CONVFIT: Conversational Fine-Tuning of Pretrained Language Models

Ivan Vulić, Pei-Hao Su, Sam Coope, Daniela Gerz,
Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and **Tsung-Hsien Wen**

PolyAI Limited
London, United Kingdom
www.polyai.com

Abstract

Transformer-based language models (LMs) pretrained on large text collections are proven to store a wealth of semantic knowledge. However, **1)** they are not effective as sentence encoders when used off-the-shelf, and **2)** thus typically lag behind conversationally pretrained (e.g., via response selection) encoders on conversational tasks such as intent detection (ID). In this work, we propose CONVFIT, a simple and efficient two-stage procedure which turns any pretrained LM into a universal conversational encoder (after Stage 1 CONVFIT-ing) and task-specialised sentence encoder (after Stage 2). We demonstrate that **1)** full-blown conversational pretraining is not required, and that LMs can be quickly transformed into effective conversational encoders with much smaller amounts of unannotated data; **2)** pretrained LMs can be fine-tuned into task-specialised sentence encoders, optimised for the fine-grained semantics of a particular task. Consequently, such specialised sentence encoders allow for treating ID as a simple semantic similarity task based on interpretable nearest neighbours retrieval. We validate the robustness and versatility of the CONVFIT framework with such similarity-based inference on the standard ID evaluation sets: CONVFIT-ed LMs achieve state-of-the-art ID performance across the board, with particular gains in the most challenging, few-shot setups.

1 Introduction and Motivation

Pretrained Transformer-based (masked) language models (LMs) such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019b), coupled with task-specific fine-tuning, offer unmatched state-of-the-art performance in a wide array of standard language understanding and conversational tasks (Wang et al., 2019a; Mehri et al., 2020). However, pretrained LMs do not produce coherent and effective sentence encodings off-the-shelf; their further adaptation is required, akin to standard task fine-tuning. For instance, Reimers and Gurevych (2019) transform monolingual English BERT with supervised natural language inference and paraphrasing data (Williams et al., 2018; Wieting and Gimpel, 2018) into a sentence encoder which excels at sentence similarity and retrieval tasks (Marelli et al., 2014; Cer et al., 2017). This transformation process supports the creation of other similar universal sentence encoders in monolingual and multilingual settings (Chidambaram et al., 2019; Wieting et al., 2020; Feng et al., 2020), and is typically based on dual-encoder architectures.

Another parallel research thread aims at learning *conversational* encoders: it validates the benefits of masked language modeling (MLM) pretraining on naturally conversational data (Wu et al., 2020; Mehri et al., 2021), as well as the benefits of transfer learning for conversational tasks which goes beyond MLM as the pretraining objective (Mehri et al., 2019; Coope et al., 2020; Henderson and Vulić, 2021, *inter alia*). In particular, response selection as a suitable pretraining task (Al-Rfou et al., 2016; Yang et al., 2018; Henderson et al., 2019b; Humeau et al., 2020) learns representations that organically capture conversational cues from conversational text data such as Reddit (Henderson et al., 2019a), again via dual-encoder architectures.

Inspired by these two research threads, we pose the following two crucial questions:

(Q1) Is it necessary to conduct full-scale expensive conversational pretraining? In other words, is it possible to simply and quickly 'rewire' existing MLM-pretrained encoders as conversational encoders via, e.g., response ranking fine-tuning on (much) smaller-scale datasets?

(Q2) If we frame conversational tasks such as intent detection as semantic similarity tasks instead of their standard classification-based formulation, is it also possible to frame supervised task-specific learning as fine-tuning of conversational sentence encoders? In other words, can we learn

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1151–1168
November 7–11, 2021.

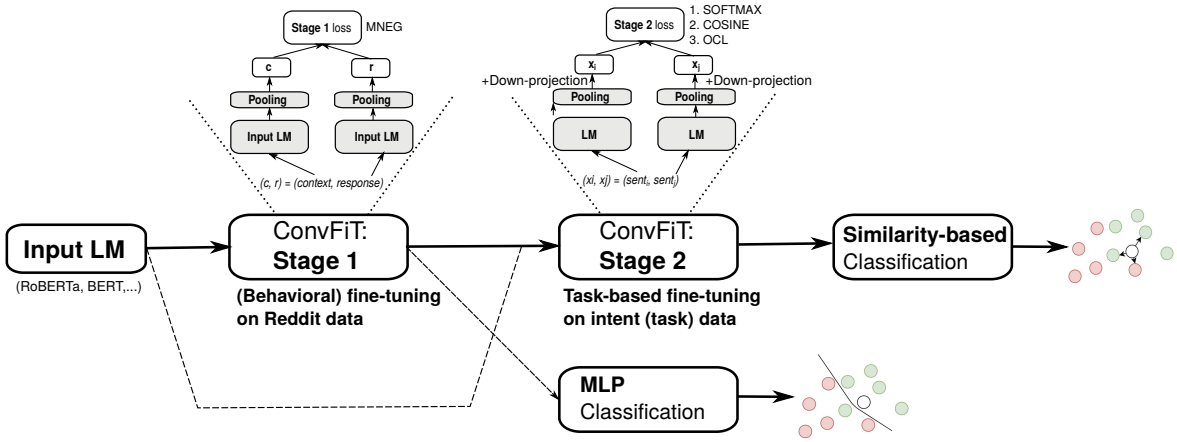
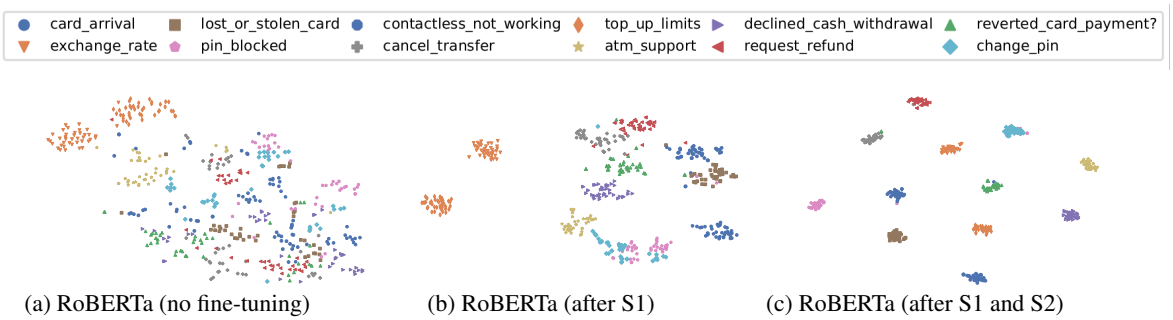


Figure 1: Illustration of the full CONVFIT framework which fine-tunes pretrained LMs such as BERT or RoBERTa in two separate stages via dual-encoder networks ("zoomed-in" parts; grey blocks denote tunable parameters), and performs intent detection with the CONVFIT-ed models via similarity-based inference. **Stage 1 (S1):** adaptive conversational fine-tuning, §2.1; **Stage 2 (S2):** task-tailored conversational fine-tuning (for intent detection), §2.2. Dashed lines denote baseline/ablation variants which skip one of the two stages: (i) we can directly task-tune the sentence encoder with the task data (Stage 2) without running Stage 1, or (ii) we can skip Stage 2, and similar to Casanueva et al. (2020), learn an MLP classifier on top of the conversational representations from Stage 1.



(a) RoBERTa (no fine-tuning) (b) RoBERTa (after S1) (c) RoBERTa (after S1 and S2)

Figure 2: t-SNE plots (van der Maaten and Hinton, 2012) of encoded utterances from the ID test set of BANKING77 (i.e., all examples are effectively unseen by the encoder models at training) associated with a selection of 12 intents, demonstrating the effects of gradual "representation specialisation funnel". The encoded utterances are created via mean-pooling based on **(a)** the original RoBERTa LM; **(b)** RoBERTa after Stage 1 (i.e., fine-tuned on 1% of the full Reddit corpus, see Figure 1); **(c)** RoBERTa after Stage 1 and Stage 2, fine-tuned with the OCL objective ($n = 3$ negatives) using the entire BANKING77 training set (see Figure 1). Additional t-SNE plots are in the Appendix.

task-specialised sentence encoders that enable *sentence similarity-based* interpretable classification?

In order to address these two questions, we propose **CONVFIT**, a two-stage **CONV**ersational **FI**ne-**T**uning procedure that turns general-purpose MLM-pretrained encoders into sentence encoders specialised for a particular conversational domain and task. Casting the end-task (e.g., intent detection) as a pure sentence similarity problem then allows us to recast task-tailored fine-tuning of a pretrained LM as gradual *sentence encoder specialisation*, as illustrated in Figures 1 and 2.

