# Zero-Shot Classification by Logical Reasoning on Natural Language Explanations

**Chi Han** [1], **Hengzhi Pei** [1], **Xinya Du** [2], **Heng Ji** [1]
[1] University of Illinois at Urbana-Champaign
[2] The University of Texas at Dallas
{chihan3,hpei4,hengji}@illinois.edu, xinya.du@utdallas.edu

## Abstract

Humans can classify data of an unseen category by reasoning on its language explanations. This ability is owing to the compositional nature of language: we can combine previously seen attributes to describe the new category. For example, we might describe a sage thrasher as "it has a slim straight relatively short bill, yellow eyes and a long tail", so that others can use their knowledge of attributes "slim straight relatively short bill", "yellow eyes" and "long tail" to recognize a sage thrasher. Inspired by this observation, in this work we tackle zero-shot classification task by logically parsing and reasoning on natural language explanations. To this end, we propose the framework CLORE (Classification by LOgical Reasoning on Explanations). While previous methods usually regard textual information as implicit features, CLORE parses explanations into logical structures and then explicitly reasons along thess structures on the input to produce a classification score. Experimental results on explanation-based zero-shot classification benchmarks demonstrate that CLORE is superior to baselines, which we further show mainly comes from higher scores on tasks requiring more logical reasoning. We also demonstrate that our framework can be extended to zero-shot classification on visual modality. Alongside classification decisions, CLORE can provide the logical parsing and reasoning process as a clear form of rationale. Through empirical analysis we demonstrate that CLORE is also less affected by linguistic biases than baselines. [1]

## 1 Introduction

Humans are capable of understanding new categories by reasoning on natural language explanations (Chopra et al., 2019; Tomasello, 2009). For example, in Figure 1, we can describe sage thrashers as "having a slim straight relatively short bill, yellow eyes and a long tail". Then when we view a real sage thrasher the first time, we can match its visual appearance with attributes "slim straight relatively short bill", "yellow eyes" and "long tail", and then logically combine these results to recognize it. This ability has been shown to be applicable to both visual objects and abstract concepts (Tomasello, 2009). Compared to learning only through examples, using language information enables humans to acquire higher accuracy in less learning time (Chopra et al., 2019).

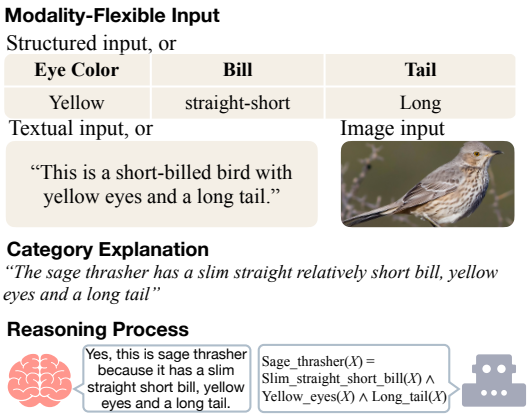


Figure 1: We propose to conduct zero-shot classification by logical reasoning on natural language explanations, just like humans do. This design encourages our approach to better utilize the compositional property in natural language explanations.

One important advantage of learning with natural language explanations is that explanations are often logical and compositional. That is, we can logically decompose the explanation of a new category into previously seen attributes (or similar ones) such as "yellow eyes" and "long tail". This enables us to reuse the knowledge on how these attributes align with visual appearances, and reduce the need for "trial-and-error". Furthermore, learning with explanations provides better interpretability which makes results more trustworthy.

Recently, there have been research efforts on using language information for zero-shot general-

[1] Code and data will be made publicly available upon publication



*Findings of the Association for Computational Linguistics: ACL 2023*, pages 8967–8981
July 9-14, 2023 

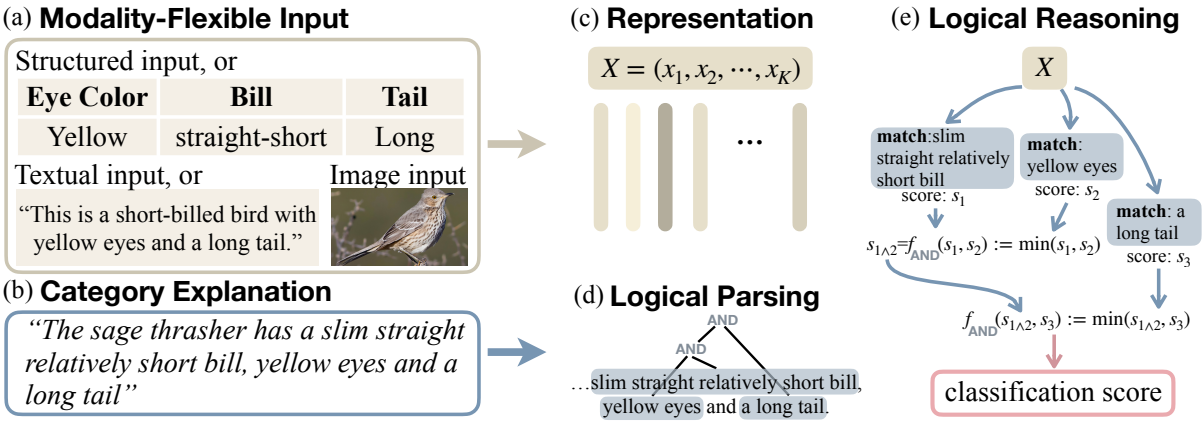


Figure 2: An illustrative figure of CLORE's working paradigm. After encoding the input (sub-figure(c)) we conduct logical parsing (sub-figure (d)) and logical reasoning (sub-figure(e)) over the explanations to obtain the classification score.

ization. Types of such language information include human-annotated explanations or task-level instructions (Menon et al., 2022; Sanh et al., 2022; Mishra et al., 2022). However, auxiliary language information is often treated merely as additional text sequences to be fed into pre-trained language models. This approach does not fully leverage the compositional nature of natural language, and does not provide sufficient interpretable rationales for its decisions.

