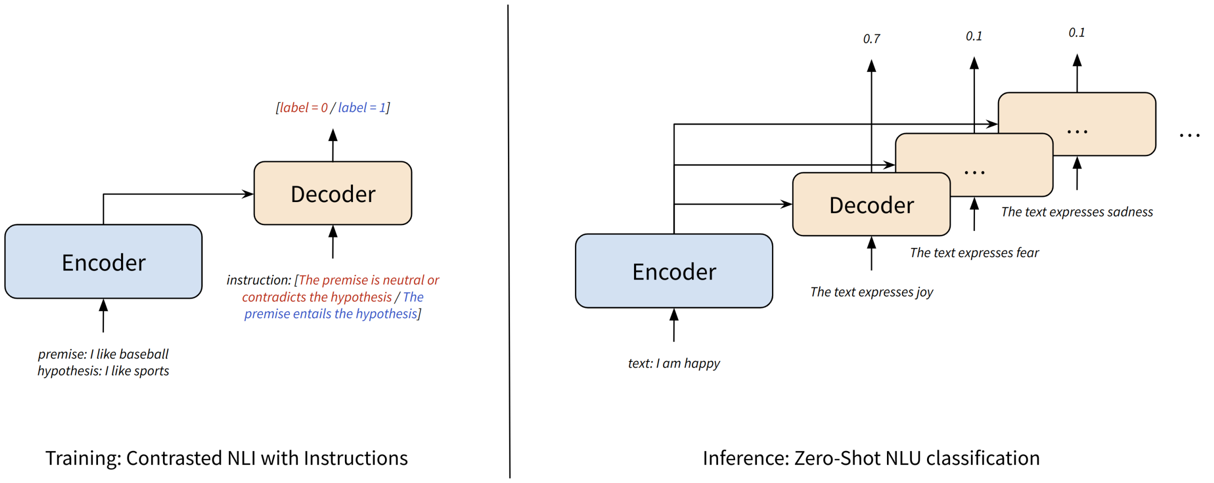


Figure 1: Overview of the proposed method. Cross-Attention layers in the Decoder are learnt using a Contrasted NLI with Instruction dataset (left). Zero-Shot NLU inference (right) uses similar input and output shapes than during training.

the input text document. The main contributions of this work are as follows:

1. We propose to use encoder-decoder architectures for zero-shot text classification to encode the input document and the class instruction text separately allowing us to leverage cross-attention layers

2. We demonstrate that training on a contrasted NLI dataset with natural language instructions is an effective source training task for the proposed architecture as well as for encoder-only architectures

3. We show through experiments that a small number of decoder layers outperform larger networks while having similar size to encoder-only methods

4. We conduct extensive experiments on a wide variety of tasks to confirm the effectiveness of the proposed method and find that the proposed method beats previous Zero-Shot methods by a large margin and achieves similar results to Few-Shot and Fine-Tuning methods.

## 2 Related Research

The problem of zero-shot learning for NLP tasks was first investigated in a pioneer study by Chang et al. (2008). Their idea was to map the input text and the labels into the same space of representation using explicit semantic analysis (Gabrilovich et al., 2007), then choose the label with the highest similarity score. Following the same approach, subsequent studies employed different methods to learn text representations and applied them for zero-shot NLP classification tasks (Song and Roth, 2014; Li et al., 2016; Veeranna et al., 2016; Yogatama et al., 2017; Rios and Kavuluru, 2018; Xia et al., 2018; Levy et al., 2017).

The emergence of LLMs revolutionized the progress in zero-shot learning for NLP, and since then, it has been an active research field in artificial intelligence (Brown et al., 2020; Schick and Schütze, 2021a,b; Gao et al., 2021; Li and Liang, 2021; Beltagy et al., 2022). There are various studies that investigated zero-shot learning for NLU, and they can be divided into two main sub-categories: methods based on transfer learning (transferring knowledge from another task) and methods based on data augmentation (creating artificial training data).

### 2.1 Transfer learning

One of the pioneering and simple method uses NLI to tackle zero-shot text classification is (Yin et al., 2019). Their main idea is to use the label itself (with a template) or to use a textual description of the label. For example, the label `SPORT`, can be converted to a sentence using the following template: *The text is about* ..., or, could be described as "an active diversion requiring physical exertion and competition". Motivated by the success of this research, Zhong et al. (2021a) extended that idea by combining data from more than 40 NLU classification tasks and converted them to a unified `YES/NO` question answering dataset. The authors reported strong zero-shot text classification accuracy across a variety of NLU tasks. Our approach is influenced by these works, but, rather than focusing on using multiple data sources, we focus on leveraging cross-attention layers in encoder-decoder models.