Our hypothesis is that the pretrained LMs, which already store a wealth of semantic knowledge, can be gradually turned into conversational task-adapted sentence encoders without expensive full pretraining. **(S1)** Stage 1 transforms pretrained LMs into universal conversational encoders via *adaptive fine-tuning* (Ruder, 2021) on (a fraction of) Reddit data (see Figure 2b), relying on a standard dual-encoder architecture with a conversational response ranking loss (Henderson et al., 2020); cf. Q1. **(S2)** Stage 2 further specializes the sentence encoder via *contrastive learning with in-task data*, that is, it learns meaningful task-related semantic clusters/subspaces. We then show that the S2 task-tailored specialisation effectively enables a simple and interpretable similarity-based classification based on nearest neighbours (NNs) in the specialised encoder space (see Q2 and Figure 2c).

The two-stage CONVFIT transformation offers new insights and contributions to representation

learning for conversational tasks. Unlike prior work which conducted large-scale conversational pretraining from scratch using large datasets, we demonstrate that full pretraining is not needed to obtain universal conversational encoders. By leveraging the general semantic knowledge already stored in pretrained LMs, we can expose (i.e., 'rewire') that knowledge (Vulić et al., 2021; Gao et al., 2021b; Liu et al., 2021b) via much cheaper and quicker adaptive fine-tuning on a tiny fraction of the full Reddit data (e.g., even using $< 0.01\%$ of the Reddit corpus). Further, the task-oriented S2 CONVFIT-ing transforms pretrained LMs into task-specialised sentence encoders. Our results with similarity-based classification, targeting the crucial conversational NLU task of intent detection (ID), reach state-of-the-art (SotA) across all standard ID datasets, with particular gains in the most challenging, few-shot setups. Importantly, we show that the gradual application of S1 and then S2 yields a synergistic effect, that is, it attains the highest ID results across the board.

Finally, CONVFIT is highly versatile: it can be used with a range of pretrained LMs and on a spectrum of text classification problems; it also allows for the simple usage of diverse fine-tuning objectives in both Stage 1 and Stage 2, beyond the ones proposed and evaluated in this work.

2 Methodology

Preliminaries. For any input text t, we obtain its encoding $\mathbf{t} = enc(t)$, where enc is a sentence encoder at any CONVFIT stage (i.e., before any fine-tuning, after S1, or after S2), or any other sentence encoder. The text t is tokenized into subwords (Schuster and Nakajima, 2012) relying on each encoder's dedicated tokeniser. The final encoding $\mathbf{t}$ is created via a *pooling* operation such as (a) using the [CLS] token, (b) or mean-pooling the output subword vectors. Following prior work (Reimers and Gurevych, 2019), we always use mean-pooling.

2.1 Stage 1: Adaptive Fine-Tuning

As in prior work on conversational pretraining (Henderson et al., 2019b, 2020; Humeau et al., 2020), Stage 1 relies on the response ranking task with Reddit data and dual-encoder architectures, which model the interaction between Reddit *(context, response)* (c, r) pairs.[1] However, unlike prior work, instead of pretraining from scratch we fine-tune an LM-pretrained encoder, which yields a much quicker conversational encoder specialisation, and does not require massive amounts of data.

Response ranking is formulated as the standard multiple negatives ranking loss (MNEG): for each positive (c_i, r_i) pair (i.e., the pair observed in the Reddit fine-tuning data), the aim is to rank the correct response r for the input c over a set of randomly sampled responses $r_j, j \neq i$ from other Reddit pairs. The similarity between c-s and r-s is quantified via the similarity function S operating on their encodings $S(\mathbf{c}, \mathbf{r})$. Following prior work, we use the scaled cosine similarity: $S(\mathbf{c}, \mathbf{r}) = D \cdot cos(\mathbf{c}, \mathbf{r})$, where D is the scaling constant. Stage 1 fine-tuning with MNEG then proceeds in batches of B positive Reddit pairs $(c_i, r_i), \ldots, (c_B, r_B)$; the MNEG loss for a single batch is computed as:

$$\mathcal{L} = -\sum_{i=1}^{B} S(\mathbf{c_i}, \mathbf{r_i}) + \sum_{i=1}^{B} \log \sum_{j=1, j\neq i}^{B} e^{S(\mathbf{c_i}, \mathbf{r_j})} \quad (1)$$

Effectively, for each batch Eq. (1) maximises the similarity score of positive context-response pairs (c_i, r_i), while it minimises the score of $B-1$ random pairs. The negative examples are all pairings of c_i with r_j-s in the current batch, where such (c_i, r_j) pairs do not occur in the Reddit data.[2]

The output of Stage 1 is the sentence encoder enc_{S1} which can be used 'as is' similarly to standard sentence encoders (Henderson et al., 2020; Casanueva et al., 2020; Feng et al., 2020): a standard ID approach stacks a Multi-Layer Perceptron (MLP) classifier on top of the fixed sentence vectors $\mathbf{t}$, and fine-tunes only the MLP parameters (Casanueva et al., 2020; Gerz et al., 2021). However, the output of S1 can also be further fed as the input encoding for CONVFIT's Stage 2 (Figure 1).

2.2 Stage 2: Task-Based Sentence Encoders

Stage 2 fine-tuning is inspired by metric-based meta-learning (Vinyals et al., 2016; Musgrave et al., 2020) and exemplar-based (also termed prototype-based) learning (Snell et al., 2017; Sung et al., 2018; Zhang et al., 2020), which is especially suited for few-shot scenarios. We assume the existence of N_a annotated in-task examples

[1] In each (c, r) pair, r is the *response* that immediately follows the preceding *context* sentence in a Reddit thread; see (Henderson et al., 2019a). The intuition is that sentences which elicit similar responses should obtain similar sentence encodings (Yang et al., 2018).

[2] We also experimented with another SotA loss function, the triplet-based multi-similarity loss (Wang et al., 2019b; Liu et al., 2021a), without any substantial performance differences.

$\{(x_1, y_1), \ldots, (x_{N_a}, y_{N_a})\}$: e.g., x-s are text sentences with y-s being their intent labels/classes; let us assume that there are N_c classes $\{C_1, \ldots, C_{N_c}\}$ in total. The aim is to fine-tune the input sentence encoder in such a way to encode all sentences associated with each particular class into coherent clusters, clearly separated from all other class-related (also coherent) clusters (see Figure 2c).[3]

Positive and Negative Pairs. We leverage the class labels only implicitly (see Figure 1), which allows us to *treat intent detection as a sentence similarity task*. CONVFIT S2 operates with two sets of pairs: **1)** PP is the set of positive pairs (x_i, x_j), where x_i and x_j are text instances associated with the same class C_i; **2)** NP contains negative pairs (x_i, x_j) where x_i and x_j are associated with two different classes C_i and C_j. We construct the set NP in a balanced way: for each positive pair $(x_i, x_j) \in PP$, we add $2 \times n$ negative pairs into NP, where n is a tunable hyper-parameter; n pairs $(x_i, x_{i,n'}), n' = 1, \ldots, n$, are constructed by randomly sampling utterances $x_{i,n'}$ which do not share the class with x_i, and we also sample n negatives $(x_{j,n'}, x_j)$ in a similar vein. We now present three different loss functions that fine-tune the input encoders towards task-specialised sentence similarity relying on the sets PP and NP. For all three S2 loss functions, we add a *down-projection* d_o-dim layer with non-linearity (*Tanh* used) after pooling, see Figure 1.[4]

SOFTMAX (SMAX) Loss. Following prior work (Reimers and Gurevych, 2019), for each input sentence pair (x_i, x_j), we concatenate their d_o-dimensional encodings $\mathbf{x_i}$ and $\mathbf{x_j}$ (obtained after passing them through the input encoder, pooling, and down-projection) with their element-wise difference $|\mathbf{x_i} - \mathbf{x_j}|$. The objective is as follows: $\mathcal{L}_{\text{SMAX}} = \text{softmax}\big(W(\mathbf{x_i} \oplus \mathbf{x_j} \oplus |\mathbf{x_i} - \mathbf{x_j}|)\big)$, where $\oplus$ denotes concatenation, and $W \in \mathbb{R}^{3d_o \times 2}$ is a trainable weight matrix of the softmax classifier, where 2 is the number of classification classes: the model must simply discern between positive pairs (from PP) and negative pairs from NP. The classifiers are optimised via standard cross-entropy.

