Multi-Label Intent Detection via Contrastive Task Specialization of Sentence Encoders

Ivan Vulić, Iñigo Casanueva, Georgios Spithourakis, Avishek Mondal, Tsung-Hsien Wen and **Paweł Budzianowski**

PolyAI Limited
London, United Kingdom
poly.ai

Abstract

Deploying task-oriented dialog (ToD) systems for new domains and tasks requires natural language understanding models that are **1)** resource-efficient and work under low-data regimes; **2)** adaptable, efficient, and quick-to-train; **3)** expressive and can handle complex ToD scenarios with multiple user intents in a single utterance. Motivated by these requirements, we introduce a novel framework for multi-label intent detection (mID): MULTI-CONVFIT (**Multi**-Label **I**ntent Detection via **Con**trastive **Conv**ersational **Fi**ne-**T**uning). While previous work on efficient single-label intent detection learns a classifier on top of a fixed sentence encoder (SE), we propose to **1)** transform general-purpose SEs into task-specialized SEs via contrastive fine-tuning on annotated multi-label data, **2)** where task specialization knowledge can be stored into lightweight adapter modules without updating the original parameters of the input SE, and then **3)** we build improved mID classifiers stacked on top of fixed specialized SEs. Our main results indicate that MULTI-CONVFIT yields effective mID models, with large gains over non-specialized SEs reported across a spectrum of different mID datasets, both in low-data and high-data regimes.

1 Introduction

Task-oriented dialog (ToD) systems allow users to interact with computer applications through conversation in order to solve a particular task with well-defined semantics (Levin and Pieraccini, 1995; Young, 2010). ToD supports a multitude of applications such as automating different customer service tasks, facilitating bookings in hospitality and travel industries, or providing assistance in healthcare or finance (Raux et al., 2003; El Asri et al., 2017; Xu et al., 2017; Budzianowski et al., 2018).

Intent detection (ID), a task of recognizing the user's intent or goal from their utterance, is a crucial component of any ToD system (Hemphill et al., 1990; Tür et al., 2010; Coucke et al., 2018). Intent detectors that adhere to industry standards should satisfy the following three requirements (Casanueva et al., 2020; Larson and Leach, 2022).[1]

(R1) Scalability and resource efficiency. They must be quickly bootstrapped for new domains and tasks. However, this process requires creating expensive annotations for each domain and task of interest, which calls for sample-efficient ID methods that achieve strong performance in *low-data scenarios*.

(R2) Lightweight design and modularity. While large laguage models (LMs) have shown strong performance in the ID task, running full-fledged fine-tuning and storing separate fine-tuned models per each domain or task yields prohibitive storage and memory costs (Ding et al., 2022). Enabling fast training of intent detectors (Casanueva et al., 2020) also speeds up ToD development cycles.

(R3) Expressiveness: supporting complex ToD *scenarios.* Previous work has typically focused on more limited single-label ID scenarios (Liu et al., 2019a; Larson et al., 2019; Wu et al., 2020; Mehri et al., 2020, *inter alia*). Such setups are not realistic in more complex industry settings and even lead to limited and simplified conversational scenarios with artificial intent sets. Intent detectors should thus tackle the more challenging *multi-label ID (mID)* task, which enables more complex 'real-life' natural language understanding for ToD (Qin et al., 2021; Hou et al., 2021; Casanueva et al., 2022).

In this work, we propose a novel framework for mID, termed MULTI-CONVFIT (**Multi**-Label **I**ntent Detection via **Con**trastive **Conv**ersational **Fi**ne-**T**uning), that satisfies the three requirements R1-R3. The framework's pipeline is illustrated in Figure 1. Previous work typically used fixed general-purpose (Henderson et al., 2020; Casanueva et al., 2020) sentence encoders (SEs) (Cer et al., 2018) combined with tunable intent

[1] See also poly.ai/modular-intent-design/.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7544–7559
December 7-11, 2022

classifiers for efficient single-label ID. In this work, we propose a modular framework that contrastively fine-tunes general-purpose encoders, using mID data annotations implicitly, to yield *task-specialized sentence encoders*. All the task-specific 'adaptation' knowledge after contrastive fine-tuning can be injected into small *adapter modules*. (Houlsby et al., 2019; Pfeiffer et al., 2020a).

Such adapter modules are then used to adapt the underlying general-purpose SE which already stores plenty of useful semantic knowledge – a single large model that serves multiple tasks and domains – into the task-specialized SE. The contrastive procedure creates a semantic space which better aligns with intent classes of a particular mID task, as demonstrated in Figure 2. Consequently, such fixed task-specialized SEs enable learning improved mID classifiers that outperform mID classifiers learnt on top of the original general SEs.

Our key results indicate effectiveness and robustness of MULTI-CONVFIT, yielding state-of-the-art results across four representative mID datasets, both in *low-data* and *high-data* scenarios, while offering modularity and efficient fine-tuning and inference. Additional analyses indicate MULTI-CONVFIT's versatility and wide applicability: it can be used with a range of pretrained SEs and LMs, and it leads to gains across different domains, dataset sizes, and intent set sizes.

2 Methodology

The full overview of MULTI-CONVFIT, aiming to satisfy the requirements R1-R3 from §1, is provided in Figure 1. In what follows, we discuss its main components in detail: contrastive task-specialization of general-purpose input encoders using annotated mID data (§2.1); a classifier for multi-label ID stacked on top of the fixed encoder (§2.2); and a more efficient variant of the framework which combines contrastive tuning with adapters (§2.3).

Preliminaries. For any input text t, we obtain its encoding $\mathbf{t} = f(t)$, where f is an encoding function of any input encoder model (i.e., LM, general-purpose or task-specialized SE). t is tokenized into subwords using each encoder's dedicated tokenizer. The final encoding $\mathbf{t}$ is created via a *pooling* operation such as (a) using the [CLS] token as the text representation, (b) or mean-pooling the output subword encodings. Following prior work (Reimers and Gurevych, 2019; Liu et al., 2021), we opt for mean-pooling as a better-performing option.

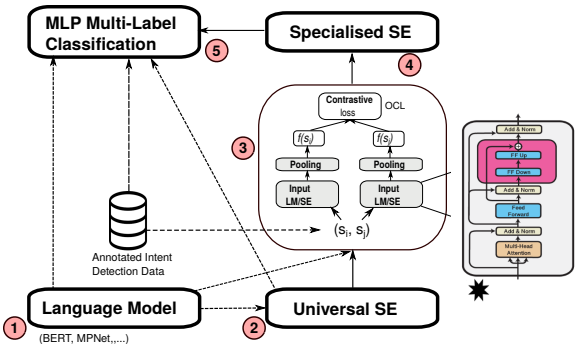


Figure 1: Overview of the full MULTI-CONVFIT framework for efficient and effective multi-label intent detection (mID), described in §2. Large language models (LMs) (①) can be transformed into universal (i.e., general-purpose) sentence encoders (SEs) (②), as done in prior work (Reimers and Gurevych, 2019; Feng et al., 2022). A contrastive fine-tuning procedure (③) can be applied on any input LM (①) or any input SE (②), relying on intent labels from the annotated mID task data. This fine-tuning yields a task-specialized SE (④). A multi-label MLP classifier (⑤) is then learnt, leveraging the same mID data as contrastive fine-tuning, on top of the sentence encodings obtained via the fixed task-specialized SE. Instead of contrastively fine-tuning the full input model, efficient *adapter modules* (Pfeiffer et al., 2020a), inserted into the input model's Transformer layers (✷, on the right), get contrastively tuned into specialized mID task adapters, while the input model remains unchanged. The MLP classifier is then trained on top of the fixed model with the fixed task adapters. The classifier can also be learnt directly (i.e., without the contrastive step) using fixed input LMs (①) or general-purpose SEs (②) as text encoders, shown with dashed lines. At inference, the trained classifier is directly applied on the encoding of any input sentence.