More recent approach leverage generative large language models (LLMs) such as GPT3, demonstrating strong capabilities in few-shot learning by scaling the number of parameters (Brown et al., 2020; Holtzman et al., 2021). Using prompts and in-context learning, few-shot text generation achieves very good results and keeps getting better (OpenAI, 2023).

Various studies attempted to alleviate the size and compute needed for those LLMs while retaining zero-shot performances on text classification tasks (Shi et al., 2022; Min et al., 2022; Hong et al., 2023; Li and Liang, 2021; Zhong et al., 2021b; Lester et al., 2021).

### 2.2 Data augmentation

Data augmentation is a technique that is commonly used when data is not highly available. It is extremely used in the fields of Computer Vision and Audio Processing but also in NLP (Feng et al., 2021). With the advances of generative LLMs, access to generated text data is relatively easy. When it comes to learning new task without available labeled data, recent methods either generate training data from label-descriptive prompts (Gao et al., 2021), use external unlabelled data to aggregate and stabilize results (Hong et al., 2023), or, use the vocabulary of the internal model as a data source to aggregate results (Zhao et al., 2023). Even though zero-shot learning methods inspired by data augmentation approaches achieve strong results, they still require to fully fine-tune the model on the synthetic datasets, which can be very time-consuming and not optimal at inference time.

## 3 Proposed approach

Out proposed method uses NLI as a source training task to perform classification on unseen tasks. In a similary way to what Yin et al. (2019) proposed, new tasks are mapped to an NLI format (premise and hypothesis) where the *premise* is the document to classify and the *hypothesis* an instruction (also called candidate label) representing the class in which the document can be classified. The format we used for the evaluated tasks are detailed in Table 1. To handle multiple sentences classification tasks, we use the markers (`text1`, `text2`, ...). Since Yin et al. (2019) did not provide any templates for multiple sentences classification tasks, we made them ourselves using the same idea.

In the following section, we detail our main contributions over previous similar works: about the usage of cross-attention layers and encoder-decoders architectures for zero-shot text classification tasks in Section 3.1, and about the contrasted NLI with instruction dataset used as the source training task in Section 3.2. Figure 1 shows an overview of the proposed method.

### 3.1 Leveraging encoder-decoders for text classification

Previous similar works (Yin et al., 2019; Min et al., 2022; Zhong et al., 2021a) use large pre-trained encoders to perform classification by leveraging the next sentence prediction (NSP) and/or mask language modeling (MLM) tasks learnt during the pre-training phase. Because, their inputs must follow the pre-training format, for zero-shot text classification, it is set as the concatenation of the input text with the candidate label into a single sequence.

On the other hand, we propose to split the encoding of the input text from the encoding of the candidate label and model their interaction using cross-attention layers. Not concatenating the input text with the candidate label has obvious practical advantages, especially when the number of candidate classes is high. However, we could think that those advantages come with a certain performance drawback. The proposed approach shows that cross-attention outperforms concatenation methods while having more practical advantages.

One of the reason we thought of doing this is the analogy with how humans execute textual tasks (specifically sentence classification tasks). The first step is usually to screen the input document (to understand it deeply) and then, resolve the task that involves the information present in that document (understand the instruction/question using the pre-processed information).

In other words, we believe that for zero-shot text classification, the cross-attention layer allows to guide the instruction, grounded by the input document like for translation or summarization tasks for generative models.

Formally, let $S = \{s_1, ..., s_N\}$ and $P = \{p_1, ..., p_M\}$ be a sequence of $N$ and $M$ tokens respectively. $S$ represents a document and $P$ the instruction (or prompt). We first map each to-

| Method / Task | Input | Instruction |
|---|---|---|
| Yin et al. (2019) | premise | hypothesis |
| Zhong et al. (2021a) | context | question |
| NLI | premise: ... hypothesis: ... | The premise {entails, contradicts, neutral} the hypothesis |
| Textual Entailment | text1: ... text2: ... | The text1 {entails, do not entails} the text2 |
| Paraphrase | text1: ... text2: ... | The text1 and the text2 are {paraphrase, not paraphrase} |
| Sentiment Analysis | text: ... | The text expresses a sentiment of {positive, negative} |
| Emotion Detection | text: ... | The text expresses an emotion of {joy, fear, ...} |
| Topic classification | text: ... | The text is about {topic1, topic2, topic3, ...} |