Inspired by these observations, in this work we explore classifying unseen categories by logically reasoning on their language explanations. To this end, we propose the framework of Classification by LOgical Reasoning on Explanations (CLORE). CLORE works in two stages: it first parses an explanation into a logical structure, and then reasons along this logical structure. Figure 2 illustrates an example of classifying sage thrashers in this way. We first encode the inputs (Figure 2 (a) → (c)) get the logical structure of explanation (Figure 2 (b) → (d)). Then we detect if the input matches attributes, and we gather the matching scores along the logical structure to output the overall classification score (Figure 2 (c),(d)→(e)). In this case the logical structure consists of AND operators over three attributes. We test the model's zero-shot capacity by letting it learn on a subset of categories, and make it categorize data from other unseen types.

We conduct a thorough set of analysis on the latest benchmark for zero-shot classifier learning with explanations, CLUES (Menon et al., 2022). Our analysis shows that CLORE works better than baselines on tasks requiring higher level of compositional reasoning, which validates the importance of logical reasoning in CLORE. CLORE also demonstrates better interpretability and robustness against linguistic biases. Furthermore, as a test on generalizability of the proposed approach on other modalities, we built a new benchmark on visual domain: CUB-Explanations. It is built upon the image dataset CUB-200-2011 (Wah et al., 2011), while we associate each category with a set of language explanations. CLORE consistently outperforms baseline models in zero-shot classification across modalities.

To sum up, our contributions are as follows:

- We propose a novel zero-shot classification framework by logically parsing and reasoning over explanations.
- We demonstrate our model's superior performance and explainability, and empirically show that CLORE is more robust to linguistic biases and reasoning complexity than black-box baselines.
- We demonstrate the universality of the proposed approach by building a new benchmarks, CUB-Explanations. It is derived from CUB-200-2011 (Wah et al., 2011) by collecting natural language explanations for each category.

## 2 Related Work

**Classification with Auxiliary Information** This work studies the problem of classification through explanations, which is related to classification with auxiliary information. For example, in the natural language processing field, Mann and McCallum (2010); Ganchev et al. (2010) incorporate side information (such as class distribution and linguistic

structures) as a regularization for semi-supervised learning. Some other efforts convert crowd-sourced explanations into pseudo-data generators for data augmentation when training data is limited (Wang et al., 2020a; Hancock et al., 2018; Wang et al., 2020b). However, these explanations are limited to describing linguistic patterns (e.g., "this is class X because word A directly precedes B"), and are only used for generating pseudo labels. A probably more related topic is using explanations for generating a vector of features for classification (Srivastava et al., 2017, 2018). However, they either learn a black-box final classifier on features or rely on observed attributes of data, so their ability of generalization is limited.

The computer vision area widely uses class-level auxiliary information such as textual metadata, class taxonomy and expert-annotated feature vectors (Yang et al., 2022; Akata et al., 2015b; Xian et al., 2016; Lampert et al., 2009; Akata et al., 2015a; Samplawski et al., 2020). However, the use of label names and class explanations is mainly limited to a simple text encoder (Akata et al., 2015b; Xian et al., 2016; Liu et al., 2021; Norouzi et al., 2014). This processing treats every text as one simple vector in similarity space or probability space, whereas our method aims to reason on the explanation and exploit its compositional nature.

**Few-shot and Zero-shot Learning with Language Guidance** This work deals with the problem of learning with limited data with the help of natural language information, which is closely related to few-shot and zero-shot learning with language guidance in NLP domain (Hancock et al., 2018; Wang et al., 2020b; Srivastava et al., 2017, 2018; Yu et al., 2022; Huang et al., 2018). Besides the discussions in the previous subsection, recent pre-trained language models (LMs) (Devlin et al., 2019; Liu et al., 2019; Tam et al., 2021; Gao et al., 2021; Yu et al., 2022) have made huge progress in few-shot and zero-shot learning. To adapt LMs to downstream tasks, common practices are to formulate them as cloze questions (Tam et al., 2021; Schick and Schütze, 2021; Menon et al., 2022; Li et al., 2022b) or use text prompts (Mishra et al., 2022; Ye et al., 2021; Sanh et al., 2022; Aghajanyan et al., 2021). These approaches hypothetically utilize the language models' implicit reasoning ability (Menon et al., 2022). However, in this work we demonstrate with empirical evidence that adopting an explicit logical reasoning approach can provide better interpretability and robustness to linguistic biases.

In computer vision, recently there has been impressive progress on vision-language pre-trained models (VLPMs) (Li et al., 2022a; Radford et al., 2021; Li et al., 2019; Kim et al., 2021). These methods are trained on large-scale high-quality vision-text pairs with contrastive learning (Radford et al., 2021; Kim et al., 2021; Li et al., 2019) or mask prediction objective (Kim et al., 2021; Li et al., 2019). However, these model mostly focus on representation learning than understanding the compositionality in language. As we will show through experiments, VLPMs fits data better at the cost of zero-shot generalization performance.

There are also efforts in building benchmarks for cross-task generalization with natural language explanations or instructions (Mishra et al., 2022; Menon et al., 2022). We use the CLUES benchmark (Menon et al., 2022) in our experiment for structured data classification, but leave Mishra et al. (2022) for future work as its instructions are focused on generally describing the task instead of defining categories/labels.

**Neuro-Symbolic Reasoning for Question Answering** is also closely related to our approach. Recent work (Mao et al., 2019; Yi et al., 2018; Han et al., 2019) has demonstrated its efficacy in question answering, concept learning and image retrieval. Different from our work, previous efforts mainly focus on question answering tasks, which contains abundant supervision for parsing natural language questions. In classification tasks, however, the number of available explanations is much more limited (100∼1000), which poses a higher challenge on the generalization of reasoning ability.

