Learning Class-Transductive Intent Representations for Zero-shot Intent Detection

Appendix

Dataset	SNIPS	CLINC
Vocabulary	10,896	6,437
Number of Samples	13,802	9,000
Average Sentence Length	9.05	8.34
Average Label Length	2.43	2.07
Number of existing Intents	5	50
Number of emerging Intents	2	10

Table 1: Dataset statistics

1 A: Datasets Statistics and Construction

The detailed statistics are shown in Table1.

CLINC includes in/out-of-scope queries covering 150 intent classes from 10 domains. We rebuild the labels as follows: We removed the intents containing abbreviations and acronyms like w2, PTO, 410k, mpg, which have no corresponding word vectors in the commonly used word embeddings. Then we remove the intents whose label names that provide little semantic information, such as ”maybe”, ”no”, etc. Next, considering that there are a large number of intents that are too similar to each other, such as “oil change how” and ”oil change when”, etc. We randomly removed some of these intents containing the same words. After the above operations, there are less than 80 intents left. We randomly selected the final 60 intents while ensuring that the selected intents cover almost all different predicates and intents of different lengths (1-4 words). We select 10 unseen intents which makes sure that there are no predicate overlap between different intent names. Among the unseen intent names only 1/3 words appear in the seen intent names. Finally, we reconstruct the CLINC dataset(50 for seen and 10 for unseen) and each intent has only 150 utterances. Therefore CLINC is a very challenging dataset for ZSID and GZSID. SNIPS contains 5 seen intents and 2 unseen intents that are pre-defined.

The label names of SNIPS and CLINC are shown in Table 2 and Table 3 respectively. For SNIPS, we show the number of utterance after the label names.

2 B: Experimental Setup

The detailed hyper-parameter settings are shown in Table 4 (CDSSM+CTIR), Table 5 (Zero-shotDNN+CTIR), Table 6 (CapsNet+CTIR), Table 7 (CNN+CTIR), Table 8 (LSTM+CTIR) and Table 9 (BERT+CTIR). The meaning of the hyper-parameters are summarized as follows:

- **kernel:** The size of each convolution kernel.
- **Cov-dim:** Then number of convolution kernels.
- **MLP-l:** The number of fully-connected layers.
- **MLP-dim:** Hidden dimension of each fully-connected layer.
- α **and** $\lambda^{'}$**:** Down-weighting coefficients that control the importance of SUID in multi-task learning.
- $\alpha\downarrow$ **and** $\lambda^{'}\downarrow$**:** Whether to decay α and $\lambda^{'}$ in the training process.
- **emb:** The type of word embdding. For SNIPS, we use 300-dim embeddings pre-trained on English Wikipedia with 30000 words. For CLINC, we use 300-dim Glove embeddings with 60000 words because some words in CLINC's intents are rare.
- D_h**:** The number of hidden units in LSTM.
- D_a**:** The hidden dimension of the multi-head attention module.
- D_p**:** The dimension of the prediction vector (in capsule network) for each intent.
- R**:** The number of attention heads.
- $Nrouting$**:** the round of Dynamic Routing iterations.
- **bs:** Batch size.
- **opt:** The type of optimizer.
- **stepsize:** The learning rate will be decayed every **stepsize** epochs.
- **gamma:** Decaying rate of the learning rate.
- m^{+}**/**m^{-} **and** $m^{'+}$**/**$m^{'-}$**:** The margins in the max-margin loss.

For the results reported in the paper, we train the models on 24GB Titan RTX GPU. We also report in Table 10 the time consumption of CTIR, which includes the entire process of data loading, model training and inference. The results of Time are reported from the CLINC dataset, which requires more training time than SNIPS. As we can see, CNN+CTIR, LSTM+CTIR and Zero-shotDNN+CTIR cost no more than three minutes. CapsNet+CTIR requires the largest amount of time because of the Dynamic Routing algorithm.

		Seen Intents		
search creative work (1,954)	search screening event (1,960)	play music (2,000)	get weather (2,001)	book restaurant (1,973)
		Unseen Intents		
	add to playlist (1,943)		rate book (1,971)	

Table 2: The unseen and seen intents used in SNIPS.

		Seen Intents		
account blocked	alarm	book flight	book hotel	calendar update
calories	car rental	change language	change user name	confirm reservation
definition	direct despost	expiration date	find phone	flip coin
ingredient substitution	insurance	insurance change	interest rate	international visa
jump start	lost luggage	make call	meaning of life	min payment
next holiday	next song	pin change	play music	plug type
reminder	repeat	restaurant suggestion	roll dice	schedule maintenance
schedule meeting	share location	spending history	taxes	tell joke
todo list	translate	update playlist	weather	what are your hobbies
what song	where are you from	whisper mode	what do you work for	who made you
		Unseen Intents		
bill due	current location	freeze account	how old are you	reset setting
cancel reservation	exchange rate	what is your name	travel alert	shopping list

Table 3: The unseen and seen intents used in CLINC.