Table 1: Templates used for the evaluated tasks. The input corresponds to the input text sentences and the instruction a textual expression of the candidate class. Yin et al. (2019) used a NLI format which inspired our method. Zhong et al. (2021a) used a QA format following Khashabi et al. (2020).

ken $s_i$ into a contextualized, $h$-dimensional vector $\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_N\} = \{Encoder(s_1, ..., s_N)\}$. We feed this contextualized sequence $\mathbf{S}$ along with the sequence $P$ into the decoder (composed of cross-attention layers) and obtain a contextualized sequence $\mathbf{P}$ conditioned on $\mathbf{S}$ as follows: $\mathbf{P} = \{\mathbf{p}_1, ..., \mathbf{p}_M\} = Decoder(\mathbf{S}; P)$. $\mathbf{S}$ is fed as the key/value sequence to each of the cross-attention layers and $P$ as the query sequence. The sequence $\mathbf{P}$ conditioned on $S$ is then mapped to a 1-dimensional vector using a simple fully-connected layer: $C = Linear(mean(\mathbf{P}))$ using the $mean - pooling$ operation. A sigmoid operation, along with a binary cross entropy loss function is applied for learning.

### 3.2 Contrasted NLI with instruction

Yin et al. (2019) first used Natural Language Inference (NLI) as the source training task for zero-shot text classification. This approach is very simple in practice and shows strong results. However, Ma et al. (2021) demonstrates that models pre-trained on the next sentence prediction (NSP) task like BERT (Devlin et al., 2018) are already good zero-shot classifiers and thus, fine-tuning on NLI does not show that much improvements. We believe that there are two reasons for this: the dataset size, and the gap between the source NLI training task and the target zero-shot text classification inference task. While some previous works focus on collecting more data from different sources to better generalize on zero-shot tasks, our proposed approach focus on reducing the training and inference gap without additional training data.

We propose to modify the NLI task into an instruction-based NLI task where a new simple *instruction* column is added to the dataset. This new column is based on the label of the original dataset. As a result, we obtain a dataset having a similar format than the target zero-shot text classification task: the (*premise, hypothesis*) set can be used as the input document and the *instruction* as the candidate label.

To further tune the decoder towards learning the interaction between the input document and instruction, we use the idea of contrastive learning where each sample has one or more negative counterpart. Applying this, the resulting dataset is a contrasted NLI with instruction dataset that can be used for training models for zero-shot text classification. Furthermore, the resulting dataset is at least 2 times bigger than the original dataset (2 times for 1 negative instruction, 3 times for 2 negative instructions, ...).

The objective of this new dataset is not to classify a pair of text (*premise, hypothesis*) into eiter `ENTAILMENT`, `CONTRADICTION` or `NEUTRAL` classes but to match an input text document with an instruction. This objective is closer than the former to the Zero-Shot Text Classification task. An example of contrasted instructions are shown in Figure 2.

For datasets with 2 classes, building negative instructions is really simple and does not require any expertise knowledge (NLI can be converted to a binary task by merging the `CONTRADICTION` and `NEUTRAL` class to a `NON-ENTAILMENT` class). The proposed method can also be applied to any 2 classes dataset (not necessarily NLI). Building other contrasted instructions datasets is left for future work.

## 4 Evaluation

The proposed method is evaluated on a variety of NLU tasks in the zero-shot setting. We report evaluation results on the GLUE benchmark (Wang et al., 2018) and on closed-set classification tasks

**Entailment input:**
*premise: Two women are embracing while holding to go packages.*
*hypothesis: Two woman are holding packages.*

**Instruction**
*The meaning of the claim is logically inferred from the meaning of the premise*
*The meaning of the claim either contradicts the meaning of the premise, is unrelated to it, or does not provide sufficient information to infer the meaning of the premise*

---

**Non-entailment input:**
*premise: A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.*
*hypothesis: A man is wearing a black shirt.*

**Instruction**
*The meaning of the claim is logically inferred from the meaning of the premise*
*The meaning of the claim either contradicts the meaning of the premise, is unrelated to it, or does not provide sufficient information to infer the meaning of the premise*

Figure 2: Two examples in the contrasted NLI with instruction dataset. Each example has a positive instruction (blue) with label 1 and a negative instruction (red) with label 0.

as previous works. Evaluated tasks include: textual entailment, sentence paraphrases, topic classification, sentiment analysis, emotion classification, and more.

### 4.1 Evaluation datasets

**GLUE** The General Language Understanding Evaluation (GLUE benchmark) by Wang et al. (2018) is a collection of resources for training, evaluating, and analyzing natural language understanding systems. The STSB task is removed from the benchmark as it is a regression task. For MRPC and QQP, we report F1, for CoLA Matthews correlation and for all other tasks accuracy. Values are in percentages (scale by 100) as standard practices.