Further, we assume that $|\mathcal{S}|$ annotated mID data examples are available: they comprise a set of pairs $\mathcal{S} = \{(s_1, L_1), \ldots, (s_i, L_i), \ldots, (s_{|\mathcal{S}|}, L_{|\mathcal{S}|})\}$, where s_i are sentences/examples, each annotated with a set of $L_i = \{l_{i,1}, \ldots, l_{i,M_i}\}$ labels, where $M_i \geq 0$. Each label l is in fact one of the $|\mathcal{C}|$ intent classes from the set $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$.

2.1 Contrastive Specialization

Motivation. The main idea is to specialize the input general-purpose encoder relying on available ID annotations so that the encoder better aligns with the actual *ID task semantics*. Such specialization of general-purpose encoders has been proven effective in prior work on single-label nonparametric ID (Zhang et al., 2020, 2021; Mehri et al., 2021; Vulić et al., 2021). Whereas the 'ID task seman-

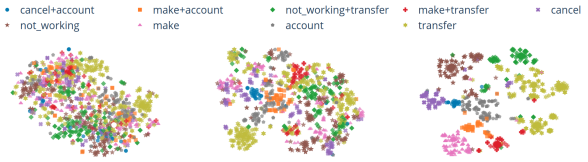


Figure 2: t-SNE plots (van der Maaten and Hinton, 2012) of encoded utterances from the mID dataset BANKING (see §3) associated with a subset of intent classes, demonstrating the effects of contrastive task specialization of the input encoder with mID data. **Left**: sentence encodings with the original 12-layer MiniLM (Wang et al., 2020); **Middle**: encodings with MiniLM transformed into a universal SE; **Right**: encodings with a task-specialized SE obtained after contrastively specializing (C-ADAPT, see §2.3) the universal MiniLM-based SE. See also Figure 7 in Appendix C.

tics' in the single-label ID scenario is clear-cut (i.e., the encoder should create coherent clusters of sentences annotated with the same *single* intent), this is not the case in more general multi-intent setups (Bi and Kwok, 2013; Qin et al., 2020). However, our hypothesis is that task-adaptive contrastive fine-tuning can still yield a task-specialized SE. This task-specialized sentence encoder should create more accurate encodings than the original universal encoder (Figure 2) for task-relevant sentences (i.e., user utterances), which should in turn help learn an improved intent classifier, stacked on top of the fixed sentence encoder (Casanueva et al., 2020, 2022; Vulić et al., 2021).

(Creating) Positive and Negative Pairs. Similar to other contrastive methods for various single-label classification tasks (Zhang et al., 2021; Gunel et al., 2021; Vulić et al., 2021), we leverage the available intent class labels only implicitly, that is, we utilize them to create sets of positive and negative learning examples. However, here we generalize the creation of such pairs for contrastive learning to the multi-label setup.

The sets of pairs in the MULTI-CONVFIT framework are constructed as follows. **1)** *PosP* is the set of positive example pairs (s_i, s_j), where the pair is added to the set if and only if s_i and s_j share at least one intent label, that is, when it holds $L_i \cap L_j \neq \varnothing$. We add (s_i, s_j) and (s_j, s_i) as separate pairs into *PosP*. **2)** The second set, *NegP*, comprises negative pairs (s_i, s_j), where s_i and s_j do not share any intent class at all, that is, $L_i \cap L_j = \varnothing$. We construct the set *NegP* in a balanced way: for each positive pair $(s_i, s_j) \in PosP$, we add n negative pairs $(s_i, s_{i,n'}), n' = 1, \ldots, n$, into *NegP*, where n is a tunable hyper-parameter. All negative instances $s_{i,n'}$ are randomly sampled from the set $\mathcal{S}$.

Contrastive Loss. We opt for the standard margin-based Online Contrastive Loss (OCL) (Reimers and Gurevych, 2019). It relies on the following standard formulation of the Contrastive Loss from Hadsell et al. (2006), which operates over the two sets of pairs *PosP* and *NegP*:

$$\mathcal{L}_{\text{OCL}} = \mathbb{1} \cdot (cosd(\mathbf{s_i}, \mathbf{s_j}))^2 + (1 - \mathbb{1}) \cdot \Big(ReLU(\delta_m - cosd(\mathbf{s_i}, \mathbf{s_j}))\Big)^2.$$

$\mathbb{1}$ is the indicator function which returns 1 iff $(s_i, s_j) \in PosP$, and 0 iff $(s_i, s_j) \in NegP$; $cosd$ is the cosine distance, and δ_m is the distance margin, set to the default value of 0.5 (Reimers and Gurevych, 2019) in all our experiments. OCL should 'attract' similar items closer together in the task-specialized space, where they should be closer by at least the margin δ_m than all other, dissimilar items (Mrkšić et al., 2017). We leverage the *online* version of the loss which operates with hard in-batch negative pairs (i.e., negatives that are close by cosine in the current semantic space) and hard in-batch positive pairs (i.e., positives that are far apart in the current space).[2] This procedure yields a task-specialized encoding function f.

Why Contrastive Specialization? The rationale is to enable the encoder to focus on parts of the sentences that yield shared labels/intent(s), and reshape the semantic space so that sentences with a larger proportion of shared intents end up more similar in the fine-tuned space. For instance, imagine a toy scenario with three classes c_x, c_y, c_z: the procedure should cluster together all single-label examples (i.e., all c_x examples should be grouped together, and another two coherent clusters are c_y and c_z examples). At the same time, all two-label examples should also create coherent clusters, and examples labeled with c_x and c_y should end up closer to the single-label c_x and c_y clusters than to the c_z cluster, etc. This effect is indeed observed with real mID data, as plotted in Figure 2. For instance, we observe that sentences labeled with intents *(cancel, account)* are indeed in encoded in a

[2] The OCL loss, among other tasks, demonstrated strong performance in single-label ID tasks in prior work (Vulić et al., 2021). The 'online' formulation typically results in quicker convergence and better performance, also confirmed in our preliminary experiments. See also www.sbert.net/docs/package_reference/losses.html. Future work will delve deeper into experimenting with other contrastive losses.

subspace between the clusters of '*cancel*-only' and '*account*-only' sentences, while the sentences with *(make, account)* are encoded in another cluster between '*make*-only' and '*account*-only' sentences.

2.2 Multi-Label Classifier

A standard approach to efficient intent detection in single-label scenarios is to stack a classifier on top of a fixed sentence encoder (Casanueva et al., 2020; Gerz et al., 2021). While it is much more lightweight than fine-tuning the entire SE (Mehri et al., 2020), this approach typically yields on-par performance in single-label ID tasks (Casanueva et al., 2020). Following prior work, our classifier is a standard Multi-Layer Perceptron (MLP), stacked on top of the fixed sentence representations $f(s)$, which were previously obtained with any *fixed* input encoder (see Figure 1). The MLP classifier comprises a single hidden layer with non-linearity, followed by a *sigmoid* layer to allow for multi-label classification. It is trained via standard binary cross-entropy loss. A tunable threshold θ determines the final classification: only intents with their probability scores $\geq \theta$ are taken as positives. This way, the threshold θ effectively controls the trade-off between precision and recall of the classifier.