Cosine (COS) Loss. The idea is to minimise the following distance, formulated as standard mean-squared error: $||\delta_l - cos(\mathbf{x_i}, \mathbf{x_j})||_2$, where cos denotes cosine similarity, and δ_l is a hyper-parameter which specifies the 'ideal' (dis)similarity margin in the specialised encoder space. Here, we rely on the default parameters from Reimers and Gurevych (2019) without any tuning: $\delta_l = 0.8$ iff $(x_i, x_j) \in PP$, and $\delta_l = 0.3$ iff $(x_i, x_j) \in NP$.

Online Contrastive Learning (OCL) Loss follows the formulation from Hadsell et al. (2006):

$$\mathcal{L}_{\text{OCL}} = \mathbb{1} \cdot (dcos(\mathbf{x_i}, \mathbf{x_j}))^2 + (1 - \mathbb{1}) \cdot \Big(ReLU(\delta_m - dcos(\mathbf{x_i}, \mathbf{x_j}))\Big)^2 \quad (2)$$

where $\mathbb{1}$ is the indicator function which returns 1 iff $(\mathbf{x_i}, \mathbf{x_j}) \in PP$, and 0 iff $(\mathbf{x_i}, \mathbf{x_j}) \in NP$; $dcos = 1 - cos$ 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. The loss 'attracts' similar items closer together in the specialised space, while 'repelling' dissimilar items (Mrkšić et al., 2017).[5]

Similarity-Based Inference. Intent detection in the specialised encoder space enc_{S2} is then performed via similarity-based classification (Zhang et al., 2020) after Stage 2.[6] Assuming the simplest case of $k = 1$ nearest neighbours (NN) classification, we select the intent class for an unseen example u as: $I_c\big(\arg\max_{\mathbf{t} \in Pool} cos(\mathbf{t}, \mathbf{u})\big)$. Here, $\mathbf{t} = enc_{S2}(t)$ refers to the sentence encoding of each example $t \in Pool$ (which is typically the pool of examples from the ID training set), and the I_c function returns the intent class of any $t \in Pool$.

Why Intent Detection as a Sentence Similarity Task? We can take the analogy of 'intent' being a latent semantic class where sentences associated with the intent are diverse surface instances of the class (i.e., language realisations of the underlying concept/intent). This means that finding the most similar labelled instances for the given unlabelled input instance/sentence can directly inform us about the underlying semantic class/intent.

[3] In other words, the encoder should learn to encode each utterance into one of such semantically well-defined clusters.

[4] A variant with down-projection yielded slightly higher scores than the one without it in our preliminary experiments.

[5] We use the *online* version of the loss that updates the loss focusing on hard negative pairs (i.e., negatives that are close by cosine in the current semantic space) and hard positives which are far apart in the current space. This typically results in quicker convergence and slightly better performance.

[6] The benefits of similarity-based classification were recently validated also in other NLP tasks such as cross-lingual abusive content detection (Sarwar et al., 2021), language modeling (Khandelwal et al., 2020; Guu et al., 2020), and question answering (Kassner and Schütze, 2020), among others.

Dataset	Intents	Examples	Domains
BANKING77	77	13,083	1 (banking)
CLINC150	150	23,700	10
HWU64	64	25,716	21

Table 1: Intent detection datasets: key statistics.

3 Experimental Setup

Input LMs. We experiment with several popular Transformer-based (Vaswani et al., 2017) LMs as input (see Figure 1), aiming to validate the robustness of CONVFIT, as well as to analyse the impact of LM pretraining on the final task performance: (i) BERT (Devlin et al., 2019) (labeled BERT henceforth); (ii) RoBERTa (ROB), as an improved variant of BERT, LM-pretrained with more data (Liu et al., 2019b); (iii) DistilRoBERTa (DROB), a distilled more compact version of RoBERTa, LM-pretrained with around 4 times fewer data than the teacher RoBERTa model (Sanh et al., 2019). The cased BASE variants are used for all input LMs: 768-dimensional Transformer layers with 12 (BERT, ROB) or 6 (DROB) attention layers. In addition, to isolate the effects of LM-pretraining and CONVFIT-ing from the mere "parameter capacity", we also experiment with a BERT/ROB architecture with RANDomly initialised parameters using the Xavier initialisation (Glorot and Bengio, 2010).

Unless noted otherwise, CONVFIT Stage 1 always proceeds with a sample comprising 2% of the full Reddit corpus from Henderson et al. (2019a).[7]

Intent Detection Datasets. As discussed in §2, the main evaluation task is intent detection (ID), with a particular focus on low-data (i.e., few-shot) scenarios. Our Stage 2 fine-tuning and the final task evaluation are based on three standard ID datasets in English, also available as part of the recently published DialoGLUE benchmark (Mehri et al., 2020): BANKING77 (Casanueva et al., 2020), HWU64 (Liu et al., 2019a), and CLINC150 (Larson et al., 2019).[8] The key statistics of all three datasets are provided in Table 1; for further details, we refer the reader to the original work and also to (Mehri et al., 2020).

Few-Shot and Full Data Setups. Prior work has recognised the importance of building intent detectors in low-data regimes (Casanueva et al., 2020; Mehri et al., 2021). Therefore, following this initiative, we evaluate the models in two **N-shot** scenarios, where we assume that only $N = 10$ or $N = 30$ annotated examples per intent are available for training the MLP classifier or for S2 fine-tuning; Figure 1.[9] The models are also evaluated in the **Full** setup, where all annotated training examples per intent are used. Note that we always report the scores on the same test set for each setup. For the few-shot scenarios, we report the scores as averages over 3 independent experimental runs.

Hyperparameters and Optimisation. CONVFIT is implemented via the *sentence-transformers (sbert)* repository (Reimers and Gurevych, 2019), which is in turn built on top of the HuggingFace repository (Wolf et al., 2020). Similar to Casanueva et al. (2020), we do not rely on any development data, and follow the general suggestions from prior work (Reimers and Gurevych, 2019; Casanueva et al., 2020) for the hyperparameter setup, which is adopted across all intent ID datasets.[10] For S1 with MNEG, we always train for 2 epochs in batches of 256 with default hparams from *sbert*.[11]

In Stage 2, with all three evaluated objective functions the batch size is 32, the maximum sequence length is 48, the output layer's dimensionality is set to $d_o = 512$. Unless stated otherwise, we always fine-tune for 10, 5, and 2 epochs for the 10-shot, 30-shot, and Full setups, respectively. For the COS and OCL variants, unless noted otherwise, we report the results with $n = 3$ negative examples per each positive in 10-shot and 30-shot setups, and with $n = 1$ (for computational tractability) in the Full setup. An analysis of the impact of n on the final ID performance is presented later in §4.

Following the suggested settings of Reimers and Gurevych (2019); Vulić et al. (2020), in both CONVFIT stages we use the AdamW optimiser (Loshchilov and Hutter, 2018); the learning rate is $2e-5$ with the warmup rate of 0.1 and linear decay

[7] The full corpus contains 700M+ *(context, response)* pairs.

[8] The datasets provide a range of diverse ID setups, covering fine-grained ID within a single domain (e.g., BANKING77), as well as coarser-grained ID spanning several well-defined domains (e.g., news, calendar, alarm, restaurant booking in HWU64 or in CLINC150). They provide a more challenging setup (and are also better aligned with the actual ID setups typically met in production) than some other well-known ID datasets such as SNIPS (Coucke et al., 2018).

[9] We use the same fixed few-shot and test sets for each intent detection dataset as released by Mehri et al. (2020).

[10] For all MLP intent classifiers, this implies relying on the empirically validated and stable setup from prior work (Casanueva et al., 2020): the best results are achieved with a 2-layer fully-connected MLP (768-dim hidden layers), trained via SGD with the high learning rate (0.5) and linear decay, and very aggressive dropout rates (0.75); training lasts for 500 epochs; batch size is 32. This setup achieved strong results in our preliminary experiments as well, and is thus adopted here.

[11] 256 is the maximum batch size with BASE BERT and RoBERTa which allows us to run Stage 1 fine-tuning on a single 12GiB GTX GPU.

afterwards, and the weight decay rate is set to 0.01.

Similarity-Based Classification. The intent class is chosen according to the $k = 1$ NNs, based on the cosine distance in the fine-tuned space.[12] Importantly, in few-shot setups we use *only* the few-shot data as the NN pool for classification.

3.1 Model Variants and Baselines

We experiment with a range of model variants enabled by the CONVFIT framework (see Figure 1), and compare their performance in the ID task against an array of cutting-edge universal and conversational sentence encoders. All the models in evaluation are summarised here for clarity.

LM+S1+S2-LOSS. Sentence encoders after running the full CONVFIT pipeline, where intent detection is based on similarity-based NN classification. LM in the label of this variant denotes the input LM, and LOSS is the loss function used in Stage 2 (i.e., SMAX, COS, or OCL).