**Topic Classification** We use the large-scale "The Yahoo! Answers topic classification" dataset from Yin et al. (2019) and the AGNews dataset from Zhang et al. (2015). Yahoo has a total of 10 classes and AGNews has 4.

**Sentiment Analysis** We use 3 well-known sentiment analysis datasets: Movie Review (MV), Customers Review (CR) and Rotten Tomatoes (RT). For these 3 datasets, we use the data provided by Min et al. (2022).

**Emotion Classification** We use the Unify Emotion dataset provided by Yin et al. (2019). It consists of 9 emotions and a "no emotion" label.

Datasets details (size, classes, domains, ...) are given in Appendix A.

### 4.2 Baselines

**NLI 0SHOT-TC** Yin et al. (2019) first proposed NLI as the source training task for Zero-Shot Text Classification. It is a simple method with robust results.

**T5** Text-To-Text Transfer Transformers (Raffel et al., 2020) is a family of models that has strong performance on a variety of NLP tasks thanks to its unified text-to-text architecture. Its large scale pre-training and ability for multi-task learning makes it a popular choice for text-to-text tasks. We use the *large* version if not specified.

**LM-BFF** Gao et al. (2021) propose a prompt-based few-shot tuning method along with an automatic prompt generation technique. With only few examples, they consistently improve over a prompt-based zero-shot baseline by better leveraging the MLM pre-training task. Although their method use few training data, it shows how well current models perform when a small portion of data is available.

**MetaQA** Zhong et al. (2021a) aggregates 43 different dataset in a question-answering (QA) format and fine-tunes a zero-shot classifier. It outperforms UnifiedQA (Khashabi et al., 2020), a model trained with less QA dataset variety.

**NPM** Min et al. (2022) fills in the [MASK] token solely from retrieving a token from a text corpus using a non-parametric masked language model and combine with contrastive training, achieving decent performance on Zero-Shot Text Classification tasks.

**Retrieval ST5** Hong et al. (2023) encodes prompted label candidates with a sentence encoder and assign it to the input text embedding with the highest similarity. It uses an external 10k corpus to compensate for poor prompt label candidates.

### 4.3 Implementation details

The proposed method (encoder-decoder) uses the pre-trained T5-large model as it proposes an encoder as well as cross-attention layers in the decoder. For the proposed encoder-only method, we use the pre-trained RoBERTa-large model and concatenate the input document with the instruction as done in previous works. The contrasted NLI with instruction dataset is instantiated from the SNLI (Bowman et al., 2015) dataset. `NEUTRAL` and `CONTRADICTION` classes are merged together to form a new `NON-ENTAILMENT` class. The final Contrasted NLI with Instruction dataset has a size of 1.1M/20k/20k for the train/dev/test split which is double the size of the original SNLI dataset (550k/10k/10k). More details on hyper-parameters are shown in Appendix B. The reported results for the proposed method are averaged on 5 runs for

| | MNLI-m (acc) | MNLI-mm (acc) | MRPC (f1) | QNLI (acc) | QQP (f1) | RTE (acc) | SST-2 (acc) | WNLI (acc) | CoLA (Matt.) | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| **Zero-Shot** | | | | | | | | | | |
| Majority | 35.4 | 35.2 | 81.2 | 50.5 | 0.0 | 52.7 | 50.9 | 56.3 | 0.0 | 40.2 |
| Prompt-based ZS | 50.8 | 51.7 | 61.9 | 50.8 | 49.7 | 51.3 | 83.6 | 49.5 | 2.0 | 50.1 |
| NLI SHOT-TC | 54.4 | 55.1 | 70.1 | 50.0 | 25.2 | 65.7 | 85.0 | 42.2 | -3.7 | 49.3 |
| Contrast-Enc (ours) | 58.5 | 58.3 | 72.9 | 51.9 | 59.9 | 79.5 | 81.5 | 58.6 | -1.2 | 57.8 |
| Contrast-EncDec (ours) | **64.0** | **64.3** | **82.2** | **67.9** | **70.3** | **<u>87.8</u>** | **92.5** | **65.9** | **11.3** | **67.4** |
| **Few-Shot and FT** | | | | | | | | | | |
| LM-BFF (FS@16) | 70.7 | 72.0 | 77.8 | 69.2 | 69.8 | 68.7 | 92.6 | 79.7 | 18.7 | 68.8 |
| T5 (FT) | <u>89.9</u> | <u>89.6</u> | <u>92.4</u> | <u>94.8</u> | <u>73.9</u> | 87.2 | <u>96.3</u> | <u>85.6</u> | <u>61.2</u> | <u>85.6</u> |