Label Smoothing. In contrast to prior work in single-label ID setups, we propose to add *label smoothing* (Müller et al., 2019) into classifier training, and later validate its impact on mID performance. This label smoothing regularization should mitigate overfitting and classification overconfidence in low-data setups (Bai et al., 2021), where such overconfidence might get even more pronounced with contrastively specialized encoders. Since the label smoothing technique has not been used in prior work on ID and mID, we provide a full description in what follows.

We leverage a standard label smoothing technique, additionally 'corrected' for multi-label classification (Hou et al., 2019). Without label smoothing, for the item $(s_i, L_i) \in \mathcal{S}$ the conversion of L_i into a $|\mathcal{C}|$-dimensional vector $Y_i = [y_{i,1}, \ldots, y_{i,|\mathcal{C}|}]$ of binary labels assigns 1-s to labeled classes from $\mathcal{C}$, and 0-s otherwise. Adding label smoothing with the value ls then means reassigning all the individual binary indicators y-s to the following y' values:

$$y'_{i,k} = \begin{cases} ls & \text{if } y_{i,k} = 1 \\ \frac{(1-ls)\cdot M_i}{|\mathcal{C}|} & \text{otherwise.} \end{cases} \quad (1)$$

M_i is the number of positive labels for the example s_i (i.e., the number of 1-s in Y_i). This reassigns some of the probability mass from the positive labels to the negative ones, this way reducing the classifier's (over)confidence (Pereyra et al., 2017).

2.3 Sentence Encoders with Adapters

We always learn the classifier on top of fixed sentence encoders. However, the contrastive task specialization must adapt the weights of the general-purpose input encoder. A standard variant, termed **C-FFT**, does full fine-tuning of all the weights, meaning that a separate full copy of the specialized model must be stored per each specialization.

However, with multiple dialog domains and tasks, requirements such as model compactness, fine-tuning and storage efficiency become crucial features; see again the main requirements listed in §1. We thus propose to combine fine-tuning of general-purpose SEs with lightweight tunable adapter modules (Houlsby et al., 2019; Pfeiffer et al., 2021). Such adapters, whose size is typically only a fraction of the size of the full input neural model, are inserted within each Transformer layer of the underlying model. At fine-tuning, only adapter parameters are updated while all the other parameters of the large model are kept fixed, which enables parameter-efficient and modular adaptation of large neural models (He et al., 2022).

Unlike prior work which typically combined adapters with off-the-shelf large LMs (Pfeiffer et al., 2020a; He et al., 2022), here we focus on inserting adapter modules directly into general-purpose sentence encoders. We create small *task-specialized modules* (Madotto et al., 2020; Pfeiffer et al., 2020b) that transform a large general-purpose SE into a particular domain- or task-specialized SE without full fine-tuning. Since a single general-purpose SE can serve multiple domains and tasks without incurring catastrophic forgetting and interference (McCloskey and Cohen, 1989; Hashimoto et al., 2017), this approach increases modularity and decreases storage demands. We note that MULTI-CONVFIT can be directly applied to the single-label ID scenario (Mehri et al., 2021; Vulić et al., 2021; Zhang et al., 2021) as a special case. The efficient adapter-based SE tuning variant, illustrated in Figure 1, is dubbed **C-ADAPT**.

3 Experimental Setup

In what follows, we outline our experimental setup, focused on evaluating and improving performance

Dataset	Domain	Abbreviation	# of Intents	# of Examples	Avg. Intents per Example
NLU++ (Casanueva et al., 2022)	e-banking	BANKING	48	2,071	2.25
NLU++ (Casanueva et al., 2022)	hotel reservations and FAQ	HOTELS	40	1,009	1.52
(internal)	insurance FAQ	INSURANCEFAQ	118	4,356	1.91
MixATIS (Qin et al., 2020)	flight info	MIXATIS	18	18k/1k/1k	2.19/2.12/2.07

Table 1: Multi-label intent detection datasets in our experiments with key statistics. MIXATIS is the only dataset with a standardized *train/dev/test* split, whereas on the other datasets we run suggested 10-fold experiments; see (Casanueva et al., 2022) and §3 for further details.

and sample-efficiency of multi-label intent detectors, relying on the standard mID benchmarks.

Input Sentence Encoders. We experiment with several representative and popular sentence encoders as input, aiming to (i) validate the robustness of the proposed methodology across different underlying encoders, as well as to (ii) analyze the impact of the chosen encoder on the final mID task performance. We opt for the following SEs that offer a good trade-off between model size and performance in sentence-level semantic tasks (Reimers and Gurevych, 2019). **1)** MLM12, **2)** MPNET, **3)** DROB are sentence encoders which transform the respective pretrained LMs – the 12-layer MiniLM (Wang et al., 2020), MPNet (Song et al., 2020), and DistilRoBERTa (Sanh et al., 2019)[3] – into SEs via standard contrastive dual-encoder (i.e., bi-encoder) frameworks (Reimers and Gurevych, 2019; Henderson et al., 2020).[4]

MLM12 (its size is 120 MB) comprises $L_T = 12$ Transformer layers, with hidden size $h_T = 384$; $L_T = 12$ and $h_T = 768$ for MPNET (490 MB); $L_T = 6$ and $h_T = 768$ for DROB (290 MB). All SEs have been obtained in prior work (Reimers and Gurevych, 2019) via contrastively fine-tuning their underlying LMs on a set of more than 1B sentence pairs, which comprises various data such as Reddit 2015-2018 comments (Henderson et al., 2019), Natural Questions (Kwiatkowski et al., 2019), PAQ (question, answer) pairs (Lewis et al., 2021), etc.[5]

Furthermore, in order to test the impact of the chosen input/underlying model (i.e., sentence encoders versus language models) we also contrastively fine-tune the original LMs instead of their SE counterparts (see Figure 1 again): we refer to the respective input LMs as MLM12-LM, MPNET-LM, and DROB-LM.

Multi-Label ID Datasets, until very recently, have been few and far between (Casanueva et al., 2022), as prior ID research predominantly focused on single-label scenarios (Larson and Leach, 2022). We experiment with a representative set of multi-label ID datasets, covering **1)** four diverse domains, **2)** ontologies with different sizes of the intent sets (from 18 to 118 intents), **3)** varied dataset size, and **4)** different average number of intents per example. A complete summary is provided in Table 1.

MIXATIS (Qin et al., 2020) is the only dataset whose 'multi-label nature' was achieved synthetically through concatenation of single-label examples from the original ATIS dataset (Hemphill et al., 1990). On the other hand, the other datasets in our evaluation are natively multi-label, relying on the concept of combining the so-called *intent modules*; see the work of Casanueva et al. (2022) for further details. Some examples of multi-label sentences from each mID dataset are provided in Appendix B.