LM+S2-LOSS. Sentence encoders optimised only via Stage 2 CONVFIT, skipping Stage 1 (see Figure 1); similarity-based intent detection.

LM+S1. The input LM is converted into a (general-purpose) conversational encoder via Stage 1 CONVFIT-ing; intent detection is performed via standard feature-based MLP classification on top of the sentence encodings as in prior work.

SotA Sentence Encoders. We evaluate three widely used state-of-the-art sentence encoders in the standard feature-based MLP classification approach to intent detection:[13] (i) *ConveRT* (Henderson et al., 2020) is a dual sentence encoder pretrained with the conversational response selection task (Henderson et al., 2019b) on the full Reddit data (Al-Rfou et al., 2016; Henderson et al., 2019a); (ii) multilingual Universal Sentence Encoder (*mUSE*) (Yang et al., 2020) is a multilingual and better-performing version of the USE model for English (Cer et al., 2018), which again relies on a standard dual-encoder framework (Henderson et al., 2019b; Humeau et al., 2020) and is pretrained on massive amounts of data; (iii) Language-agnostic BERT Sentence Embedding (*LaBSE*) (Feng et al., 2020) adapts pretrained multilingual BERT (mBERT) (Devlin et al., 2019) into a sentence encoder using a dual-encoder framework (Yang et al., 2019) with larger embedding capacity (i.e., it provides a shared multilingual vocabulary spanning 500k subwords).[14]

4 Results and Discussion

The main results are summarised in Table 2, and further results and analyses are available in §4.1, with additional results in the Appendix.[15] These results offer multiple axes of comparison, succinctly discussed in what follows.

MLP versus Similarity-Based ID. First, we note that CONVFIT-ed LMs achieve peak ID scores across all three ID datasets, and in all data setups, with ROB+S1+S2-OCL being the highest-performing model variant overall. Running Stage 1 does transform input LMs into effective (universal) conversational encoders already: for MLP-based ID, we observe competitive or even improved performance (cf., the results on BANKING77 and HWU64 as two more challenging evaluation sets) with the ROB+S1 and BERT+S1 variants against current state-of-the-art (conversational) sentence encoders such as ConveRT, USE, and LaBSE.

Importantly, the results after Stage 2 'unanimously' suggest the effectiveness of treating ID as a semantic similarity task, and additional task-specific specialisation of the sentence encoders with in-task data. Put simply, it seems more effective to use the in-task training data to 'task-specialise' the sentence encoder space than to learn a standard (MLP) classifier, which directly maps from the feature space to intent labels (Sarwar et al., 2021). The gains are especially pronounced in few-shot setups (e.g., see 10-shot BANKING77).

We speculate that dual-encoder contrastive learning surpasses MLP-based approaches especially in few-data scenarios because it learns from finer-grained and more abundant information in such low-data scenarios: i.e., we learn to contrast between pairs of instances rather than simply learning an MLP-based mapping from an instance to its underlying class intent/class. This formulation can also capture some subtle cross-instance (dis)similarities which cannot be captured by MLP.

[12] Very similar results are observed with $k = 3$ and $k = 5$.

[13] For more technical details regarding each sentence encoder, we refer the reader to the original work.

[14] LaBSE is the current SotA encoder across a wide array of languages (Feng et al., 2020; Litschko et al., 2021; Gerz et al., 2021). Besides dual-encoder training, LaBSE leverages standard self-supervised objectives used in pretraining of mBERT and XLM: masked and translation language modeling (Conneau and Lample, 2019); see the original work.

[15] For brevity, in the main paper we report the results with the two better-performing S2 losses: COS and OCL.

Model Variant	BANKING77			CLINC150			HWU64		
	10	30	Full	10	30	Full	10	30	Full
Similarity-Based Classification									
ROB+S1+S2-COS	86.48	91.33	**94.35**	92.87	95.91	97.20	85.06	**90.46**	**92.98**
BERT+S1+S2-COS	84.32	90.91	93.91	91.80	95.58	96.56	85.13	89.41	91.93
DROB+S1+S2-COS	85.13	90.75	94.06	91.64	95.48	97.00	83.64	89.68	92.94
ROB+S1+S2-OCL	**87.38**	**91.36**	94.16	**92.89**	**96.42**	**97.34**	**85.32**	90.06	92.42
BERT+S1+S2-OCL	85.97	90.65	93.77	91.53	95.53	96.82	85.04	89.41	92.21
DROB+S1+S2-OCL	86.04	90.78	93.89	91.98	95.60	97.04	83.64	89.50	92.84
ROB+S2-COS	84.96	90.81	94.19	91.56	95.64	96.78	84.52	89.87	92.19
BERT+S2-COS	81.27	90.32	93.73	89.58	95.08	96.54	82.90	89.12	91.78
DROB+S2-COS	83.28	90.58	93.91	89.47	95.32	86.78	82.43	89.41	92.10
ROB+S2-OCL	85.78	90.98	93.77	92.64	95.40	96.87	84.76	89.31	92.01
BERT+S2-OCL	82.28	89.77	93.54	90.71	95.07	96.62	83.09	88.94	92.57
DROB+S2-OCL	82.60	90.65	93.38	90.78	95.02	96.69	81.69	88.75	92.38
Baselines: MLP Classification									
ROB+S1	83.08	90.16	93.38	90.98	94.12	96.42	81.13	87.73	91.44
BERT+S1	82.69	89.82	93.67	89.88	94.07	96.33	82.25	88.01	91.12
CONVERT*	83.32	89.37	93.01	92.62	95.78	97.16	82.65	87.88	91.24
USE*	84.23	89.74	92.81	90.85	93.98	95.06	83.75	89.03	91.25
USE (ours)	82.95	89.09	92.81	90.27	93.54	94.91	82.71	88.20	91.64
LABSE	81.69	88.96	92.60	90.89	93.41	95.12	81.60	86.15	90.99

Table 2: Accuracy scores ($\times 100\%$) on the three ID data sets with varying number of training examples (**10** examples per intent; **30** examples per intent; **Full** training data). $n = 3$ negatives are used in Stage 2 for 10-shot and 30-shot setups, $n = 1$ for the Full setup (see §3). The peak scores per column are in bold, the second best is underlined. *The scores were taken directly from prior work, and computed on different 10/30-shot samples (and are thus not directly comparable, Zhao et al. 2021). For clarity, we show only a subset of (arguably most informative) model variants; the complete table with additional evaluated variants is available in the Appendix.

Extending beyond pure absolute performance, decisions based on k-NN similarity-based ID in the specialised space are also easy to interpret (Simard et al., 1992; Wallace et al., 2018).

Stage 1 + Stage 2? The scores in Table 2 indicate that Stage 2 alone already transforms pretrained LMs into very strong task-specialised sentence encoders. However, a more careful comparison of LM+S1+S2-LOSS versus LM+S2-LOSS variants reveals that Stage 1 fine-tuning is universally useful (regardless of the chosen loss function in S2), and yields ID performance gains. In other words, the coarser-grained adaptive fine-tuning already exposes some conversational knowledge from the pretrained LMs, and such knowledge does have substantial impact on task-specialised S2 tuning. In sum, this finding is line with prior work in other domains and NLP tasks (Gururangan et al., 2020; Glavaš et al., 2020; Ruder, 2021): both domain-adaptive (our S1) and task-adaptive additional tuning (our S2) of general-purpose LMs have a synergistic positive impact on the final task performance.

The impact of the gradual two-stage sentence encoder transformation is also clearly visible from the t-SNE visualisation in Figure 2. Besides this, a standard quantitative measure of cluster coherence, the Silhouette coefficient σ (Rousseeuw, 1987) also points in the same direction: $\sigma = 0.067$ for the test examples and model variant from Figure 2a, $\sigma = 0.188$ (Figure 2b), and $\sigma = 0.698$ (Figure 2c).[16] [17]

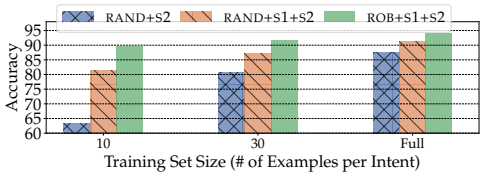


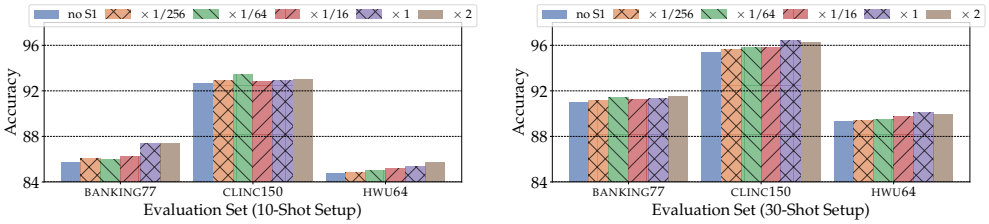
Figure 3: A comparison of a randomly initialised RoBERTa (RAND) against LM-pretrained RoBERTa after S2 CONVFIT-ing with OCL; BANKING77.