## 3 Logical Parsing and Reasoning

Explanation-based classification is, in essence, a bilateral matching problem between inputs and explanations. Instead of simply using similarity or entailment scores, in this work we aim at better utilizing the logical structure of natural language explanations. A detailed illustration of our proposed model, CLORE, is shown in Figure 2. At the core of the approach is a 2-stage logical matching process: logical parsing of the explanation (Figure 2(d)) and logical reasoning on explanation and inputs to obtain the classification scores (Figure 2(e)). Rather than using sentence embeddings, our approach focuses more on the logical structure of language

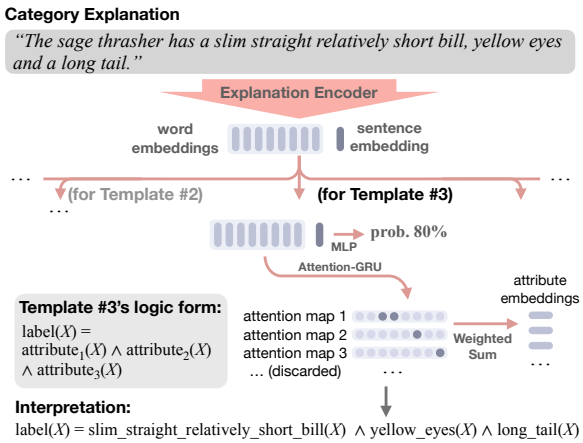


Figure 3: We parse each explanation into its logical structure. For each template, we predict its probability and attribute embeddings given by attention-based weighted sum.

explanations, setting it apart from logic-agnostic baselines such as ExEnt and RoBERTa-sim (which is based on sentence embedding similarity). To the best of our knowledge, ours is the first attempt to utilize logical structure in zero-shot classification benchmarks, and it also serves as a proof of concept for the importance of language compositionality. In the following part of this section we will describe these two stages. More implementation details including input representation can be found at Section 4 and 5.

### 3.1 Logical Parsing

This stage is responsible for detecting attributes mentioned in an explanation as well as recovering the logical structure on top of these attributes. (Figure 2(b) to Figure 2(d)). A more detailed illustration is given in Figure 3. We divide this parsing into 2 steps:

**Step 1: Selecting attribute Candidates** We deploy a attribute detector to mark a list of attribute candidates in the explanations. Each attribute candidate is associated with an attention map as in Figure 3. First we encode the explanation sentence with a pre-trained language encoder, such as RoBERTa (Liu et al., 2019). This outputs a sentence embedding vector and a sequence of token embedding vectors. Then we apply an attention-based Gated Recurrent Unit (GRU) network (Qiang et al., 2017). Besides the output vector at each recurrent step, attention-based GRU also outputs an attention map over the inputs that is used to produce the output vector. In this work, we use the sentence embedding vector as the initialization vector $h^0$ for GRU, and word embeddings as the inputs. We run GRU for a maximum of $T$ (a hyper-parameter) steps, and get $T$ attention weight maps. Finally we adopt these attention maps to acquire weighted sums of token features $\{w_t | t \in [1..T]\}$ as attribute embeddings.

**Step 2: Parsing Logical Structure** The goal of this step is to generate a logical structure over the attribute candidates in the previous step. As shown in Figure 2(d), the logical structure is a binary directed tree with nodes being logical operators `AND` or `OR`. Each leaf node corresponds to an attribute candidate. In this work, we need to deal with the problem of undetermined number of attributes, and also allow for differentiable optimization. To this end, we define a fixed list of tree structures within maximum number of $T$ leaf nodes, each resembling the example in Figure 3. A complete list is shown in Appendix A.2. We compute a distribution on templates by applying an multi-layer perceptron (MLP) with soft-max onto the explanation sentence embedding. This provides a non-negative vector $\boldsymbol{p}$ with sum 1, which we interpret as a distribution over the logical structure templates. If the number of attributes involved in the template is fewer than $T$, we discard the excessive candidates in following logical reasoning steps.

### 3.2 Logical Reasoning

After getting attribute candidates and a distribution over logical structures, we conduct logical reasoning on the input to get the classification score. An illustration is provided in Figure 2(e).

**Step 1: Matching attributes with Inputs** We assume that the input is represented as a sequence of feature vectors $X = (x_1, x_2, \cdots, x_K)$. First we define a matching score between attribute embedding $w_t$ and input $X$ as the maximum cosine similarity:

$$\text{sim}(X, w_t) := \max_k \cos(x_k, w_t).$$

**Step 2: Probabilisitc Logical Reasoning** This step tackles the novel problem of reasoning over logical structures of explanations. During reasoning, we iterate over each logical tree template and walk along the tree bottom-up to get the intermediate reasoning scores node by node. First, for leaf nodes in the logical tree (which are associated with attributes), we use the attribute-input matching scores in the previous step as their intermediate

| Top-1 acc/% | CLUES-Real | + pre-training |
|---|---|---|
| **ExEnt** | 54.8 | 52.7 |
| **RoBERTa-sim** | 45.1 | 46.3 |
| **CLORE-plain** | 45.8 | 49.8 |
| **CLORE** | **57.4** | **55.2** |

Table 1: Cross-task generalization results on CLUES dataset (Menon et al., 2022). The first row of results are acquired by only fine-tuning on CLUES-Real, and the second row shows results with additional pre-training on CLUES-Synthetic.