Low-Data versus High-Data Setups. Due to the high cost of task-specific annotations, prior work on single-label ID has recognized the importance of building and bootstrapping intent detectors in low-data regimes (Casanueva et al., 2020; Mehri et al., 2021; Vulić et al., 2021), and the same naturally holds also for multi-label ID. In order to understand the behaviour of MULTI-CONVFIT in such scenarios versus setups with more abundant task-annotated data, we conduct experiments in two standard data setups: **1)** *low-data* and **2)** *high-data*.

For the BANKING, HOTELS, and INSURANCE-FAQ datasets (see Table 1), we adopt the standard 10-fold cross-validation (Casanueva et al., 2022). Then, in *low-data* setups we use 1 fold as our training data for contrastive fine-tuning *and* MLP train-

[3] MiniLM and DistilRoBERTa are more compact distilled versions of the standard BERT-base (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019b) LMs, respectively.

[4] See `www.sbert.net/docs/pretrained_models.html`, Appendix A, and the relevant SE and LM model cards in the HuggingFace Transformers repository (Wolf et al., 2020) for further details on architectures and training data.

[5] In a nutshell, the contrastive task is formulated as follows: given a 'query' sentence from each sentence pair, and a set of R randomly sampled negatives plus 1 true positive (the sentence from the same pair), the model should predict which sentence from the set of $R + 1$ sentences is actually paired with the query sentence in the dataset; see e.g. `huggingface.co/sentence-transformers/all-mpnet-base-v2` for the full overview of the fine-tuning data and setup.

ing (see Figure 1), and test the model on the remaining 9 folds. The *high-data* setups are effectively the same, but with swapped training and test data: we now use merged data from 9 folds as training data, and test on the single held-out fold. All the reported scores are averages over all folds. The *folding* evaluation comes with several benefits: **1)** we avoid overfitting to any particular test set; **2)** we reach more stable results with smaller training and test data (i.e., simulating low-data regimes typically met in production) through averaging over different folds; **3)** variations in results due to potentially different random seeds are reduced.

For MIXATIS, we leverage its development portion for our *low-data* experiments, and its training portion for the *high-data* setup (without leveraging development data for model selection; see next for our hyper-parameter selection procedure).

Contrastive Specialization Setup. Following the suggested settings (Reimers and Gurevych, 2019; Vulić et al., 2021), we use the AdamW optimizer (Loshchilov and Hutter, 2018). The learning rate for the C-FFT variants is set to the standard value of $2e$-5, while a higher learning rate of $4e$-4 is used for the adapter-based C-ADAPT variants. For C-ADAPT, we opt for a standard efficient *bottleneck* adapter configuration, following Pfeiffer et al. (2021): ReLU activation (Nair and Hinton, 2010), with the adapter reduction factor of 4.[6]

The warmup rate of 0.1 with cosine decay is used; weight decay rate is 0.02. We fine-tune for 10 epochs in *low-data*, and for 3 epochs in *high-data* setups, with the number of negatives n=2;[7] batch size is 32, and max sequence length is 128.

Classifier Setup. We adopt the MLP classifier architecture from Casanueva et al. (2020): it contains 1 hidden layer of size 512 with ReLU as non-linear activation. The values for the hyperparameters were largely adopted from prior work (Reimers and Gurevych, 2019; Casanueva et al., 2020; Vulić et al., 2021), both for contrastive fine-tuning and MLP training. We further fine-tuned them relying solely on one randomly sampled fold (Fold 5) from BANKING with MLM12 and MLM12-LM in the *low-data* setup, and applied the same hyperparameters across all other models, setups, and runs. The classifiers' dropout rate is fixed to 0.4, and the threshold θ is fixed to 0.3 in all runs. We again train with AdamW, with the standard warmup rate of 0.1, weight decay of 0.02, and the learning rate is set to 0.003; 600 epochs with the batch size of 32. Unless noted otherwise, we always apply label smoothing with $ls = 0.95$.

Evaluation Details. All the reported scores are averaged across three runs with three different random seeds. We report standard ID evaluation metrics: F_1 and exact match accuracy (*Acc*).

4 Results and Discussion

The main results are summarized in Table 2 and Figure 3, while further results and analyses are available in §4.1, with additional results in Appendix C. These results offer multiple axes of comparison and analysis, discussed in what follows.

Impact of Contrastive Task Specialization. First, the results clearly demonstrate substantial and consistent gains achieved via contrastive task-specialization. The strong performance boosts are present across the board, and span all input encoders (including both SEs and LMs), both fine-tuning variants (C-FFT and C-ADAPT), all mID datasets, both *low-data* and *high-data* scenarios, and also all groups of examples with a different number of intents per example (see also Figure 9 in Appendix C). While prior work has proven that even general-purpose SEs can support effective and efficient (single-label) ID (Casanueva et al., 2020; Gerz et al., 2021; Zhang et al., 2021), here we demonstrate that **1)** the efficient SE-based approach is also beneficial for the multi-label ID task, and **2)** large performance improvements are achieved by transforming/adapting such general-purpose SEs into task-specialized sentence encoders.

C-FFT versus C-ADAPT. Importantly, the comparison in Figure 3 validates that the efficient C-ADAPT variant maintains strong performance across the board, offering on-par or even slightly improved performance across different setups. It indicates that task-specialized modules can be combined with large sentence encoders to obtain their task specialization. The strong performance with C-ADAPT is maintained over both data setups and using different input SEs. The C-ADAPT variant in

[6] See, e.g., (Pfeiffer et al., 2020a) for the definition of the reduction factor. When combined with the MLM12 SE, this adapter config requires only 3.5 MB additional parameters for each task specialization of the base MLM12 model.

[7] We also experimented with higher n values, which substantially increase fine-tuning time while offering diminishing/negligible performance gains in the mID task in our preliminary experiments. A similar finding for single-label ID scenarios was reported by Vulić et al. (2021).

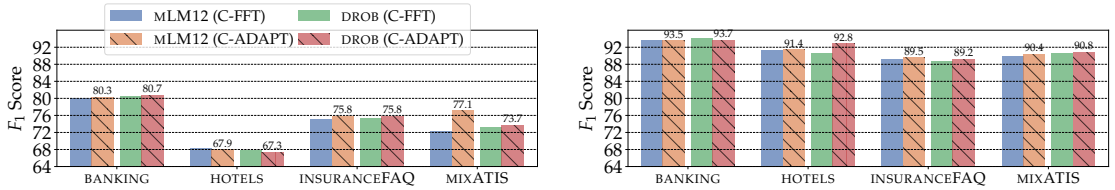
SE↓ / ID Dataset→	BANKING		HOTELS		INSURANCEFAQ		MIXATIS	
	low-data	*high-data*	*low-data*	*high-data*	*low-data*	*high-data*	*low-data*	*high-data*
MLM12	70.8 / 31.2	86.7 / 58.1	64.2 / 46.0	80.3 / 65.8	64.8 / 33.2	85.2 / 64.0	58.1 / 22.0	78.6 / 47.5
MLM12 +C-FFT	**80.1 / 46.6**	**93.6 / 79.5**	**68.2 / 47.6**	**91.3 / 80.2**	**75.2 / 47.1**	**89.2 / 74.1**	**72.2 / 37.3**	**89.9 / 77.9**
DROB	70.6 / 31.0	86.7 / 60.0	65.3 / 46.8	80.9 / 65.1	64.6 / 32.8	85.3 / 64.1	58.5 / 21.2	78.4 / 46.9
DROB +C-FFT	**80.4 / 47.6**	**94.0 / 78.0**	**67.9 / 47.9**	**90.5 / 79.2**	**75.4 / 47.2**	**88.8 / 71.9**	**73.1 / 44.8**	**90.6 / 77.5**
MPNET	70.5 / 30.0	85.8 / 59.5	64.7 / 47.5	79.9 / 64.2	64.7 / 32.2	85.6 / 64.8	58.6 / 21.6	77.6 / 49.6
MPNET +C-FFT	**81.9 / 49.1**	**94.3 / 80.5**	**70.2 / 51.1**	**93.4 / 84.0**	**77.3 / 49.9**	**90.1 / 75.1**	**76.5 / 51.4**	**91.5 / 81.1**

(a) Results in the multi-label ID task. Contrastive fine-tuning starts from (general-purpose) sentence encoders (SEs).