Impact of Input LMs. While the results suggest that the CONVFIT framework is applicable and effective with any pretrained LM, the choice of the input LM naturally impacts the absolute ID performance. As expected, the CONVFIT variants with RoBERTa achieve the highest scores across the board. A comparison between DROB and BERT reveals that the pretraining data size and regime seem to play a more critical role than the parameter capacity: the more compact DROB LM is competitive with or even outscores BERT-based variants.[18]

[16] Higher σ scores are desirable as they imply more coherent and compact clusters, and a stronger inter-cluster separation.

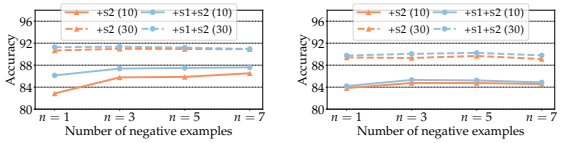
[17] Stage 2 tuning with more in-task data also naturally yields a better separation of examples into coherent clusters , which then naturally improves NN-based classification. For instance, running the ROB+S1+S2-OCL ($n = 3$) variant in 10-shot, 30-shot, and Full data setups yields the respective σ scores for the same set of test examples from Figure 2: $\sigma_{10} = 0.378$, $\sigma_{30} = 0.548$, $\sigma_{Full} = 0.698$, validating the intuition.

[18] Given the versatility of CONVFIT, in future work we plan to extend the experiments to other pretrained LMs such as



(a) 10-shot (ROB+S1+S2-OCL) (b) 30-shot (ROB+S1+S2-OCL)

Figure 4: Varying the amount of Reddit data for Stage 1 CONVFIT; $\times 1$ refers to the Reddit size used in all our other Stage 1 fine-tuning experiments ($\approx$2% of the full Reddit corpus from Henderson et al. (2019a)), while other Reddit data sizes are relative to this corpus size (e.g., $\times 1/32$ means that we use $2\%/32 \approx 0.0625\%$ of the full Reddit corpus). Similar plots (with similar findings) using the COS loss in Stage 2 are available in the Appendix.



(a) BANKING77 (b) HWU64

Figure 5: Impact of the number of negative examples n in 10-shot and 30-shot setups. The CONVFIT variants are ROB+S2-OCL and ROB+S1+S2-OCL (labelled +S1 and +S1+S2 in the figures, respectively).

Variant	10	30	Full
ROB+S1+S2-COS	**82.37**	**94.39**	**98.12**
ROB+S2-COS	70.71	92.14	97.42
MLP-Based			
ROB+S1	48.26	85.49	97.16
USE	47.25	87.21	97.31
LABSE	43.10	87.32	97.42

Table 3: Results on English ATIS (Accuracy $\times 100$).

Importance of LM Pretraining is illustrated by Figure 3. The trend is quite straightforward: semantic knowledge acquired by LM-pretraining is particularly important in the fewest-shot (i.e., 10-shot) setups, and the gap gets reduced with more in-task data available for S2 tuning. However, the gap remains substantial even in the Full setups.

Figure 3 also reveals that the strength of CONVFIT Stage 1 is in adapting the knowledge acquired at LM pretraining: S1 fine-tuning of RAND with smaller amounts of Reddit data cannot match ROB as the input LM, although the gap does become smaller with more in-task data for S2.

Stage 2: Fine-Tuning Losses. Table 2 reveals that strong ID performance after S2 tuning is achieved with different loss functions from §2.2, with different input LMs, even without any careful tuning of hyper-parameters for single settings. This verifies the versatility and robustness of CONVFIT. Both COS and OCL yield consistently strong results, and we expect that even higher absolute scores might be achieved by applying more sophisticated (contrastive learning) loss functions from prior work (Hermans et al., 2017; Liu et al., 2021a) in Stage 2.

4.1 Further Discussion

Stage 1: Amount of Reddit Examples. We now analyse what amount of Reddit data is required to turn input LMs into conversational encoders, by reducing S1 fine-tuning data through subsampling. The scores over different sizes are provided in Figure 4, and we note that they extend to other CONVFIT variants (see §3.1). As expected, having more Reddit data does yield better results on average, but even a small sample of Reddit data (e.g., $\approx$50K (c, r) pairs) **1)** transforms the input LM into an effective sentence encoder (e.g., its MLP-based ID results are on par with those achieved with USE, LaBSE, and ConveRT), and **2)** improves over the CONVFIT variant that skips S2 completely. This implies that perhaps more careful domain-driven data sampling in the future might yield even more domain-adapted conversational encoders after S1.

Amount of Negative Examples in Stage 2 has only a moderate to negligible impact on the final performance, as shown in Figure 5. Small gains when moving from $n = 1$ to $n = 3$ are observed only for the 10-shot setup: there, having more negatives may implicitly play the role of data augmentation for fine-tuning. However, with more in-task examples, the dependence on n becomes inconsequential, and the performance saturates quickly (e.g., see the curves in the 30-shot setups).

Stage 2: Few-Shot versus Full. Framing the ID task a sentence similarity seems especially beneficial for few-shot scenarios, as the model can leverage prototype-based (or instance-based) similarities (Snell et al., 2017) in the specialised encoder space. However, the strong performance with fully CONVFIT-ed models persists also in Full setups.

ELECTRA (Clark et al., 2020) and T5 (Raffel et al., 2020).

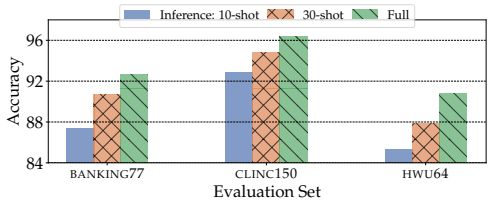


Figure 6: Impact of the number of data instances at inference. The ROB+S1+S2-OCL variant is tuned in 10-shot setups in S2, and additional data (30-shot or Full) is used only at inference without any S2 retuning.

This finding is further corroborated with the results on another standard ID dataset, English ATIS (Hemphill et al., 1990; Xu et al., 2020), see Table 3. There, we observe even more prominent differences in favour of similarity-based ID enabled by CONVFIT, again especially in the two low-data setups. The proposed prototype-based learning and inference holds promise to boost few-shot performance even more in future work, through additional metric learning (Zhang et al., 2020) or data augmentation techniques (Lee et al., 2021).

One limitation of CONVFIT, especially prominent in Full scenarios, is its quadratic time complexity. Future work will look into effective sampling strategies and adaptations towards more sample-efficient and quicker fine-tuning (Tran et al., 2019; Tian et al., 2020; O'Neill and Bollegala, 2021).

Data Augmentation for Inference. Adding more data instances for similarity-based inference, serving as exemplars/prototypes, is likely to boost the final intent detection performance *without the need to retrain the model*. The intuition is that additional instances can provide finer-grained prototypes for inference, semantically more similar to the input query sentences than the original training data. To test this hypothesis, we conduct a simple probing experiment, where we train the ROB+S1+S2-OCL ($n = 3$) variant in the 10-shot setup, but then run inference (i) with the same 10 shots; (ii) in the 30-shot setup (i.e., effectively performing the inference-time data augmentation, relying on 20 more data instances per intent class at inference); (iii) in the Full setup.

The scores are summarised in Figure 6. They clearly indicate that performance does rise with more data instances at inference, even without any model retraining/re-tuning, confirming that increased semantic variability helps at inference. This finding is salient for all three evaluation sets.[19]

As expected, the absolute performance of 30-shot or Full inference when the model is trained in 10-shot setups is lower than in the setup where the more abundant data is additionally used for CONVFIT Stage 2 task-tuning.

Based on these findings, we restate that a promising path for future research concerns investigating and 'task-adapting' automatic paraphrase generation models (Krishna et al., 2020; Dopierre et al., 2021; Schick and Schütze, 2021) such as the one that rely on prompting large models (e.g., GPT-3, T5) (Gao et al., 2021a). Such paraphrases might provide a richer and semantically more varied set of data instances for CONVFIT task-tailored fine-tuning and similarity-based inference.