scores. Then, for a non-leaf node, if it is associated with an AND operator, we define its intermediate score as $\min(s_1, s_2)$ with $s_1$ and $s_2$ following common practice (Mao et al., 2019). If the non-leaf node is associated with an OR operator instead, we use $\max(s_1, s_2)$ as the intermediate score. The intermediate score of the root node $s_{root}$ serves as the output reasoning score. Note that we generated a *distribution over logical structures* rather than a deterministic structure. Therefore, after acquiring the reasoning scores on each structure, we use the probability distribution weight $\boldsymbol{p}$ to sum up the scores $\boldsymbol{s}$ of all structures. The resulting score is then equivalent to probabilistically logical reasoning over a distribution of logical structures.

$$s_{expl} = \boldsymbol{p}^\top \boldsymbol{s}$$

We also consider that some explanations might be more or less certain than others. When using words like "maybe", the explanation is less certain than another explanation using word "always". We model this effect by associating each explanation with a certainty value $c_{certainty}$, which is produced by another MLP on the explanation sentence embedding. So we scale the score $s_{expl}$ with $c_{certainty}$ in logit scale:

$$s_{scaled} = \sigma(c_{certainty} \cdot \text{logit}(s_{expl}))$$

Intuitively, the training phase will encourage the model to learn to assign each explanation a certainty value that best fits the classification tasks.

**Step 3: Reasoning over Multiple Explanations** There are usually multiple explanations associated with a category. In this case, we take the maximum $s_{scaled}$ over the set of explanations as the classification score for this category.

## 4 Experiments on Zero-Shot Classification

In this section we conduct in-depth analysis of our proposed approach towards zero-shot classification with explanations. We start with a latest benchmark, CLUES (Menon et al., 2022), which evaluates the performance of classifier learning with natural language explanations. CLUES focuses on the modality of structured data, where input data is a table of features describing an item. This data format is flexible enough for computers on a wide range of applications, and also benefits quantitative analysis in the rest part of this section.

### 4.1 CLUES benchmark

CLUES is designed as a cross-task generalization benchmark on structured data classification. It consists of 36 real-world and 144 synthetic multi-class classification tasks, respectively. The model is given a set of tasks for learning, and then evaluated on a set of unseen tasks. The inputs in each task constitute a structured table. Each column represents an attribute type, and each row is one input datum. In each task, for each class, CLUES provides a set of natural language explanations.

We follow the data processing in Menon et al. (2022) and convert each input into a text sequence. The text sequence is in the form of "`odor | pungent [SEP] ... [SEP] ring-type | pendant`", where "`odor`" is the attribute type name, and "`pungent`" is the attribute value for this input, so on and so forth. For CLORE, we encode the sentence with RoBERTa (Liu et al., 2019)[2] and use the word embeddings as input features $X$. More implementation details can be found in Appendix A.1. We use ExEnt as a baseline, which is an text entailment model introduced in the CLUES paper. ExEnt uses pre-trained RoBERTa as backbone. It works by encoding concatenated explanations and inputs, and then computing an entailment score. We also introduce a similarity-based baseline, RoBERTa-sim, which uses cosine between RoBERTa-encoded inputs and explanations as classification scores. Finally, we compare with CLORE-plain as an ablation study, which ignores the logical structure in CLORE and plainly addes all attribute scores as the overall classification scofre.

[2] https://huggingface.co/roberta-base

| Task | Natural Language Explanation | Interpreted Logical Structure |
|---|---|---|
| car-evaluation | *Cars with higher safety and capacity are highly acceptable for resale.* | Label$(X)$ = with_higher_safety $(X) \wedge$ and_capacity $(X)$ |
| indian-liver-patient | *Age group above 40 ensures liver patient* | Label$(X)$ = group_above_40 $(X) \wedge$ ensures_liver $(X)$ |
| soccer-league-type | *If the league is W -PSL then its type is women's soccer* | Label$(X)$ = league_is_W $(X)$ |
| award-nomination-result | *If the name of association has 'American' in it then the result was mostly won.* | Label$(X)$ = association_has_'American' $(X)$ |

Table 2: Examples of interpreted logical structures learned by CLORE. We randomly select 5 tasks from CLUES dataset, and use the alphabetically first explanation for interpretation. In each logical structure, the words corresponding to the detected attributes are colored in the explanation.

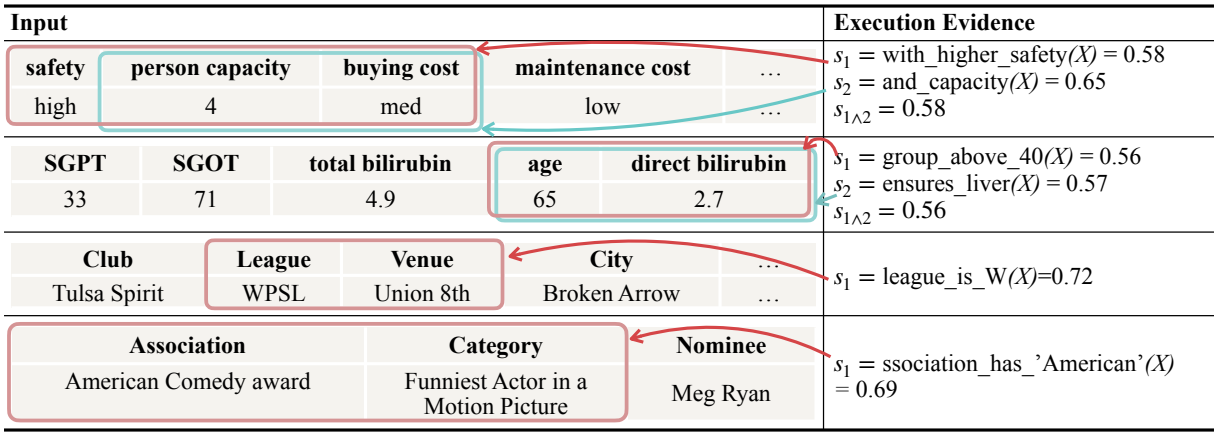


Figure 4: Examples of logical reasoning evidence. The evidence table cells are linked to attributes with colored arrows.