LM↓ / ID Dataset→	BANKING		HOTELS		INSURANCEFAQ		MIXATIS	
	low-data	*high-data*	*low-data*	*high-data*	*low-data*	*high-data*	*low-data*	*high-data*
DROB-LM	65.9 / 27.2	86.0 / 59.5	54.7 / 39.1	79.4 / 64.2	60.8 / 30.6	84.3 / 63.1	56.2 / 18.9	77.9 / 44.9
DROB-LM +C-FFT	**75.4 / 39.9**	**93.1 / 76.1**	**62.1 / 43.9**	**88.6 / 75.5**	**71.2 / 42.0**	**88.8 / 71.8**	**70.8 / 37.1**	**90.1 / 77.4**
MPNET-LM	62.2 / 23.9	84.9 / 56.6	54.6 / 37.5	78.2 / 59.4	58.2 / 27.4	81.6 / 57.9	56.0 / 20.5	79.6 / 49.0
MPNET-LM +C-FFT	**76.4 / 41.3**	**94.2 / 80.4**	**66.2 / 47.8**	**89.9 / 79.3**	**72.9 / 44.6**	**89.8 / 74.3**	**75.6 / 44.2**	**90.2 / 80.3**

(b) Results in the multi-label ID task. Contrastive fine-tuning starts from (general-purpose) language models (LMs).

Table 2: F_1 ($\times 100\%$) and exact match Accuracy scores (Acc; $\times 100\%$), in the format F_1/Acc, in the multi-label ID task with *full model fine-tuning* via supervised contrastive learning (+C-FFT). **(a)** C-FFT starts from a sentence encoder; **(b)** C-FFT starts from a language model (results with MLM12-LM, following the same trends, omitted for brevity). **Bold** numbers indicate a better-scoring configuration per each individual SE or LM architecture, whereas underlined numbers denote the best overall performance in each column (which includes both sub-tables).



(a) *low-data* scenarios (b) *high-data* scenarios

Figure 3: Comparison of full-model (C-FFT) and adapter-based (C-ADAPT) contrastive fine-tuning, demonstrating the competitive performance of much more parameter-efficient C-ADAPT. F_1 scores shown; see also Appendix C.

fact allows for building efficient, high-performing and modular multi-label intent detectors, satisfying the motivating requirements from §1.

Here, we use standard bottleneck adapters, but we believe that it is possible to strike an even better trade-off between parameter-efficiency and mID performance. A further exploration of and efficiency optimization with different adapter configurations (Pfeiffer et al., 2020a; He et al., 2022), including more efficient variants (Li and Liang, 2021; Mahabadi et al., 2021; Ansell et al., 2021, 2022), is beyond the scope of this paper, and we leave it for future work.

Input Encoders. As expected, the choice of the input encoder also impacts final mID performance. First, we mark that starting from SEs yields higher performance than starting from their LM-based counterparts (e.g., MPNET +C-FFT outperforms MPNET-LM +C-FFT, and the same holds with other models and fine-tuning variants). The gap is substantial in *low-data* setups, and it also exists even in *high-data* setups. This finding suggests the usefulness of conducting the adaptive fine-tuning step (Mehri et al., 2019; Henderson et al., 2020; Ruder, 2021), transforming LMs into general-purpose SEs through more suitable objectives such as response selection and paraphrase detection. Our finding corroborates a similar result in single-label ID scenarios (Vulić et al., 2021). Contrastively fine-tuning LMs with task-annotated mID data does yield large benefits in the mID task, but they cannot reach performance peaks of SEs as starting encoders.

A comparison of different input SEs reveals that the SE with the highest capacity (MPNET) yields highest absolute scores across the board. However, even the most lightweight input SE (MLM12) displays very competitive performance in all the experiments, also with the efficient C-ADAPT specialization variant. Similarly, when we start from LMs instead of SEs, MPNET-LM has a slight edge over DROB-LM. Again, we stress that applying the MULTI-CONVFIT specialization yields large

benefits regardless of the starting input encoder.

Low-Data versus High-Data Setups. MULTI-CONVFIT yields performance boosts in both data setups. As expected, absolute improvements with contrastive learning are higher in *low-data* setups (e.g., +12.6 F_1 in *low-data* versus +4.5 in *high-data* with MPNET +C-FFT on INSURANCEFAQ; +9.5 versus 6.9 with MLM12 +C-ADAPT on BANKING). However, we observe prominent boosts even in *high-data* setups with several thousand annotated instances (e.g., more than 4k for INSURANCEFAQ), rendering task specialization of SEs as universally useful for multi-label intent detection.

What is more, while the primary focus of MULTI-CONVFIT is the trade-off of performance and efficiency, the results in *high-data* setups on BANKING, HOTELS, and MIXATIS are current state-of-the-art results on all these datasets. The scores on BANKING increase from F_1 of 93.0 (Casanueva et al., 2022) to 94.3, while the previous high score on HOTELS of 86.7 is superseded by F_1 scores from Table 2 and Figure 3, reaching up to 93.4 F_1. The previous high scores were obtained via QA-based intent models (Namazifar et al., 2021; Casanueva et al., 2022) which require much more computationally demanding and slower training and inference regimes. The previous best-reported results on MIXATIS (*high-data*) are F_1 of 81.2 (Qin et al., 2020) and Acc of 76.3 (Qin et al., 2021), while we report respective peak scores of 91.5 and 81.1.[8]

In brief, our results illustrate the important aspect of sample efficiency of the MULTI-CONVFIT framework. On top of offering better scores in *high-data* setups, it also allows for reaching strong mID performance relying on smaller amounts of the most 'precious' resource: annotated task data (cf. the scores in *low-data* scenarios).

4.1 Further Discussion

We now analyze other important aspects of MULTI-CONVFIT, running a series of side experiments. Due to a large number of experiments and to avoid clutter, we plot results from a representative subset of possible experimental configurations (i.e., encoders, fine-tuning variants, datasets, data setups), but we note that very similar patterns in results have been observed with other configurations.

[8]Moreover, the previous state-of-the-art models on MIXATIS are also less lightweight than MULTI-CONVFIT: they model ID and the slot labelling task in a joint framework (Chen et al., 2019b), effectively leveraging additional annotations available for another task, and performing model selection and hyper-parameter search on the dedicated development set.