5 Conclusion and Future Work

We proposed CONVFIT, a two-stage *conversational fine-tuning* procedure that transforms pretrained LMs (e.g., BERT, RoBERTa) into universal (after Stage 1) and task-specialised conversational sentence encoders (after Stage 2) through dual-encoder architectures. The semantic knowledge already stored in the pretrained LMs gets 'rewired' for a particular domain and task. We demonstrated that such task-specialised sentence encoders enable casting intent detection (ID) as simple sentence similarity; CONVFIT-ed encoders yield strong ID results across diverse ID datasets and setups.

The CONVFIT framework is very versatile and opens up many future research paths and further extensions and experimentation beyond the scope of this paper. For instance, it is possible to replace the current contrastive loss functions with other recent effective contrastive losses (van den Oord et al., 2018; Gunel et al., 2021, *inter alia*), or mine hard (instead of using random) negative examples (Lauscher et al., 2020; Kalantidis et al., 2020; Robinson et al., 2021). We will also extend CONVFIT to other pretrained models, experiment with automatic paraphrasers for data augmentation, and port the framework to other conversational tasks (e.g., slot labelling for dialogue), as well as to other, non-dialogue text classification tasks.

Acknowledgements

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

[19] The same trends persist with other CONVFIT variants.

References

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *CoRR*, abs/1606.00372.

Iñigo Casanueva, Tadas Temcinas, 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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval 2017*, pages 1–14.

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.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 250–259.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR 2020*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS 2019*, pages 7057–7067.

Sam Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of ACL 2020*, pages 107–121.

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. *arXiv preprint arXiv:1805.10190*, pages 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.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. ProtAugment: Intent detection meta-learning through unsupervised diverse paraphrasing. In *Proceedings of ACL-IJCNLP 2021*, pages 2454–2466.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *Proceedings of ACL-IJCNLP 2021*, pages 3816–3830.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP 2021*.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michal 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*.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of COLING 2020*, pages 6350–6365.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of AISTATS 2010*, pages 249–256.

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*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL 2020*, pages 8342–8360.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of ICML 2020*, pages 3929–3938.

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

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. 2019a. 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.

Matthew Henderson and Ivan Vulić. 2021. ConVEx: Data-efficient and few-shot slot labeling. In *Proceedings of NAACL-HLT 2021*.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019b. Training neural response selection for task-oriented dialogue systems. In *Proceedings of ACL 2019*, pages 5392–5404.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of ICLR 2020*.

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

Nora Kassner and Hinrich Schütze. 2020. BERT-kNN: Adding a kNN search component to pretrained language models for better QA. In *Findings of EMNLP 2020*, pages 3424–3430.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *Proceedings of ICLR 2020*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of EMNLP 2020*, pages 737–762.

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.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of COLING 2020*, pages 1371–1383.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *CoRR*, abs/2102.01335.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2021. Evaluating multilingual text encoders for unsupervised cross-lingual retrieval. In *Proceedings of ECIR 2021*, pages 342–358.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pre-training for biomedical entity representations. In *Proceedings of NAACL-HLT 2021*.

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

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

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*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*, pages 216–223.

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*.

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 ACL*, 5:314–325.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Proceedings of ECCV 2020*, pages 681–699.

James O'Neill and Danushka Bollegala. 2021. Semantically-conditioned negative samples for efficient contrastive learning. *CoRR*, abs/2102.06603.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

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*.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

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.

Sheikh Muhammad Sarwar, Dimitrina Zlatkova, Momchil Hardalov, Yoan Dinkov, Isabelle Augenstein, and Preslav Nakov. 2021. A neighbourhood framework for resource-lean content flagging. *CoRR*, abs/2103.17055.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *CoRR*, abs/2104.07540.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Patrice Y. Simard, Yann LeCun, and John S. Denker. 1992. Efficient pattern recognition using a new transformation distance. In *Proceedings of NeurIPS 1992*, pages 50–58.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of NeurIPS 2017*, pages 4077–4087.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of CVPR 2018*, pages 1199–1208.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *Proceedings of ICLR 2020*.

Viet-Anh Tran, Romain Hennequin, Jimena Royo-Letelier, and Manuel Moussallam. 2019. Improving collaborative metric learning with efficient negative sampling. In *Proceedings of SIGIR 2019*, pages 1201–1204.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*, pages 6000–6010.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proceedings of NeurIPS 2016*, pages 3630–3638.

Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021. LexFit: Lexical fine-tuning of pretrained language models. In *Proceedings of ACL-IJCNLP 2021*, pages 5269–5283.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of EMNLP 2020*, pages 7222–7240, Online.

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 136–144.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS 2019*, pages 3261–3275.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019b. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of CVPR 2019*, pages 5022–5030.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of ACL 2018*, pages 451–462.

John Wieting, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A bilingual generative transformer for semantic sentence embedding. In *Proceedings of EMNLP 2020*, pages 1581–1594.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT 2018*, pages 1112–1122.

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.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of EMNLP 2020*, pages 5052–5063.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of ACL 2020: System Demonstrations*, pages 87–94.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of IJCAI 2019*, pages 5370–5378.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-Yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of the 3rd Workshop on Representation Learning for NLP*, pages 164–174.

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.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of ACL-IJCNLP 2021*, pages 5751–5767.

A Additional Experiments and Results

Additional experiments and analyses that further support the main claims of the paper have been relegated to the appendix for clarity and compactness of our presentation in the main paper. They largely follow the trends observed in the results which are provided in the main paper. In sum, we provide the following additional results and information, which offer further empirical support of our main claims in this paper:

Table 4 provides the results with all input LMs in our comparison in all the CONVFIT variants discussed in §3.1 across different data setups on all three intent detection datasets. It can be seen as a full (i.e., expanded) version of Table 2 provided in the main paper.

Figure 8 (COS loss in Stage 2) and **Figure 9** (OCL loss in Stage 2) demonstrate the impact of using LM-pretrained Transformers versus randomly initialised Transformers in the CONVFIT framework (both in the full S1+S2 setup, as well as in the setup where only task-tuning (S2) is employed). The patterns in the results, presented over all three evaluation sets, largely follow the patterns observed in Figure 3, which is provided in the main paper.

Figure 10 plots how the amount of Reddit data in Stage 1 impacts the final intent detection performance when the COS loss is used for task-tuning in Stage 2. The observed trends in results are very similar to the ones obtained with the OCL loss, presented in the main paper (see Figure 4).

Figure 11 presents the impact of the number of negative examples n during Stage 2 fine-tuning with the COS loss; the observed trends are very similar to the ones with the OCL loss, presented in the main paper (see Figure 5).

Figure 12 provides t-SNE plots with varying amounts of task data for Stage 2 task-tuning (10-shot versus 30-shot versus Full data setups), demonstrating that very tight and coherent clusters emerge even in the 10-shot setups. **Figure 13** shows t-SNE plots after 10-shot Stage 2, when varying amounts of Reddit data for Stage 1 fine-tuning are used (e.g., skipping Stage 1 completely versus using ≈50k *(context, response)* Reddit pairs). Finally, **Figure 14** demonstrates that the patterns which emerge after Stage 1 and Stage 2 CONVFIT-ing do not depend on the chosen input LM, and on the chosen loss function in Stage 2: the trends very similar to Figure 2 (provided in the main paper) are also observed with *distilRoBERTa* as the input LM, and COS as the S2 loss. **Figure 7** shows visible impact of adaptive Stage 1 fine-tuning even when only $50k$ Reddit *(context, response)* pairs are used.

B Models and Evaluation Data

URLs to the models are provided in Table 6. The intent detection evaluation data is available online:

1. BANKING77, CLINC150, and HWU64 intent detection data have been downloaded from the DialoGLUE repository:
`github.com/alexa/dialoglue`

We use the 10-shot data provided in the repository, and use their script to generate 30-shot setups for all three datasets.

2. The English ATIS intent detection dataset is extracted from the recently published MultiATIS++ dataset (Xu et al., 2020), available here:
`github.com/amazon-research/multiatis`

For reproducibility, we will release the generated 10-shot and 30-shot data splits.