### 4.2 Zero-Shot Classification Results

Zero-shot classification results are listed in Table 1. CLORE outperforms the baseline methods on main evaluation metrics. To understand the effect of backbound model, we need to note that ExEnt also uses RoBERTa as the backbone model, so the CLORE and baselines do not exhibit a significant difference in basic representation abilities. The inferior performance of RoBERTa-sim compared to ExEnt highlights the complexity of the task, indicating that it demands more advanced reasoning skills than mere sentence similarity. Furthermore, as an ablation study, CLORE outperforms CLORE-plain, which serves as initial evidence on the importance of logical structure in reasoning.

### 4.3 Effect of Explanation Compositionality

What causes the difference in performance between CLORE and baselines? To answer this question, we investigate into how the models' performance varies with the compositionality of each task on CLUES. Table 3 provides a pair of examples. An explanations is called "simple explanation" if it only describes one attribute, e.g., "If safety is high, then the car will not be unacceptable.". Other explanations describe multiple attributes to define a class, e.g., "Cars with higher safety and medium luggage boot size are highly acceptable for resale.". We define the latter type as "compositional explanation". In Figure 7 we plot the classification accuracy against the proportion of compositional explanations in each subtask's explanation set. Intuitively, with more compositional explanations, the difficulty of the task increases, so generally we should expect a drop in performance. Results show that, on tasks with only simple explanations (x-value = 0), both models perform similarly. However, with higher ratio of compositional explanations, CLORE's performance generally remains stable, but ExEnt's performance degrades. This validates our hypothesis that CLORE's performance gain mainly benefits from its better compositional reasoning power.

To further explore the effect of logical reasoning on model performance. Figure 5 plots the performance regarding the maximum number of

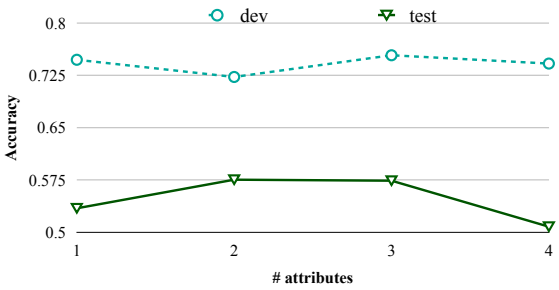


Figure 5: The effect of maximum number of attributes $T$ on the classification performance. When $T = 1$ the model reduces to a simple similarity-based model.

| **Compositional Explanation** |
|---|
| *Cars with higher safety and medium luggage boot size are highly acceptable for resale.* |
| **Simple Explanation** |
| *If safety is high, then the car will not be unacceptable.* |

Table 3: Examples of a compositional explanation and a simple one in CLUES dataset.

attributes $T$. Generally speaking, when $T$ is larger, CLORE can model more complex logical reasoning process. When $T = 1$, the model reduces to a simple similarity-based model without logical reasoning. The figure shows that when $T$ is 2∼3, the model generally achieves the highest performance, which also aligns with our intuition in the section 3. We hypothesize that a maximum logical structure length up to 4 provides insufficient regularization, and CLORE is more likely to overfit the data.

### 4.4 Interpretability

CLORE is interpretable in two senses: 1) it parses logical structures to explain how the explanations are interpreted, and 2) the logical reasoning evidence serves as decision making rationales. To demonstrate the interpretability of CLORE, in Table 2 and Figure 4 we present examples of the parsed logical structure and reasoning process.

The first example of Table 2 shows that CLORE selects "*with higher safety*" and "*and capacity*" as attributes candidates, and uses an AND operator over the attributes. In Figure 4 correspondingly, two attributes match with columns 1∼3 and 2∼3, respectively. This example is correctly classified by our model, but mis-classified by the ExEnt baseline.

To quantitatively evaluate the learned attributes, we manually annotate keyword spans for 100 out of 344 explanations. These spans describe the key attributes for making the explanation. When

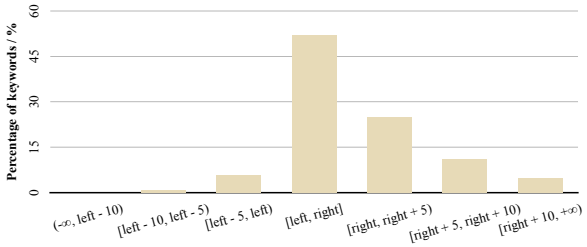


Figure 6: The position of detected attributes relative to the expert-annotated keyword spans. Y-axis is the proportion of explanations. Each interval category on x-axis denotes a position range relative to the keyword span in the explanation.

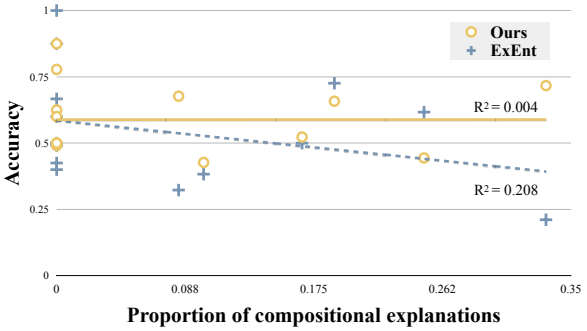


Figure 7: The classification accuracy on zero-shot tasks in CLUES plotted against the proportion of compositional explanations. (There are multiple tasks with only simple explanations, so there are multiple points at $x = 0$ position.)

there are multiple attributes detected, we select the one closest to the keyword span. Then we plot the histogram of the relative position between top-attention tokens and annotated keyword spans in Figure 6. From the figure we can see that the majority of top-attention tokens (52%) fall within the range of annotated keyword spans. The ratio increases to 81% within distance of 5 tokens from the keyword span, and 95% within distance of 10 tokens.