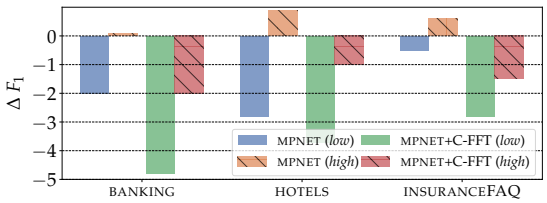


Figure 4: Change in F_1 performance when no label smoothing is used versus the standard variant with label smoothing (ΔF_1 on the y-axis), with all other parts kept equal. Similar trends are observed with other input models and with C-ADAPT, and also with Acc scores (see Appendix C).

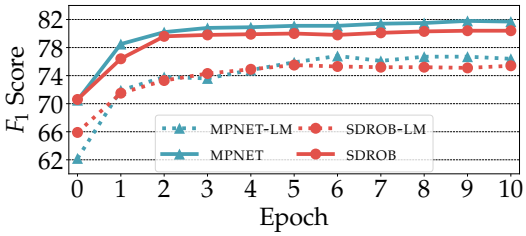
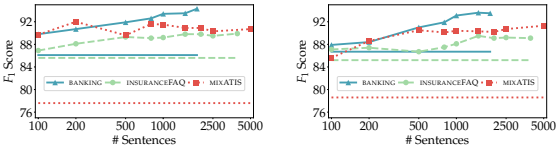


Figure 5: Impact of contrastive specialization duration (i.e., the number of epochs) on the final mID performance in *low-data* scenarios on BANKING. F_1 scores; C-FFT. Very similar trends are observed on the other ID datasets, with C-ADAPT, and with other SEs and LMs.



(a) MPNET (C-FFT) (b) MLM12 (C-ADAPT)

Figure 6: Impact of (random) sampling of positive examples (*high-data* scenarios) for contrastive SE specialization on the final mID performance (F_1 scores shown). **(a)** MPNET is the underlying SE, C-FFT; **(b)** MLM12 is the underlying encoder, C-ADAPT. x-axis is in the log-scale for clarity; straight lines of the same color and style refer to respective model configurations without any contrastive specialization.

Impact of Label Smoothing. Figure 4 suggests the importance of applying label smoothing, especially in *low-data* scenarios (e.g., drops in F_1 scores even up to 4-5 points) and with contrastively tuned encoders, where there is a higher chance of overfitting that leads to classification overconfidence. Switching off label smoothing (i.e., effectively setting $ls = 1.0$) is less severe in *high-data* setups, but our results render it almost universally useful for different MULTI-CONVFIT model variants.

Training Duration. Figure 5 indicates that the highest gains in mID performance are achieved in the first few epochs of contrastive fine-tuning. There is a large leap already after a single epoch of C-FFT or C-ADAPT, with more gains achieved in subsequent epochs before the procedure starts converging. Figure 5 also illustrates similar learning patterns both for SEs and LMs: the starting gap between SEs and their corresponding LMs does not get mitigated through contrastive specialization, again suggesting the importance of using SEs instead of LMs as the underlying text encoders.

Subsampling Positive Examples. The complexity of contrastive fine-tuning scales quadratically with the number of task-annotated examples, i.e., its complexity is $O(|\mathcal{S}|^2)$. Therefore, despite observed gains in *high-data* setups, the procedure might become prohibitively expensive if the datasets are too large. We investigate if a model variant where (i) we randomly subsample a smaller number of examples from the set $\mathcal{S}$ for the creation of sets *PosP* and *NegP*, while (ii) keeping the full set for the much less expensive part of the model, MLP training, still maintains the benefits stemming from contrastive specialization.

The impact of such random sampling of positive examples is illustrated in Figures 6a (C-FFT) and 6b (C-ADAPT). The plots suggest several findings. **1)** As expected, relying on more positive examples yields higher absolute scores, but the large increase in training time does come with diminishing returns in terms of performance (i.e., F_1 scores start saturating already with 500-1000 examples. **2)** Even a small number of examples (100-200) already yields large benefits over the model variant that does not apply any contrastive specialization, suggesting that it is possible to trade off a fraction of performance for large efficiency benefits. Finally, similar patterns are again observed for C-FFT and for C-ADAPT.

5 Conclusion and Future Work

We proposed MULTI-CONVFIT, a contrastive fine-tuning framework for multi-label intent detection (mID) that transforms general-purpose language models and sentence encoders (SEs) into task-specialized SEs. Such specialized SEs facilitate efficient learning of mID classifiers stacked on top of the fixed sentence encodings (Casanueva et al., 2020). Moreover, we demonstrate how to combine SEs with lightweight adapter modules, resulting in a modular multi-tenant design of the MULTI-CONVFIT framework. We demonstrate effectiveness and robustness of contrastive mID task specialization across a representative set of mID datasets, different input encoders, with large improvements especially in the most demanding low-resouce scenarios. We hope that MULTI-CONVFIT will inspire more work on sample-efficient, modular, and highly adaptable multi-label intent detectors.

There are multiple avenues for future research that can further improve various aspects of the proposed MULTI-CONVFIT framework. For instance, in this work, for simplicity and clarity, we rely on globally set fixed threshold values θ, while such thresholds can also be adaptable, with differently calibrated values for different (sets of) intents (Hou et al., 2021). Further, this work relied on a particular class of parameter-efficient fine-tuning methods, bottleneck adapters, as one of the most established methods available. However, as mentioned in §4, future work should also explore other parameter-efficient methods (Pfeiffer et al., 2020a; Ding et al., 2022), aiming to achieve an even better trade-off between performance and parameter-efficiency. In particular, driven by the efficiency requirements, we will investigate parameter-efficient methods that do not increase the model size at all, and thus maintain the same time efficiency at inference, such as methods based on low-rank adaptation (Hu et al., 2022) or sparse fine-tuning (Sung et al., 2021; Ansell et al., 2022).

Acknowledgements

We are grateful to our colleagues at PolyAI for many fruitful discussions and their encouragement to pursue this project. We also thank the anonymous reviewers for their helpful suggestions.

Limitations

We believe there is room for enhancing the underlying contrastive fine-tuning technique. In this work we evaluated only a single contrastive loss, OCL, following the suggestions and empirical analyses from prior work (see §2.1) as well as our preliminary experiments, demonstrating its strong performance. However, other contrastive losses can also be applied within the MULTI-CONVFIT framework. Further, we relied on random sampling of negative examples, as well as random sampling of positive examples in §4.1: we believe that additional performance gains might be achieved through more

sophisticated and semantically guided sampling strategies (Kalantidis et al., 2020; Robinson et al., 2021, *inter alia*).

This work focused only on multi-label intent detection as a well-defined downstream application, and the methodology was inspired by the desiderata of (efficient) mID. While the proposed methodology is not tied to the mID task and should be equally applicable to other multi-label sentence classification tasks, we did not evaluate the capacity and usefulness of the proposed methods in other tasks as part of this paper.

Finally, we point to the limitations of the current mID datasets and their design which cannot be mitigated solely through improving mID models: current mID datasets provide user utterances without any previous dialog context, and they still fail to distinguish between very subtle meaning differences in more difficult examples (e.g., using a toy example of sentences *No, I want the booking.* and *No, I don't want the booking.*, both sentences will be annotated with labels *deny* and *booking*).

References

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of ACL 2022*, pages 1778–1796.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of EMNLP 2021*, pages 4762–4781.

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. 2021. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. In *Proceedings of ICML 2021*, pages 566–576.