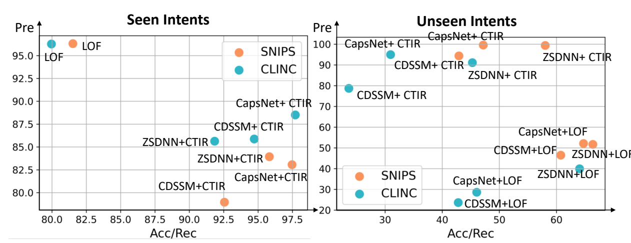

Figure 1: Trade-off between Pre and Acc/Rec. We only have one LOF result each dataset for seen intents because the three systems use the same LOF model in Phase1.

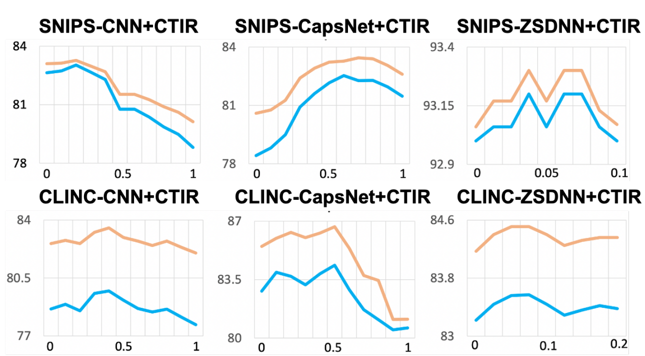

Figure 2: Overall GZSID ACC (orange) and F1 (blue) with the variation of down-weighting coefficients.

3 C: Trade-off Between Precision and Recall

We consider two evaluation metrics, namely Precision (Pre) and Recall (Rec). They are computed with the average value weighted by their support on each class. Therefore, Rec and Acc are exactly the same. As Figure 1 shows, +LOF and CTIR reveal a trade-off between Pre and Acc/Rec. +LOF recalls more unseen intent utterances but at the same time mistakenly classifies some seen intent utterances into y_{unseen}, which hurts Pre on unseen intents and Acc/Rec on seen intents. By contrast, CTIR classifies less utterances to the unseen classes, but at higher precision. Meanwhile, the Acc/Rec on seen intents obvious outstrips +LOF. Table 11 shows the trade-off between Pre and Acc/Rec by inspecting into the performance in each class.

4 D: The Effect of Down-Weighting Coefficients

As shown in Figure 2, the performance varies with the increase of α and λ'. For different models and datasets, the comfortable region is different, but generally, the scores first increases and then declines. This suggests that distinguishing seen and unseen intent is beneficial but paying too much attention to this objective can hurt the final performance.

Datasets	Task	lr	bs	opt	kernel	Cov-dim	MLP-l	MLP-dim	α	$\alpha\downarrow$	emb
SNIPS	ZSID	0.1	256	SGD	3	1000	1	300	0.4	No	wiki
	GZSID	0.1	256	SGD	3	1000	1	300	0.6	No	wiki
CLINC	ZSID	0.15	256	SGD	3	1000	1	300	0.85	Yes	glove
	GZSID	0.1	64	SGD	3	1000	1	300	0.0125	Yes	glove

Table 4: Details of the experimental setup of CDSSM+CTIR of ZSID and GZSID in both datasets.

Datasets	Task	lr	bs	opt	MLP-l	MLP-dim	α	$\alpha\downarrow$	emb
SNIPS	ZSID	0.01	128	Adam	2	300,128	1	No	wiki
	GZSID	0.01	128	Adam	2	300,128	0.0125	No	wiki
CLINC	ZSID	0.001	128	Adam	2	300,128	0.05	No	glove
	GZSID	0.001	64	Adam	2	300,128	0.05	Yes	glove

Table 5: Details of the experimental setup of ZSDNN+CTIR of ZSID and GZSID in both datasets.