### 4.5 Robustness to linguistic bias

Linguistic biases are prevalent in natural language, which can subtly change the emotions and stances of the text (Field et al., 2018; Ziems and Yang, 2021). Pre-trained language models have also been found to be affected by subtle linguistic perturbations (Kojima et al., 2022) and hints (Patel and Pavlick, 2021).

In this section we investigate how different models are affected by these linguistic biases in inputs. To this end, we experiment on 3 categories of linguistic biases. *Punctuated*: inspired by discussions about linguistic hints in (Patel and Pavlick, 2021),

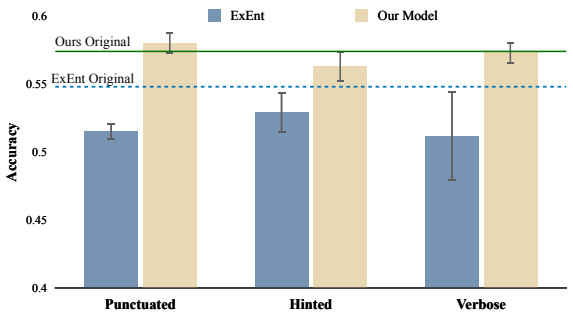


Figure 8: The effect of linguistic biases on classifiers. *Punctuated*, *Hinted* and *Verbose* are three types of biasing strategies. The two horizontal lines denote the original performance. Error bars denote standard deviation.

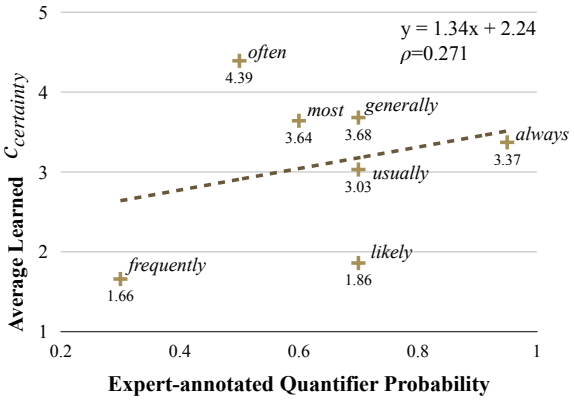


Figure 9: Comparison between the learned certainty coefficients $c_{certainty}$ in CLORE and expert annotations in Srivastava et al. (2018)..

we append punctuation such as "?" and "..." to the input in order to change its underlying tone. *Hinted*: we change the joining character from "|" to phrases with doubting hints such as "is claimed to be". *Verbose*: Transformer-based models are found to attend on a local window of words (Child et al., 2019), so we append a long verbose sentence ($\approx$ 30 words) to the input sentence to perturb the attention mechanism. These changes are automatically applied.

Results are presented in Figure 8. Compared with the original scores without linguistic biases (the horizontal lines), CLORE's performance is not significantly affected. But ExEnt appears to be susceptible to these biases with a large drop in performance. This result demonstrates that ExEnt also inherits the sensitivity to these linguistic biases from its PLM backbone. By contrast, CLORE is encouraged to explicitly parse explanations into its logical structure and conduct compositional logical reasoning. This provides better inductive bias for classification, and regulates the model from leveraging subtle linguistic patterns.

### 4.6 Linguistic Quantifier Understanding

Linguistic quantifiers is a topic to understand the degree of certainty in natural language (Srivastava et al., 2018; Yildirim et al., 2013). For example, humans are more certain when saying something *usually* happens, but less certain when using words like *sometimes*. We observe that the certainty coefficient $c_{certainty}$ that CLORE learns can naturally serve the purpose the of modelling quantifiers. We first detect the existence of linguistic quantifiers like *often* and *usually* by simply word matching. Then we take the average of $c_{certainty}$ on the matched explanations. We plot these values against expert-annotated "quantifier probabilities" in (Srivastava et al., 2018) in Figure 9. Results show that $c_{certainty}$ correlates positively with "quantifier probabilities" with Pearson correlation coefficient value of 0.271. In cases where they disagree, our quantifier coefficients also make some sense, such as assigning *often* a relatively higher value but giving *likely* a lower value.

### 4.7 Linguistic Quantifier Understanding

Linguistic quantifiers is a topic to understand the degree of certainty in natural language (Srivastava et al., 2018; Yildirim et al., 2013). For example, humans are more certain when saying something *usually* happens, but less certain when using words like *sometimes*. We observe that the certainty coefficient $c_{certainty}$ that CLORE learns can naturally serve the purpose the of modelling quantifiers. We first detect the existence of linguistic quantifiers like *often* and *usually* by simply word matching. Then we take the average of $c_{certainty}$ on the matched explanations. We plot these values against expert-annotated "quantifier probabilities" in (Srivastava et al., 2018) in Figure 9. Results show that $c_{certainty}$ correlates positively with "quantifier probabilities" with Pearson correlation coefficient value of 0.271. In cases where they disagree, our quantifier coefficients also make some sense, such as assigning *often* a relatively higher value but giving *likely* a lower value.

## 5 Extending to Visual Inputs

Natural language explanations are prevalent in other applications as well. Taking this observation, in this section we evaluate whether CLORE

| | Model | $ACC_U$ | $ACC_S$ | $ACC_{\mathbf{H}}$ |
|---|---|---|---|---|
| w/o VLPMs | TF-VAEGAN$_{expl}$ | 4.7 | 39.1 | 8.3 |
| | CLORE (ours) | **6.6** | **51.1** | **11.7** |
| w/ VLPMs | CLIP$_{linear}$ | 34.3 | 41.2 | 37.4 |
| | CLIP$_{finetuned}$ | 29.9 | **66.9** | 41.3 |
| | CLORE$_{CLIP}$ (ours) | **39.1** | 65.8 | **49.1** |

Table 4: Generalized zero-shot classification results (in percentage) on CUB-Explanations dataset.

can be extended to visual domain.