Wei Bi and James Tin-Yau Kwok. 2013. Efficient multi-label classification with many labels. In *Proceedings of ICML 2013*, pages 405–413.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP 2018*, pages 5016–5026.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Iñigo Casanueva, Ivan Vulić, Georgios P. Spithourakis, and Paweł Budzianowski. 2022. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of NAACL-HLT 2022*, pages 1998–2013.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of EMNLP 2018*, pages 169–174.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019a. Deep short text classification with knowledge powered attention. In *Proceedings of AAAI 2019*, pages 6252–6259.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019b. BERT for joint intent classification and slot filling. *CoRR*, abs/1902.10909.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. SNIPS Voice Platform: An embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190:12–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *CoRR*, abs/2203.06904.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of SIGDIAL 2017*, pages 207–219.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of ACL 2022*, pages 878–891.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of EMNLP 2021*, pages 7468–7475.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *Proceedings of ICLR 2021*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of CVPR 2006*, pages 1735–1742.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of EMNLP 2017*, pages 1923–1933.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *Proceedings of ICLR 2022*.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS Spoken Language Systems Pilot Corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, pages 96–101.

Matthew Henderson, Pawel Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the 1st Workshop on Natural Language Processing for Conversational AI*, pages 1–10.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of EMNLP 2020*, pages 2161–2174.

Jinhui Hou, Huanqiang Zeng, Lei Cai, Jianqing Zhu, Jing Chen, and Kai-Kuang Ma. 2019. Multi-label learning with multi-label smoothing regularization for vehicle re-identification. *Neurocomputing*, 345:15–22.

Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. 2021. Few-shot learning for multi-label intent detection. In *Proceedings of AAAI 2021*, pages 13036–13044.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of ICML 2019*, pages 2790–2799.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR 2022*.

Yannis Kalantidis, Mert Bülent Sariyildiz, Noé Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. In *Proceedings of NeurIPS 2020*, pages 21798–21809.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Stefan Larson and Kevin Leach. 2022. A survey of intent classification and slot-filling datasets for task-oriented dialog. *CoRR*, abs/2207.13211.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1311–1316.

Esther Levin and Roberto Pieraccini. 1995. Chronus, the next generation. In *Proceedings of the ARPA Workshop on Spoken Language Technology*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL-IJCNLP 2021*, pages 4582–4597.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of EMNLP 2021*, pages 1442–1459.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of IWSDS 2019*, pages 165–183.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *Proceedings of ICLR 2018*.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Findings of EMNLP 2020*, pages 2422–2433.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. In *Proceedings of NeurIPS 2021*, pages 1022–1035.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tür. 2020. DialoGLUE: A natural language understanding benchmark for task-oriented dialogue. *CoRR*, abs/2009.13570.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tür. 2021. Example-driven intent prediction with observers. In *Proceedings of NAACL-HLT 2021*, pages 2979–2992.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of ACL 2019*, pages 3836–3845.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:314–325.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Proceedings of NeurIPS 2019*, pages 4696–4705.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of ICML 2010*, pages 807–814.

Mahdi Namazifar, Alexandros Papangelis, Gokhan Tur, and Dilek Hakkani-Tür. 2021. Language model is all you need: Natural language understanding as question answering. In *Proceedings of ICASSP 2021*, pages 7803–7807.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *Proceedings of ICLR 2017: Workshop Track*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of EACL 2021*, pages 487–503.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 46–54.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of EMNLP 2020*, pages 7654–7673.

Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanxiang Che, and Ting Liu. 2021. GL-GIN: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling. In *Proceedings of ACL-IJCNLP 2021*, pages 178–188.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of EMNLP 2020*, pages 1807–1816.

Antoine Raux, Brian Langner, Alan W. Black, and Maxine Eskénazi. 2003. LET's GO: Improving spoken dialog systems for the elderly and non-natives. In *Proceedings of EUROSPEECH 2003*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP 2019*, pages 3982–3992.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *Proceedings of ICLR 2021*.

Sebastian Ruder. 2021. Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Proceedings of NeurIPS 2020*, pages 16857–16867.

Yi-Lin Sung, Varun Nair, and Colin Raffel. 2021. Training neural networks with fixed sparse masks. In *Proceedings of NeurIPS 2021*, pages 24193–24205.

Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2010. What is left to be understood in ATIS? In *Proceedings of SLU 2019*, pages 19–24.

Laurens van der Maaten and Geoffrey E. Hinton. 2012. Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of EMNLP 2021*, pages 1151–1168.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained Transformers. In *Proceedings of NeurIPS 2020*, pages 5776–5788.

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 EMNLP 2020: System Demonstrations*, pages 38–45.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of EMNLP 2020*, pages 917–929.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of CHI 2017*, pages 3506–3510.

Steve Young. 2010. Cognitive user interfaces. *IEEE Signal Processing Magazine*.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. 2021. Few-shot intent detection via contrastive pre-training and fine-tuning. In *Proceedings of EMNLP 2021*, pages 1906–1912.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of EMNLP 2020*, pages 5064–5082.

A Models and Data

URLs to the models used in this paper are provided in Table 3.

Our code is based on PyTorch, and relies on the following two widely used repositories:

- `sentence-transformers`: `www.sbert.net`
- `huggingface.co/transformers/`

Publicly available mID data can be accessed following these links:

- `github.com/PolyAI-LDN/task-specific-datasets/tree/master/nlupp` (NLU++: BANKING and HOTELS)
- `github.com/LooperXX/AGIF/tree/master/data` (MIXATIS)

(Due to concerns with privacy and security, the INSURANCEFAQ dataset cannot be publicly released in full.)

B Examples of Multi-Label Sentences

Some examples from the four multi-label ID datasets in our evaluation (see §3 and Table 1 in the main paper) are provided in Table 4.

C Additional Results

Additional empirical evidence and analyses, which further support the claims in the main paper, have been relegated to the appendix for clarity and compactness of the presentation in the main paper. These results to a large extent follow the trends observed in the results which are presented in the main paper, or offer additional supporting evidence for the main claims. In summary, we provide the following additional results:

Figure 7 is similar to Figure 2 in the main paper; it shows t-SNE plots that illustrate the effects of contrastive task specialization relying on three related encoders: MPNET-LM, MPNET, and MPNET +C-FFT, and on a subset of intent classes from the mID dataset HOTELS. Unlike this figure, Figure 2 in the main paper focuses on another set of encoders, and relies on +C-ADAPT contrastive fine-tuning, and plots examples from the BANKING dataset. Both of them demonstrate the desirable effect on the semantic space achieved by contrastive task specialization.

Table 5 shows the exact mID scores (F_1 and Acc) for several C-ADAPT variants on all four mID datasets in both data setups; see also related Figure 3 in the main paper.

Table 6 demonstrates the impact of duration of contrastive task specialization on mID Accuracy scores; Figure 5 in the main paper demonstrates the impact on mID F_1 scores.

Figure 8 demonstrates the impact of disabling label smoothing on final mID Accuracy scores (while Figure 4 in the main paper demonstrates the impact on mID F_1 scores).