Table 2: GLUE results. Prompt-based ZS and LM-BFF are from Gao et al. (2021). NLI 0SHOT-TC is using Yin et al. (2019). T5 is from Raffel et al. (2020). For our methods, Contrast-Enc uses RoBERTa while Contrast-EncDec uses T5. Approaches are grouped into those not using training examples (Zero-Shot) and those using training examples (Few-Shot and Fine-Tuning). The greatest values for Zero-Shot models are in **bold**, and the overall greatest values are <u>underlined</u>.

stability (see Appendix C for detailed results).

# 5 Results

## 5.1 GLUE Benchmark

The results for the GLUE benchmark are shown in Table 2.

The proposed method using the encoder-decoder model is on average +27 absolute points above the majority baseline showing that obtain results are not random. It is also almost on par with LM-BFF, a few-shot method that uses $K = 16$ examples for each class in each task showing that the source contrasted NLI training dataset generalizes well to unseen tasks. Our method even achieves better results than a fully fine-tuned model on the RTE dataset and achieves close results on QQP and SST2. Results on a variety of GLUE dataset shows the wide effective range of the proposed method.

Compared to the previous most similar work by Yin et al. (2019), the proposed method achieves more than +18 absolute points improvements (a 36 % increase) while using the same source training task (NLI). We are able to show drastic improvements without collecting any additional data.

We also reported the proposed method using an encoder-only model and it also outperforms previous works with the same encoding strategy (i.e., concatenation). It is on average +8 absolute points (17% increase) over Yin et al. (2019). These results show that the contrasted NLI training has a positive impact whether we are using encoder-only or encoder-decoder as the architecture. On top of this, separating the encoding of the input document from the instruction has an even greater positive impact on the results since encoder-decoder models perform better than encoder-only.

## 5.2 Closed-set text classification

To further investigate the performance of the proposed method, we evaluate our model on various closed-set text classification tasks. The results are shown in Table 3.

We first want to note that the results in Table 3 are quite sparse due to the fact that there are no benchmarks for closed-set text classification. In that setting, direct comparison is better than average comparison.

Evaluation shows that first, the proposed method using an encoder-only model under performs the baseline showing that using a contrasted NLI dataset with instruction combined with concatenation does not help on the evaluated closed-set text classification datasets. However, the proposed method (encoder-decoder) outperforms, with a large margin, every previous zero-shot methods by Yin et al. (2019), Zhong et al. (2021a), and, Hong et al. (2023) on every dataset. Hong et al. (2023) is only not beaten on Yahoo which could be explained by the large number of classes in this dataset. When comparing with the most similar work by Yin et al. (2019), evaluation is improved from 67.4 to 73.3 (almost +7 absolute points, a +8.7% increase).

The proposed method also significantly outperforms NPM (Min et al., 2022) that uses an external 10k size corpus during inference. It even beats LM-BFF (Gao et al., 2021), a few-shot method.

| | AGNews (acc) | Yahoo (acc) | UnifyEmotion (f1) | MR (acc) | RT (acc) | CR (acc) | AVG |
|---|---|---|---|---|---|---|---|
| **Zero-Shot** | | | | | | | |
| Majority | | | | | | | |
| NLI 0SHOT-TC | 74.6 | 53.3 | 27.0 | 78.5 | 80.5 | 90.7 | 67.4 |
| MetaQA | 82.0 | 54.3 | 28.5 | - | - | - | |
| Retrieval ST5 | 76.6 | **57.4** | - | 81.7 | 82.4 | 87.4 | - |
| Contrast-Enc (ours) | 70.8 | 46.0 | 26.6 | 75.7 | 74.3 | 85.2 | 63.1 |
| Contrast-EncDec (ours) | **87.7** | 55.0 | **29.9** | **86.7** | **87.3** | **92.0** | **73.3** |
| **Few-Shot and FT** | | | | | | | |
| NPM (corpus) | 74.5 | 53.9 | - | 83.7 | 86.0 | 81.2 | - |
| LM-BFF (FS@16) | - | - | - | 86.6 | - | 90.2 | - |
| RoBERTa (FT) | - | - | - | 90.8 | - | 89.4 | - |

Table 3: Zero-Shot results on closed-set classification tasks. NLI 0SHOT-TC is using (Yin et al., 2019). MetaQA is from (Zhong et al., 2021a), Retrieval ST5 from (Hong et al., 2023). NPM and RoBERTa are from (Min et al., 2022). LM-BFF is from Gao et al. (2021). For our methods, Contrast-Enc uses RoBERTa while Contrast-EncDec uses T5. Approaches are grouped into those not using training examples (Zero-Shot) and those using training examples (Few-Shot and Fine-Tuning). The greatest values for Zero-Shot models are in **bold**, and the overall greatest values are underlined.