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

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

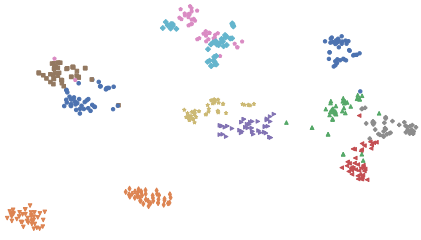
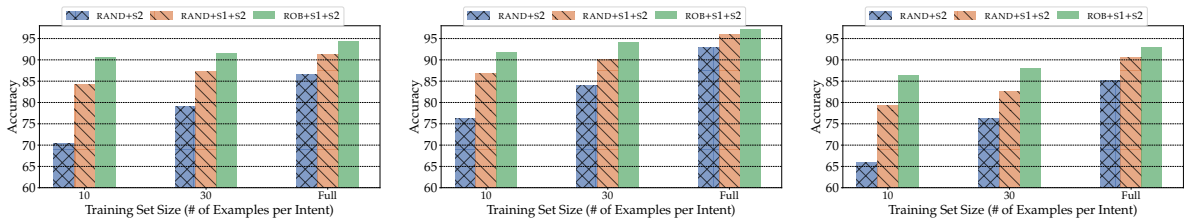


Figure 7: t-SNE plots of encoded utterances from the test set of BANKING77 (a subset of 12 intents, see the legend in Figure 2) after Stage 1 fine-tuning of RoBERTa using only ≈50k *(context, response)* pairs from Reddit; cf., Figure 2a.

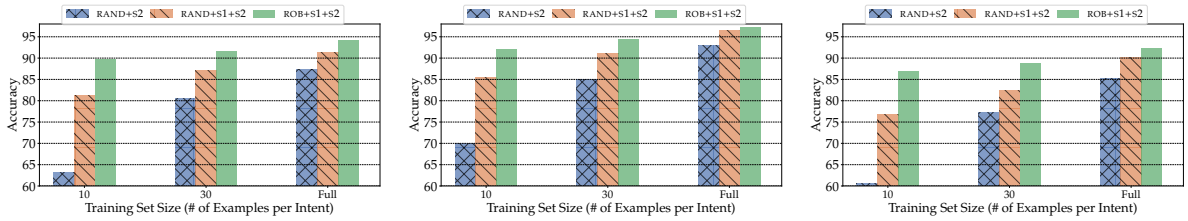
Model Variant	BANKING77			CLINC150			HWU64		
	10	30	Full	10	30	Full	10	30	Full
Similarity-Based Classification									
ROB+S1+S2-COS	86.48	91.33	**94.35**	92.87	95.91	97.20	85.06	**90.46**	**92.98**
BERT+S1+S2-COS	84.32	90.91	93.91	91.80	95.58	96.56	85.13	89.41	91.93
DROB+S1+S2-COS	85.13	90.75	94.06	91.64	95.48	97.00	83.64	89.68	92.94
RAND+S1+S2-COS	79.03	87.37	91.69	83.96	89.98	94.12	76.30	82.62	88.20
ROB+S1+S2-OCL	**87.38**	**91.36**	94.16	**92.89**	**96.42**	**97.34**	**85.32**	90.06	92.42
BERT+S1+S2-OCL	85.97	90.65	93.77	91.53	95.53	96.82	85.04	89.41	92.21
DROB+S1+S2-OCL	86.04	90.78	93.89	91.98	95.60	97.04	83.64	89.50	92.84
RAND+S1+S2-OCL	80.62	87.01	91.49	84.91	90.98	94.44	77.23	82.99	88.85
ROB+S2-COS	84.96	90.81	94.19	91.56	95.64	96.78	84.52	89.87	92.19
BERT+S2-COS	81.27	90.32	93.73	89.58	95.08	96.54	82.90	89.12	91.78
DROB+S2-COS	83.28	90.58	93.91	89.47	95.32	86.78	82.43	89.41	92.10
RAND+S2-COS	70.32	84.16	90.75	76.31	86.69	91.76	65.89	79.18	86.43
ROB+S2-OCL	85.78	90.98	93.77	92.64	95.40	96.87	84.76	89.31	92.01
BERT+S2-OCL	82.28	89.77	93.54	90.71	95.07	96.62	83.09	88.94	92.57
DROB+S2-OCL	82.60	90.65	93.38	90.78	95.02	96.69	81.69	88.75	92.38
RAND+S2-OCL	63.15	81.30	89.71	69.91	85.53	92.18	60.48	76.67	86.90
ROB+S1+S2-SMAX	86.27	90.58	94.06	92.44	95.62	96.76	85.87	88.83	92.48
BERT+S1+S2-SMAX	84.44	90.16	93.09	90.31	93.84	95.91	83.28	88.18	92.29
DROB+S1+S2-SMAX	83.32	89.85	93.47	90.42	94.13	96.47	83.36	88.75	92.57
RAND+S1+S2-SMAX	76.79	85.55	90.97	82.22	87.69	92.91	76.30	81.51	88.85
ROB+S2-SMAX	84.61	90.49	93.66	91.89	95.17	96.71	83.46	88.57	92.57
BERT+S2-SMAX	81.33	89.44	92.63	89.69	93.38	96.12	81.51	87.83	91.58
DROB+S2-SMAX	82.60	89.31	93.54	89.44	93.96	96.04	82.53	87.36	91.91
RAND+S2-SMAX	73.38	83.67	90.32	76.71	85.53	92.62	68.77	79.74	88.94
Baselines: MLP Classification									
ROB+S1	83.08	90.16	93.38	90.98	94.12	96.42	81.13	87.73	91.44
BERT+S1	82.69	89.82	93.67	89.88	94.07	96.33	82.25	88.01	91.12
DROB+S1	82.95	89.55	93.34	89.76	93.46	96.02	81.23	87.64	90.91
CONVERT*	83.32	89.37	93.01	92.62	95.78	97.16	82.65	87.88	91.24
USE*	84.23	89.74	92.81	90.85	93.98	95.06	83.75	89.03	91.25
USE (ours)	82.95	89.09	92.81	90.27	93.54	94.91	82.71	88.20	91.64
LABSE	81.69	88.96	92.60	90.89	93.41	95.12	81.60	86.15	90.99
Baselines: Full Fine-Tuning									
BERT (BASE)**	79.87	–	93.02	89.52	–	95.93	81.69	–	89.97

Table 4: Accuracy scores (×100%) on the three intent detection data sets with varying number of training examples (**10** examples per intent; **30** examples per intent; **Full** training data). As mentioned in §3, $n = 3$ negatives are used in Stage 2 for 10-shot and 30-shot setups, $n = 1$ for the Full setup. The peak scores per column are in bold, the second best is underlined. *The scores were taken directly from prior work, and computed on different 10/30-shot samples (and are thus not directly comparable, Zhao et al. 2021) **The scores achieved by full (regular) fine-tuning of BERT (BASE) have been taken directly from Mehri et al. (2020), and were not available for the 30-shot setup.



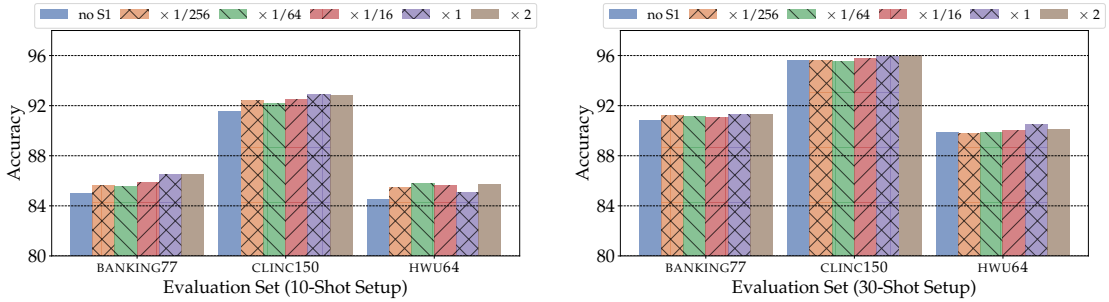
(a) BANKING77 (b) CLINC150 (c) HWU64

Figure 8: A comparison of a randomly initialized BERT or RoBERTa architecture (RAND) with LM-pretrained RoBERTa after Stage 2 CONVFIT-ing; evaluation on all three intent detection datasets; the COS loss used in S2. Figure 9 shows the similar plots with the OCL loss used in S2.



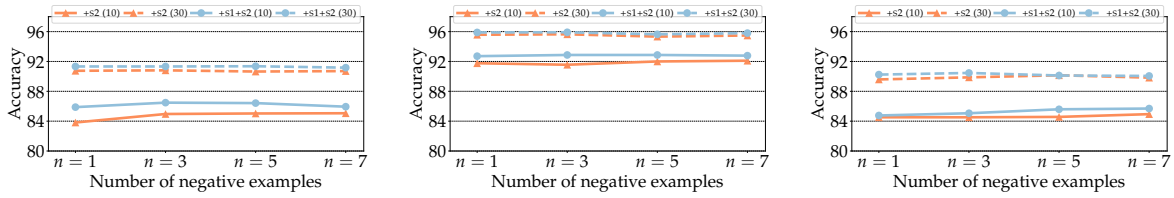
(a) BANKING77 (b) CLINC150 (c) HWU64

Figure 9: A comparison of a randomly initialized BERT or RoBERTa architecture (RAND) with LM-pretrained RoBERTa after Stage 2 CONVFIT-ing; evaluation on all three intent detection datasets; the OCL loss used in S2.