Intents per Example. Finally, Figure 9 shows F_1 scores over examples with a different number of intents, while a similar plot with Acc scores is provided in Figure 10. Contrastive task specialization leads to pronounced improvements over all groups of examples, especially in *low-data* setups. Interestingly, while Acc scores are naturally higher for the groups with a fewer number of intents (see Figure 10), the 1-label group displays lower F_1 scores than 2-label or 3-label groups on BANKING and INSURANCEFAQ. We attribute it to the modular ontology design (Casanueva et al., 2022): as a consequence, 1-label examples in those mID datasets are typically very short sentences (e.g., 1-3 word tokens), which are known to pose a challenge for sentence encoders (Chen et al., 2019a).

Name	Abbreviation	URL
	Language Models	
MiniLM-L12-H384-uncased	MLM12-LM	huggingface.co/microsoft/MiniLM-L12-H384-uncased
mpnet-base	MPNET-LM	huggingface.co/microsoft/mpnet-base
distilroberta-base	DROB-LM	huggingface.co/distilroberta-base
	Sentence Encoders	
all-MiniLM-L12-v2	MLM12	huggingface.co/sentence-transformers/all-MiniLM-L12-v2
all-mpnet-base-v2	MPNET	huggingface.co/sentence-transformers/all-mpnet-base-v2
all-distilroberta-v1	DROB	huggingface.co/sentence-transformers/all-distilroberta-v1

Table 3: URLs, abbreviations of the language models and sentence encoders used in this work.

Dataset	Sentence	Labels
BANKING	I want to apply for a loan, what should I do?	*loan, make, request_info*
BANKING	The pin for my card is not the same as the one for my account, right?	*pin, account, card, request_info*
HOTELS	Cancel the restaurant reservation for 18:45 under Jane Doe.	*cancel_close, restaurant, booking*
HOTELS	I have a reservation and I need to change the number of adults	*change, existing, booking*
INSURANCEFAQ	I'm stuck at the tax identification number.	*tax_id, not_working*
INSURANCEFAQ	How do I reset my security questions?	*how, change, security_question*
MIXATIS	what is the distance between Pittsburgh airport and downtown Pittsburgh and what are my meal options from Boston to Denver?	*atis_distance, atis_meal*
MIXATIS	what is the code for business class and show me the cities served by nationair	*atis_abbreviation, atis_city*

Table 4: Examples from multi-label ID datasets in our evaluation (see Table 1 in §3 of the main paper).

SE↓ / ID Dataset→	BANKING		HOTELS		INSURANCEFAQ		MIXATIS	
	low-data	*high-data*	*low-data*	*high-data*	*low-data*	*high-data*	*low-data*	*high-data*
MLM12 +C-ADAPT	80.3 / 46.7	93.5 / 78.7	67.9 / 47.4	91.4 / 80.4	75.8 / 47.6	89.5 / 73.9	77.3 / 40.6	90.4 / 78.3
DROB +C-ADAPT	80.7 / 47.9	93.7 / 77.2	67.3 / 47.6	92.8 / 84.9	75.8 / 47.2	89.2 / 73.9	73.7 / 44.6	90.8 / 78.3

Table 5: Results in the multi-label ID task for several C-ADAPT variants. F_1 / Acc scores.

Encoder↓ / Epoch →	0	1	2	3	4	5	6	7	8	9	10
MPNET-LM	23.9	33.3	36.1	35.8	38.0	40.0	41.1	39.9	40.3	40.5	41.3
MPNET	30.0	42.7	46.5	47.0	47.4	47.9	48.0	48.3	48.6	49.1	49.1
DROB-LM	27.2	34.1	36.7	37.9	39.1	39.5	40.1	39.4	39.7	39.2	39.9
DROB	31.0	39.2	44.5	45.8	45.9	46.7	47.0	47.5	47.1	47.9	47.6

Table 6: Impact of contrastive specialization duration (i.e., the number of epochs) on the final mID performance in *low-data* scenarios on BANKING. Acc scores; C-FFT. Very similar trends are observed on the other ID datasets, with C-ADAPT, and with other SEs and LMs.

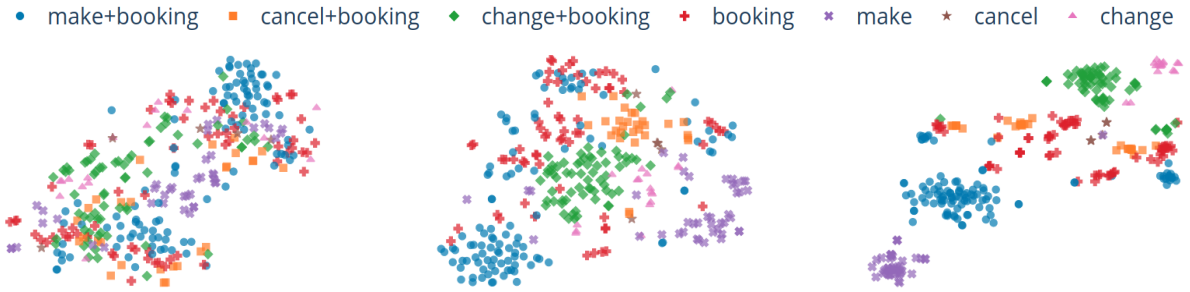


Figure 7: t-SNE plots (van der Maaten and Hinton, 2012) of encoded utterances from the mID dataset HOTELS (see §3) associated with a subset of intent classes, demonstrating the effects of contrastive task specialization of the input encoder with mID data. **Left**: sentence encodings with the original MPNet LM (Song et al., 2020); **Middle**: encodings with MPNet transformed into a universal SE (Reimers and Gurevych, 2019); **Right**: encodings with a task-specialized SE obtained after contrastively fine-tuning the universal MPNet-based SE.

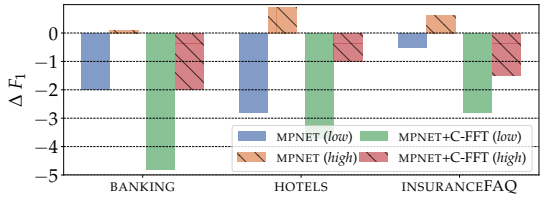
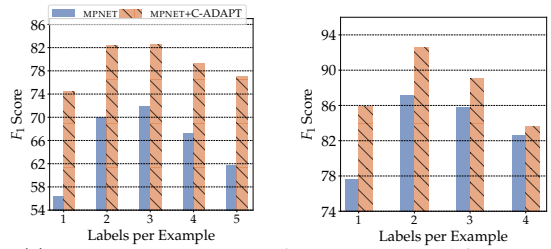
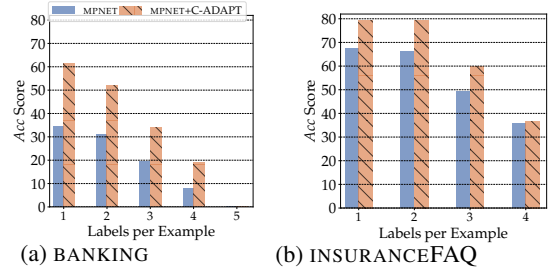


Figure 8: Change in Acc performance when no label smoothing is used, with all other parts kept equal. Similar trends are observed with other input models and with C-ADAPT.



(a) BANKING (b) INSURANCEFAQ

Figure 9: F_1 scores over examples with a particular number of intents; **(a)** *low-data*, **(b)** *high-data*.



(a) BANKING (b) INSURANCEFAQ

Figure 10: *Acc* scores over examples with a particular number of intents; **(a)** *low-data*, **(b)** *high-data*.