### 5.3 Contrasted NLI with instruction

One of our core proposal is the contrasted NLI with instruction dataset that is used to train our models. As said in Section 3.2, the dataset is simply build using already existing NLI datasets. To prove the effectiveness of this dataset for our models, we propose to compare 4 different settings including the original dataset:

- **3-way**: original dataset with `ENTAILMENT`, `NEUTRAL`, and `CONTRADICTION` classes
- **Binary**: *3-way* dataset where `NEUTRAL` and `CONTRADICTION` classes are merged
- **Instruct**: *binary* dataset with the addition of (positive) instructions
- **Contrast**: *binary* dataset with contrasted (positive and negative) instructions

Results on closed-set classification tasks are shown in Figure 3

Results on the evaluated datasets show that: 2-classes datasets (*binary, instruct, contrast*) are on average better than *3-way*. Adding instructions (*instruct, contrast*) has a more significant positive impact with +3 points for *instruct* and +6 points for *contrast* compared to *3-way*. We think that this is thanks to the gap reduction between the training and inference tasks.

The difference between *instruct* and *contrast* in Figure 3 is interesting. We remark that on the MR, RT, and CR datasets, the two methods are similar while on the others, positive instructions only is worse than without any instructions. Because the latter often happens on datasets where even the baseline produces good results, properties of these datasets (sequence length, number of classes, ...) could explain this trend. We also noticed that this happened with sentiment analysis datasets so there could be a link. Further evaluation could explain this trend.

The proposed method (*contrast*) consistently outperforms every other method on all evaluated datasets with a large margin without showing a similar trend than the former *contrast* dataset. Adding contrasted instructions mitigates errors and does not show saturation while being consistent.

### 5.4 Number of cross-attention layers

Leveraging cross-attention layers for zero-shot text classification is one of our main proposal. Previous works focus only on using self-attention layers in the encoder, by concatenating the input document with the instruction (candidate label). Table 2 and Table 3 show the effectiveness of using cross-attention layers (i.e., encoder-decoder) for this kind of task. In this section, we propose to dive deeper on the usage of these cross-attention layers by experimenting different decoder size (i.e., different number of cross-attention layers). We experiment 1, 6, 12 and 24 cross-attention layers. Results are shown in Table 4.

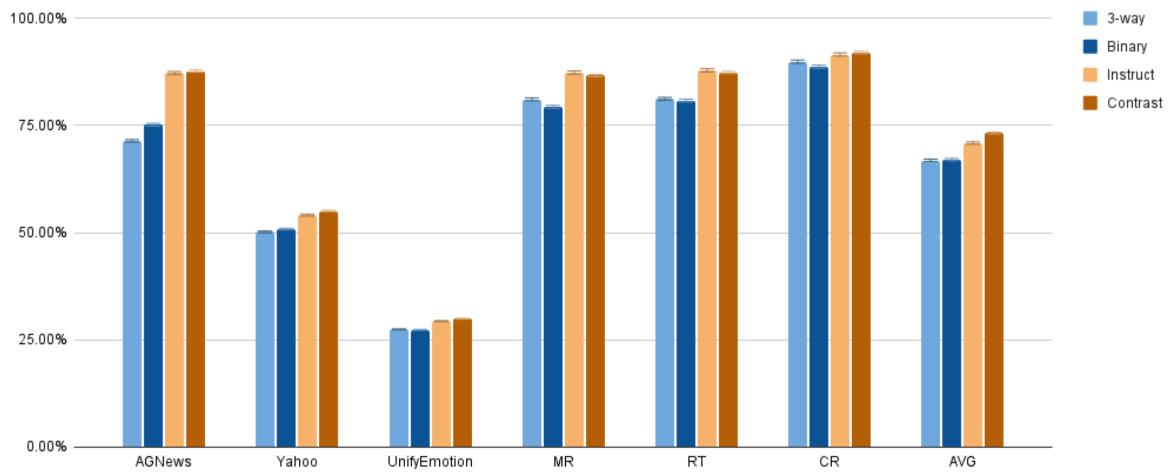


Figure 3: Zero-Shot results on closed-set classification tasks with different training dataset using the proposed model. The contrast dataset performs the best on average.

| **# Layers** | **GLUE** (average) | **Closed-Set** (average) | $\Delta$ (average) |
|---|---|---|---|
| 1 | 67.4 | 73.3 | 0.0 |
| 6 | 67.5 | 73.2 | 0.0 |
| 12 | 67.0 | 73.1 | -0.3 |
| 24 | 66.8 | 73.0 | -0.5 |

Table 4: Effect of the number of cross-attention layers (i.e., decoder size) on evaluated tasks. $\Delta$ represents the average difference compared to the smallest model.