### 5.1 Datasets

Due to lack of datasets on evaluating zero-shot classification with compositional natural language explanations, we augment a standard visual classification datasets with manually collected explanations. Specifically, we select CUB-200-2011 (Wah et al., 2011), a bird image classification, as the recognition of birds benefits a lot from their compositional features (such as colors, shapes, etc.).

**CUB-Explanations** We build a CUB-Explanations dataset based on CUB-200-2011, which originally includes $\sim$ 12k images with 200 categories of birds. 150 categories are used for training and other 50 categories are left for zero-shot image classification. In this work, we focus on the setting of zero-shot classification using natural language explanations. Natural language explanations of categories are more efficient to collect than the crowd-sourced feature annotations of individual images. They are also similar to human learning process, and would be more challenging for models to utilize. To this end, we collect natural language explanations of each bird category from Wikipedia. These explanations come from the short description part and the *Description*, *Morphology* or *Identification* sections in the Wikipedia pages. We mainly focus on the sentences that describe visual attributes that can be recognized in images (e.g. body parts, visual patterns and colors). Finally we get 1$\sim$8 explanation sentences for each category with a total of 991 explanations.

For evaluation, we adopt the three metrics commonly used for generalized zero-shot learning: $ACC_U$ denotes accuracy on unseen categories, $ACC_S$ denotes accuracy on seen categories, and their harmonic average $ACC_{\mathbf{H}} = \frac{2ACC_U ACC_S}{ACC_U + ACC_S}$.

### 5.2 Experiment Setting and Baselines

On CUB-Explanations dataset, we use a pretrained visual encoder to obtain image patch representation vectors. These vectors are then flattened as a sequence and used as visual input $X$. We use ResNet (He et al., 2016) as visual backbone for CLORE. For baselines, we make comparisons in two groups. The first group of models does not use parameters from pre-trained vision-language models (VLPMs). We adapt TF-VAEGAN (Narayan et al., 2020)[3], a state-of-the-art model on the CUB-200 zero-shot classification task, to use RoBERTa-encoded explanations as auxiliary information. This results in the baseline TF-VAEGAN$_{expl}$. The second group of models are those using pre-trained VLPMs. The main baseline we compare with is CLIP (Radford et al., 2021)[4], which is a well-performed pretrained VLPM. We build two of its variants: CLIP$_{linear}$, which only fine-tunes the final linear layer and CLIP$_{finetuned}$, which fine-tunes all parameters on the task. For fairer compasion, in this group we also replace the visual encoder with CLIP encoder in our model and get CLORE$_{CLIP}$.

### 5.3 Classification Results

Results are listed in Table 4 . On CUB-Explanations CLORE achieves the highest $ACC_U$ and $ACC_{\mathbf{H}}$ both with and without pre-trained vision-language parameters. Note that fine-tuning all parameters of CLIP makes it fit marginally better on seen classes, but sacrifices its generalization ability. Fine-tuning only the final linear layer (CLIP$_{linear}$) provides slightly better generalizability on unseen categories, but it is still lower than our approach.

## 6 Conclusions and Future Work

In this work, we propose a multi-modal zero-shot classification framework by logical parsing and reasoning on natural language explanations. Our method consistently outperforms baselines across modalities. We also demonstrate that, besides being interpretable, CLORE also benefits more from tasks that require more compositional reasoning, and is more robust against linguistic biases.

There are several future directions to be explored. The most intriguing one is how to utilize pre-trained generative language models for explicit logical reasoning . Another direction is to incorporate semantic reasoning ability in our approach, such as reasoning on entity relations or event roles.

[3] https://github.com/akshitac8/tfvaegan

[4] https://github.com/openai/CLIP

## Limitations

The proposed approach focuses more on logical reasoning on explanations for zero-shot classification. The semantic structures in explanations, such as inter-entity relations and event argument relations, are less touched (although the pre-trained language encoders such as BERT provides semantic matching ability to some extent). Within the range of logical reasoning, our focus are more on first-order logic, while leaving the discussion about higher-order logic for future work.

## Ethics Statement

This work is related to and partially inspired by the real-world task of legal text classification. As legal matters can affect the life of real people, and we are yet to fully understand the behaviors of deep-learning-based models, relying more on human expert opinions is still a more prudent choice. While the proposed approach can be utilized for automating the process of legal text, care must be taken before using or referring to the result produced by any machine in legal domain.


### Acknowledgements

We would like to thank anonymous reviewers for valuable comments and suggestions. This work was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.


## References


Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2015a. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438.

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. 2015b. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2927–2936.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Sahil Chopra, Michael Henry Tessler, and Noah D Goodman. 2019. The first crank of the cultural ratchet: Learning and transmitting concepts through language. In *CogSci*, pages 226–232.

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.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*.

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.

Chi Han, Jiayuan Mao, Chuang Gan, Josh Tenenbaum, and Jiajun Wu. 2019. Visual concept-metaconcept learning. *Advances in Neural Information Processing Systems*, 32.

Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 1884. NIH Public Access.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022a. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429.

Sha Li, Liyuan Liu, Yiqing Xie, Heng Ji, and Jiawei Han. 2022b. Piled: An identify-and-localize framework for few-shot event detection. *arXiv preprint arXiv:2202.07615*.

Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. 2021. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3794–3803.

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. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Gideon S. Mann and Andrew McCallum. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11(32):955–984.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*. International Conference on Learning Representations, ICLR.

Rakesh R Menon, Sayan Ghosh, and Shashank Srivastava. 2022. Clues: A benchmark for learning classifiers using natural language explanations. *arXiv preprint arXiv:2204.07142*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487.

Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. 2020. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision*, pages 479–495. Springer.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. 2014. Zero-shot learning by convex combination of semantic embeddings. In *2nd International Conference on Learning Representations, ICLR 2014*.

Roma Patel and Ellie Pavlick. 2021. "was it "stated" or was it "claimed"?: How linguistic bias affects generative language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10080–10095.

C Qiang, W Shu, H Yan, and W Liang. 2017. A hierarchical contextual attention-based gru network for sequential recommendation. *Neurocomputing*.

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 Samplawski, Erik Learned-Miller, Heesung Kwon, and Benjamin M Marlin. 2020. Zero-shot learning in the presence of hierarchically coarsened labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 926–927.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021. 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.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1527–1536.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2018. Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Tomasello. 2009. *The cultural origins of human cognition*. Harvard university press.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. *Computation and Neural Systems Technical Report*.

Ziqi Wang, Yujia Qin, Wenxuan Zhou, Jun Yan, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, and Xiang Ren. 2020a. Learning from explanations with neural execution tree. In *ICLR*.

Ziqi Wang, Yujia Qin, Wenxuan Zhou, Jun Yan, Qinyuan Ye, Leonardo Neves, Zhiyuan Liu, and Xiang Ren. 2020b. Learning from explanations with neural execution tree. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. 2016. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 69–77.

Guanyu Yang, Zihan Ye, Rui Zhang, and Kaizhu Huang. 2022. A comprehensive survey of zero-shot image classification: methods, implementation, and fair evaluation. *Applied Computing and Intelligence*, 2(1):1–31.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Ilker Yildirim, Judith Degen, Michael Tanenhaus, and Florian Jaeger. 2013. Linguistic variability and adaptation in quantifier meanings. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35.

Pengfei Yu, Zixuan Zhang, Clare Voss, Jonathan May, and Heng Ji. 2022. Building an event extractor with only a few examples. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 102–109.

Caleb Ziems and Diyi Yang. 2021. To protect and to serve? analyzing entity-centric framing of police violence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 957–976.

# A Appendix

## A.1 Configuration and Experiment Setting

We build CLORE on publicly available packages such as HuggingFace Transformers[5], where we used model checkpoints as initialization. We train CLORE for 30 epochs in all experiments. In the image classification task on CUB-Explanations, we adopt a two-phase training paradigm: in the first phase we fix both visual encoders and Explanation encoders in $E_\Phi$, and in the second phase we finetune all parameters in CLORE.

Across experiments in this work we use the AdamW (Loshchilov and Hutter, 2017) optimizer widely adopted for optimizing NLP tasks. For hyper-parameters in most experiments we follow the common practice of learning rate$= 3e-5$, $\beta_1 = 0.9, \beta_2 = 0.999$, $\epsilon = 1e-8$ and weight decay$= 0.01$. An exception is the first phase in image classification where, as we fix the input encoder, the learnable parameters become much less. Therefore we use the default learning rate$= 1e-3$ in AdamW. For randomness control, we use random seed of 1 across all experiments.

In Figure 7, there are multiple data points at $x$-value of 0. Therefore, the data variance on data at $x = 0$ is intrinsic in data, and is unsolvable theoretical for *any* function fitting the data series. This causes the problem when calculating $R^2$ value, as $R^2$ measures the extent to which the data variance are "explained" by the fitting function. So $R^2$ can be upper bounded by: $R^2 \leq 1 - \frac{Var_{intrinsic}}{Var_{total}}$. To deal with this problem when measuring $R^2$ metric, we removed the intrinsic variance in data point set $D$ by replacing data points $(0, y_i) \sim D$ with $(0, \frac{1}{n}\sum_{(0,y_i)\sim D} y_i)$ in both series in Figure 7 before calculating $R^2$ value.

## A.2 Logical Structure Templates

As the number of valid logical structure templates grows exponentially with maximal attribute numbers $T$, we limit $T$ to a small value, typically 3. We list the logical structure templates in Table 5.

| |
|---|
| label$(X)$ = attribute$_1(X)$ |
| label$(X)$ = attribute$_1(X) \wedge$ attribute$_2(X)$ |
| label$(X)$ = attribute$_1(X) \vee$ attribute$_2(X)$ |
| label$(X)$ = attribute$_1(X) \wedge$ attribute$_2(X) \wedge$ attribute$_3(X)$ |
| label$(X)$ = attribute$_1(X) \vee$ attribute$_2(X) \vee$ attribute$_3(X)$ |
| label$(X)$ = (attribute$_1(X) \wedge$ attribute$_2(X)) \vee$ attribute$_3(X)$ |
| label$(X)$ = (attribute$_1(X) \vee$ attribute$_2(X)) \wedge$ attribute$_3(X)$ |

Table 5: The list of logical structure templates at maximum attribute number $T = 3$.

## A.3 Resources

We use one Tesla V100 GPU with 16GB memory to carry out all the experiments. The training time is 1 hour for tabular data classification on CLUES, 2 hours for image classification on CUB-Explanations.

[5] https://huggingface.co, https://github.com/huggingface/transformers

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section "Limitations", page 10*

☑ A2. Did you discuss any potential risks of your work?
*Section "Ethics Statement", page 10*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Section "Abstract" and Section 1 Introduction, page 1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☒ Did you use or create scientific artifacts?

*Left blank.*

☐ B1. Did you cite the creators of artifacts you used?
*No response.*

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*No response.*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*No response.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*No response.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*No response.*

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*No response.*

### C ☑ Did you run computational experiments?

*Section 4 and 5*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Section A.3*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section A.1*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 4, Figure 8*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Section 4, 5, and A.1*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*No response.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*No response.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*No response.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No response.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*No response.*