Datasets	Task	lr	bs	opt	$\lambda^{'}$	$\lambda^{'}\downarrow$	D_h	D_a	D_P	R	$Nrouting$	m^+ / m^-	$m^{'+} / m^{'-}$	emb
SNIPS	ZSID	0.0001	64	Adam	0.5	No	32	20	10	3	2	0.1,0.9	0.01,0.99	wiki
	GZSID	0.0001	64	Adam	0.5	No	32	20	10	3	2	0.1,0.9	0.01,0.99	wiki
CLINC	ZSID	0.001	256	Adam	0.05	No	256	60	30	3	2	0.1,0.9	0.01,0.99	glove
	GZSID	0.001	256	Adam	0.05	No	256	60	30	3	2	0.1,0.9	0.01,0.99	glove

Table 6: Details of the experimental setup of CapsNet+CTIR of ZSID and GZSID in both datasets.

Datasets	Task	lr	bs	opt	kernel	Cov-dim	MLP-l	MLP-dim	α	$\alpha\downarrow$	emb
SNIPS	ZSID	0.01	128	Adam	3	300	2	32,16	0.05	No	wiki
	GZSID	0.001	128	Adam	3	300	2	32,16	0.025	No	wiki
CLINC	ZSID	0.003	256	Adam	3	300	2	128,96	0.025	No	glove
	GZSID	0.003	256	Adam	3	300	2	128,96	0.025	No	glove

Table 7: Details of the experimental setup of CNN+CTIR of ZSID and GZSID in both datasets.

Datasets	Task	lr	bs	opt	MLP-l	MLP-dim	α	$\alpha\downarrow$	D_h	emb
SNIPS	ZSID	0.001	128	Adam	2	64,32	0.05	No	64	wiki
	GZSID	0.01	128	Adam	2	128,32	0.05	No	128	wiki
CLINC	ZSID	0.01	256	Adam	2	64,32	0.05	No	64	glove
	GZSID	0.01	256	Adam	2	64,32	0.05	No	64	glove

Table 8: Details of the experimental setup of LSTM+CTIR of ZSID and GZSID in both datasets.

Datasets	Task	lr	bs	opt	stepsize	gamma	α	$\alpha\downarrow$
SNIPS	ZSID	0.0001	256	Adam	10	0.01	0.05	No
	GZSID	0.0001	256	Adam	10	0.01	0.05	No
CLINC	ZSID	0.0001	512	Adam	5	0.5	0.1	No
	GZSID	0.0001	512	Adam	5	0.5	0.1	No

Table 9: Details of the experimental setup of BERT+CTIR of ZSID and GZSID in both datasets.

Model	**CNN+CTIR**	**LSTM+CTIR**	**CapsNet+CTIR**	**Zero-shotDNN+CTIR**
Parameters	81M	70M	95M	69M
Time	2.9min	2.85min	23min	1.25min

Table 10: The computational requirements of four typical CTIR models on a 2.20GHz Intel Xeon CPU. The Time here measures the entire process including data loading, model training and inference.

Intent	CapsNet Pre		CapsNet Acc/Rec		CDSSM Pre		CDSSM Acc/Rec		Zero-shotDNN Pre		Zero-shotDNN Acc/Rec	
	+LOF	+CTIR	+LOF	+CTIR	+LOF	+CTIR	+LOF	+CTIR	+LOF	+CTIR	+LOF	+CTIR
SearchCreativeWork	**95.06**	56.32	74.87	**93.68**	**95.06**	48.23	74.87	**81.54**	**95.06**	61.87	74.87	**91.95**
SearchScreeningEvent	**98.31**	92.50	77.02	**96.59**	**98.31**	92.01	77.02	**88.40**	**98.31**	95.74	77.02	**91.98**
PlayMusic	**89.85**	75.44	80.50	**99.50**	**89.85**	81.60	80.50	**95.49**	**89.85**	76.55	80.50	**99.50**
GetWeather	**100.00**	97.84	85.32	**99.83**	**100.00**	81.48	85.32	**95.42**	**100.00**	93.27	85.32	**98.64**
BookRestaurant	**98.30**	97.55	89.79	**99.67**	**98.30**	92.44	89.79	**98.00**	**98.30**	96.59	89.79	**99.33**
AddToPlaylist	78.60	**100.00**	36.22	**45.27**	40.88	**98.00**	**85.03**	67.64	43.38	**99.48**	**84.69**	65.23
RateBook	41.49	**100.00**	**96.08**	47.12	51.85	**89.73**	**35.84**	22.35	61.46	**99.32**	**50.34**	50.00
Overall	86.04	**88.52**	77.15	**83.22**	82.15	**83.37**	75.56	**78.55**	83.86	**88.97**	77.56	**85.35**

Table 11: Per class performance of GZSID in SNIPS. “AddToPlaylist” and “RateBook” are unseen intents and the others are seen intents. The Overall scores are reported using the average value weighted by their support on each class.