On average, increasing the number of cross-attention layers does not result in higher performances unlike other trends in NLP (CITE). We see that having a small number of layers actually performs better than having a high number of layers, showing a saturation at around $N = 6$ layers.

To explain these number, we hypothesis that the contrastive training strategy is able to train smaller models with final good performance effectively. Indeed, we believe that this comes from the negative instruction examples in the training dataset. We believe that these examples force the cross-attention layers to effectively learn the meaning of the instruction, grounded by the meaning of the input document. During training, the same input document is seen twice (for a single epoch) but with different instructions. Thus, one of the input of the cross-attention layer stays the same while the other changes. This difference seems to be the key to effectively learn cross-attention layers in a contrasted way.

Another reason that could explain these results are the fact that the instructions are rather simple English sentences compared to the input document, so it would need less layers to learn its meaning.

This trend show that the add of a small decoder (1 to 6 layers) show significant improvements while adding only a few number of parameters compared to the full encoder-decoder model. Compared to encoder-only models, this results in a 4% increase for a single layer jumping from 356M to 370M parameters while being way more effective as shown in Table 2 and Table 3.

## 6 Conclusion

We propose to use cross-attention layers combined with a contrasted NLI dataset for zero-shot text classification. The proposed method allows the separation of the encoding of the input document and the candidate label at inference time unlike previous methods that concatenate them to form a single sequence. Evaluation on a large panel of NLU task including the GLUE benchmark and closed-set classification tasks demonstrates the effectiveness of our approach. Thanks to the nature of the contrasted training, we also showed that the proposed method do not need a large decoder to achieve strong results, close to few-shot or fine-tuning methods.

## 7 Limitations and Risks

The proposed method is still instruction (prompt) dependent and does not propose any strategy to improve them. Because the used instructions were generally short, the effect of longer instructions has not been evaluated and could be a topic for further research. It goes the same with longer documents. The evaluated datasets did not contain very long documents (i.e., longer than 512 tokens) and thus the robustness of the proposed method on longer inputs documents is still left unexplored.

The proposed method uses a contrasted NLI dataset that is twice the size of the original NLI dataset. This means the training time for a single epoch is also doubled with the same computation resources. This can be seen as a drawback even though training time is usually less important than inference time.

For multi-class classification problems, even thought inference should be faster than previous works, the decoder has to be run for every class which can be unpractical if the number of class is very high. Batch inference neglect this but at a certain computational cost.

Finally, because LLMs are pre-trained on large web corpus, we can not guarantee that some evaluated dataset were not present in the pre-training dataset. In that sense, expected results can vary depending on pre-training strategy. On top of this, as the datasets used for training includes bias, using different dataset may have a large impact on the results.

## References


Iz Beltagy, Arman Cohan, Robert L Logan IV, Sewon Min, and Sameer Singh. 2022. Zero-and few-shot nlp with pretrained language models. *ACL 2022*, page 32.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI*, volume 2, pages 830–835.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp.

Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.

Jimin Hong, Jungsoo Park, Daeyoung Kim, Seongjae Choi, Bokyung Son, and Jaewook Kang. 2023. Empowering sentence encoders with prompting and label retrieval for zero-shot text classification.

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

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Yuezhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia Sycara. 2016. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2678–2688.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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, Online. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Nonparametric masked language modeling. *arXiv e-prints*, pages arXiv–2212.

OpenAI. 2023. Gpt-4 technical report.

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.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 3132. NIH Public Access.

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.

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.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. knn-prompt: Nearest neighbor zero-shot inference, 2022b. *URL https://arxiv.org/abs/2205.13792*.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sappadla Prateek Veeranna, Jinseok Nam, Eneldo Loza Mencıa, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *Proceeding of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium: Elsevier*, pages 423–428.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S Yu Philip. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *EMNLP*.