(a) 10-shot (ROB+S1+S2-COS) (b) 30-shot (ROB+S1+S2-COS)

Figure 10: Varying the amount of Reddit data for Stage 1 CONVFIT; $\times 1$ refers to the Reddit size used in all our other Stage 1 fine-tuning experiments ($\approx$2% of the full Reddit corpus from Henderson et al. (2019a)), while other Reddit data sizes are relative to this corpus size (e.g., $\times 1/32$ means that we use $2\%/32 \approx 0.0625\%$ of the full Reddit corpus). Stage 2 loss is COS ($n = 3$).

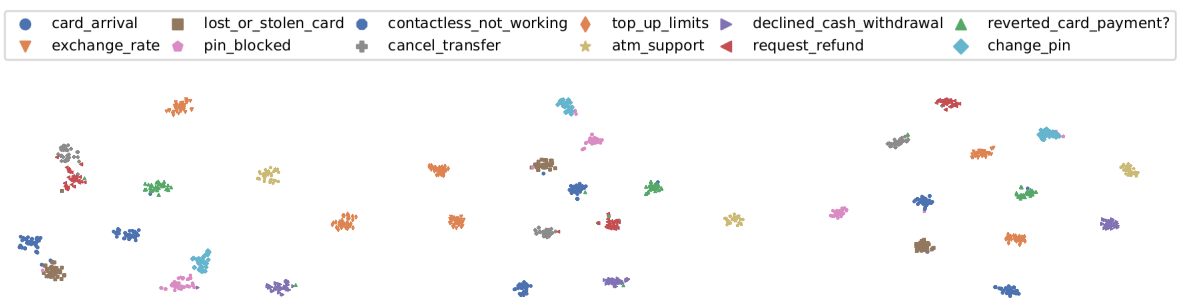


(a) BANKING77 (b) CLINC150 (c) HWU64

Figure 11: Impact of the number of negative examples n on intent detection performance in 10-shot and 30-shot setups. The CONVFIT model variants are ROB+S2+COS and ROB+S1+S2+COS, that is, RoBERTa is the input LM in all experiments, and the results show model variants with the COS loss in Stage 2, without and with S1 fine-tuning (labelled +S2 and +S1+S2 in the figures, respectively).

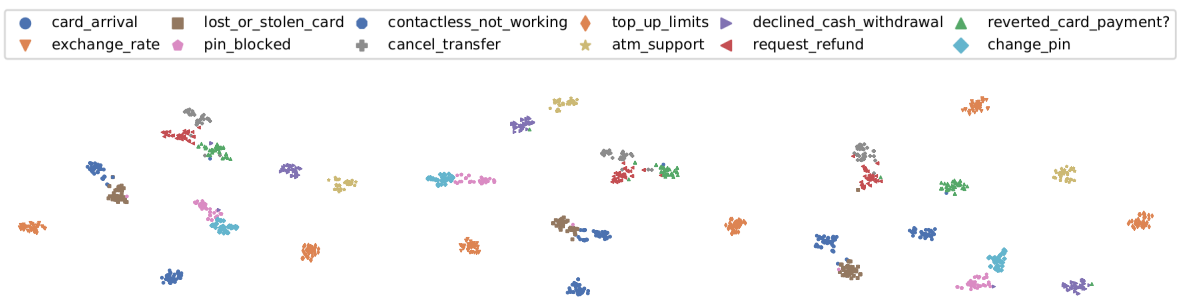
	BANKING77		CLINC150		HWU64	
After	**10**	**30**	**10**	**30**	**10**	**30**
Epoch 1	86.30	91.40	92.80	96.02	**86.15**	**90.33**
Epoch 2	**87.38**	91.36	92.89	**96.42**	85.32	90.06
Epoch 5	87.28	**91.46**	**93.29**	96.32	85.69	89.98

Table 5: Impact of longer Stage 2 CONVFIT-ing on the final performance; ROB+S1+S2-OCL.



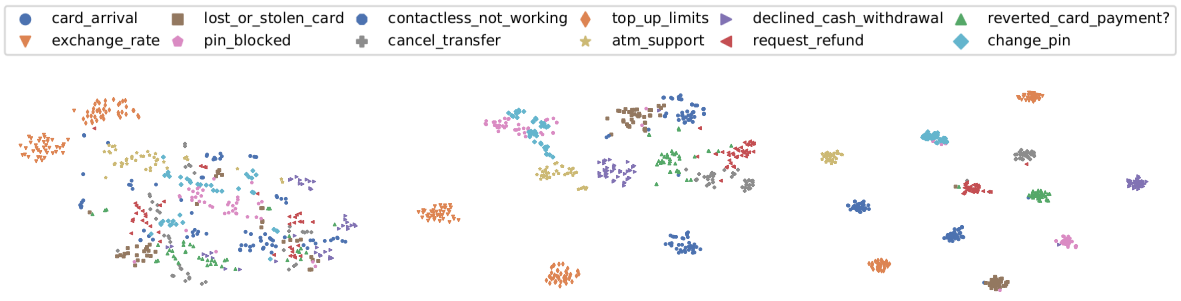
(a) RoBERTa (10-shot S2) (b) RoBERTa (30-shot S2) (c) RoBERTa (Full S2)

Figure 12: t-SNE plots (van der Maaten and Hinton, 2012) of encoded utterances from the test set of BANKING77 (i.e., all examples are effectively unseen by the encoder models at training) associated with a selection of 12 intents. The encoded utterances are created via mean-pooling based on fine-tuned RoBERTa encoders which underwent Stage 1 plus Stage 2 in the **(a)** 10-shot Stage 2 setup (i.e., 10 examples per intent); **(b)** 30-shot setup; **(c)** Full setup (see also §3). Stage 2: fine-tuning with the OCL objective ($n = 3$ negatives). The results suggest that even in 10-shot setups it is possible to learn coherent clusters and clear cluster separations; however, the clusters become less and less compact, and less separated in the semantic space as we fine-tune with fewer in-task instances (e.g., compare the clusters in the 10-shot versus Full setup), and the fine-tuned encoder model is more prone to incorrect cluster assignments. This (initially) visual observation is also supported by the Silhouette coefficient scores (higher is better): (a) $\sigma = 0.378$, (b) $\sigma = 0.548$, (c) $\sigma = 0.698$.



(a) RoBERTa (10-shot S2) (b) RoBERTa (30-shot S2) (c) RoBERTa (Full S2)

Figure 13: t-SNE plots of encoded utterances from the test set of BANKING77 (i.e., all examples are effectively unseen by the encoder models at training) associated with a selection of 12 intents. The encoded utterances are created via mean-pooling based on RoBERTa as the input LM: **(a)** without any Stage 1 fine-tuning with Reddit data; **(b)** Stage 1 fine-tuning with only 50k *(context, response)* Reddit pairs; **(c)** Stage 1 fine-tuning with 2% of the full Reddit corpus of Henderson et al. (2019a) (≈15M pairs). Stage 2 in all three cases is performed in 10-shot setups with the OCL objective ($n = 3$ negatives). The respective Silhouette coefficient scores (higher is better): (a) $\sigma = 0.320$, (b) $\sigma = 0.338$, (c) $\sigma = 0.378$.



(a) DistilRoBERTa (no fine-tuning) (b) DistilRoBERTa (after S1) (c) DistilRoBERTa (after S1 and S2)

Figure 14: t-SNE plots of encoded utterances from the test set of BANKING77 (i.e., all examples are effectively unseen by the encoder models) associated with a selection of 12 intents. The encoded utterances are created via mean-pooling based on **(a)** the original DistilRoBERTa LM; **(b)** DistilRoBERTa after Stage 1 (i.e., fine-tuned on 2% of the full Reddit corpus, see Figure 1); **(c)** DistilRoBERTa after Stage 1 and Stage 2, fine-tuned with the COS objective ($n = 3$ negatives) using the entire BANKING77 training set (see Figure 1).

Name	Abbreviation	URL
bert-base-cased	BERT	huggingface.co/bert-base-uncased
roberta-base	ROB	huggingface.co/roberta-base
distilroberta-base	DROB	huggingface.co/distilroberta-base
LaBSE	LaBSE	huggingface.co/sentence-transformers/LaBSE
multilingual USE	USE	tfhub.dev/google/universal-sentence-encoder-multilingual-large/3

Table 6: URLs of the language models used in this work.