D Yogatama, C Dyer, W Ling, and P Blunsom. 2017. Generative and discriminative text classification with recurrent neural networks. In *Thirty-fourth International Conference on Machine Learning (ICML 2017)*. International Machine Learning Society.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2023. Pre-trained language models can be fully zero-shot learners.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021a. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Conference on Empirical Methods in Natural Language Processing*.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021b. Factual probing is [mask]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033.

# A Datasets

Table 5 shows the list of the datasets we used in our Zero-Shot evaluation. In total, we used six datasets, two of them are Topic Classification (AGNews and Yahoo), three are Sentiment Analysis (Movie Reviews, Rotten Tomatoes, and Customer Reviews), and the last one is Emotion Classification (Unify Emotion).

# B Result with Standard deviation

Our models are trained for 1 epoch with a batch size of 64 and maximum sequence length of 128. AdamW optimizer (Loshchilov and Hutter, 2019) is used with a constant learning rate of 1e-4. Experiments are done on consumer GPUs for reproducibility: we use a single NVIDIA GeForce RTX 3090 Ti (24Gb of VRAM) GPU with QLoRA ($R = 64, alpha = 16$) (Dettmers et al., 2023) using HuggingFace's transformers (Wolf et al., 2020) and PEFT (Mangrulkar et al., 2022) libraries.

# C Result with Standard deviation

Table 6 and Table 7 show the standard deviation over 5 runs for our proposed models on the GLUE and closed-set classification datasets.

# D Fully Fine-tuned Results on GLUE

Table 8 shows results of RoBERTa and T5 models when fine-tuned on each dataset of the GLUE benchmark. Overall, RoBERTa leads to betters results in terms of number of parameters since its architecture is made for sequence classification tasks.

| Dataset name | Task | Size | Classes | Class names |
|---|---|---|---|---|
| AGNews | Topic | 7.6k | 4 | world, sports, business, sci-tech |
| Yahoo | Topic | 60k | 10 | Society & Culture, Science & Mathematics, Health, Education & Reference, Computers & Internet, Sports, Business & Finance, Entertainment & Music, Family & Relationships, Politics & Government |
| Movie Reviews | Sentiment | 2k | 2 | positive, negative |
| Rotten Tomatoes | Sentiment | 2k | 2 | positive, negative |
| Customer Reviews | Sentiment | 2k | 2 | positive, negative |
| Unify Emotion | Emotion | 15.6k | 10 | fear, joy, sadness, shame, guilt, disgust, anger, surprise, love, noemo |

Table 5: Details for the datasets used for zero-shot evaluation

| | MNLI-m (acc) | MNLI-mm (acc) | MRPC (f1) | QNLI (acc) | QQP (f1) | RTE (acc) | SST-2 (acc) | WNLI (acc) | CoLA (Matt.) | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| Contrast-Enc (ours) | 0.4 | 0.5 | 6.6 | 1.5 | 3.5 | 1.8 | 1.7 | 4.1 | 2.4 | 0.8 |
| Contrast-EncDec (ours) | 3.1 | 3.6 | 0.2 | 3.9 | 0.8 | 0.8 | 0.5 | 3.8 | 4.4 | 1.2 |

Table 6: Standard deviation over 5 runs for our methods, Contrast-Enc uses RoBERTa while Contrast-EncDec uses T5.

| | AGNews (acc) | Yahoo (acc) | UnifyEmotion (f1) | MR (acc) | RT (acc) | CR (acc) | AVG |
|---|---|---|---|---|---|---|---|
| Contrast-Enc (ours) | 4.4 | 4.2 | 1.4 | 2.8 | 2.9 | 3.8 | 2.0 |
| Contrast-EncDec (ours) | 0.5 | 0.9 | 0.6 | 0.8 | 0.5 | 0.3 | 0.3 |

Table 7: Standard deviation over 5 runs for our methods, Contrast-Enc uses RoBERTa while Contrast-EncDec uses T5.

| Model | Params | MNLI (acc) | MNLI-mm (acc) | MRPC (acc) | QNLI (acc) | QQP (acc) | RTE (acc) | SST-2 (acc) | WNLI (acc) | CoLA (Matt.) | AVG |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RoBERTa | 356M | **90.2** | **90.2** | **90.9** | 94.7 | **92.2** | 86.6 | **96.4** | **91.3** | **68.0** | **88.9** |
| T5 | **755M** | 89.9 | 89.6 | 89.9 | **94.8** | 89.9 | **87.2** | 96.3 | 85.6 | 61.2 | 87.2 |

Table 8: GLUE results for RoBERTa-large (356M) and T5-large (755M) model when fully fine-tuned on each task.