Intent classification with five examples

Arijit Mukherjee
arijitmukh007@gmail.com

Abstract

Intent classification is an important part of most modern day dialogue agents. Though deep learning based methods have showed promising results in NLP tasks, such method requires large amount of labeled data. For intent classification collecting large number of labeled data for every domain is not possible, so most of the dialogue agents rely on traditional feature based methods for intent classification. In this paper we show that we can achieve comparable results with fully supervised in-domain trained state of the art network for intent classification using currently available tools, while using as low as 5 examples from a new domain.

1 Introduction

Intent classification is a core part of most modern day dialogue agents. Dialogue agents when deployed to a new domain requires data labeled with their intents for the new domain. Most dialogue agents framework uses traditional statistical methods to classify intents due to the fact that very less amount of examples are available from the new domain. Although deep learning based methods have performed well in several NLP tasks, they require large amount of labeled data. The amount of data restricts deep learning based text classification frameworks to be used in such cases like intent classification. To counter the problem of extending the label space or easily generalizing to new label space few-shot learning aims to learn a network which can generalize well to samples which are not seen while training. Metric learning is one of the methods for few-shot learning which try to learn a similarity on a representation that can generalize over unseen samples. Recent availability of large scale pre-trained sentence encoder like Deep bidirectional transformers for language understanding (BERT) (Devlin et al., 2018) have made it possible to approach NLP tasks with small data. BERT is trained on a tweaked language modeling objective, have shown that a general purpose model can perform well when fine-tuned on a particular task.

Sentence	Intent
i want to fly from baltimore to dallas round trip	atis_flight
which airlines fly from boston to washington dc via other cities	atis_airline
listen to westbam alumb allergic on google music	PlayMusic
i give this current textbook a rating value of 1 and a best rating of 6	RateBook

Table 1: Examples from ATIS and SNIPS, (1)(2) belongs to ATIS and (3)(4) belongs to SNIPS

In this paper we combine the best of few-shot learning and general purpose sentence encoder like BERT to a real life application of few-shot classification. Traditionally in case of intent classification we trained our model on a given train set to classify the intent on a held out test set. In this fully supervised setup the label space is same in train and set set. The domain from which train and test set is collected is also the same. We propose a method where we show that we train a network on ATIS data-set which collected from air traffic related queries while test our network on SNIPS test set which comes from SNIPS personal assistant queries and still achieve comparable result accuracy with the fully supervised setup. Our proposed setup uses already existing tools like Prototypical Network (Snell et al., 2017) for metric learning and BERT (Devlin et al., 2018) to encode the utterances to fixed size vectors. We show that not

only our method is able to generalize to unseen classes but also it is able to provide the power to generalize to unseen samples coming from completely different domain.

2 Related Works

Our problem have four main aspects few-shot classification, sentence representation, intent classification and few-shot intent classification. We will briefly review the work done on each one of the area.

2.1 Few-shot Classification

Few-shot classifier tries to generalize an already learnt model to a new set of classes. There has been a lot of work done on few-shot learning. We are particularly interested in one approach called metric-learning. In metric learning the network learns to compare two samples and gives a score of similarity between them. More specifically it learns to represent the sample in such a way that if two samples are similar the distance metric between them is small and vice-versa. Once we have a network trained using this methodology we can easily extend the class space to new classes with very small number of examples. In matching network (Vinyals et al., 2016) used cosine similarity. Prototypical Network (Snell et al., 2017) used euclidean distance from the mean of the class representation to assign class labels. Relation Network (Sung et al., 2017) uses a CNN to map the relation between the query embedding and support set embedding.

2.2 Sentence representation

Traditional methods to represent sentences to fixed size vectors included extracting features like tf-idf to various other non-neural methods. Later methods like word embedding (Turian et al., 2010) played a significant role in NLP research. Neural based sentence representation have come a long way starting from ELMO (Peters et al., 2018) which gave contextual representation of words to OpenAI GPT (Radford et al., 2019) which showed a general purpose model trained on large corpus can then be fine-tuned to downstream tasks to achieve state of the art results with minimal task specific addition to the network. Further BERT (Devlin et al., 2018) improved upon OpenAI GPT and achieved state of the art results on various downstream tasks. In our approach we use BERT to encode variable length utterances to fixed sized vectors.

2.3 Intent classification

Intent classification is key task for language understanding. Various deep learning based methods are explored in this area. Slot-filling is a very closely related task with intent classification. Some approaches try to model intent classification independently from slot filling. While others try to model both task jointly. Independent intent classification includes character level convolution (Zhang et al., 2015), LSTM based utterance classifier (Ravuri and Stolcke, 2015).While (Guo et al., 2014) modeled intent classification and slot-filling jointly using recursive neural networks. (Goo et al., 2018) proposed a sloted-gated attention based model for the same. Recently (Chen et al., 2019) jointly modeled intent classification and slot-filling with BERT and achieved state of the art results in two data-sets, ATIS (Guo et al., 2014) and SNIPS (Coucke et al., 2018).

2.4 Few-shot text classification

Few-shot classification in NLP is a relatively new area. (Yu et al., 2018) used weighted combination of different metric and showed improvement on intent classification, (Bailey and Chopra, 2018) used pre-trained word embedding with human in the loop for few-shot text classification. (Xia et al., 2018) used Capsule Network originally proposed by (Sabour et al., 2017) for zero-shot intent classification. Recently (Geng et al., 2019) used Induction Network for few-shot text classification.

Traditionally few-shot learning methods showed good results when the label space was extended in the same domain, while we show comparable results with fully supervised setup when label space is extended that too in a different domain. Such generalizing is due to having BERT in the whole setup.

3 Our Approach

We propose a network combining the representation power of BERT with the metric learning capacity of Prototypical Network to classify sentences to intent. We first explain how a Prototypical Network works and then how we use BERT with the setup.

3.1 Prototypical Network

Any metric learning method relies on computing the distance/similarity of a query point from few examples of support points. We denote the support points as support set. There can be N_w number of classes and N_s number of samples for each class in the support set. We call the whole setup as N_w way N_s shot.

Prototypical network computes a M-dimensional representation c_k for each class k in the support set and q for the query example. Where c_k is calculated as,

$$c_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} f_\phi(x_i)$$

where S_k is the support set of class k. While q is calculated as follows,

$$q = f_\phi(x_q)$$

Finally $Pr(y_q = k|x_q)$ is soft-max over the euclidean distance between the query prototype q and class specific support prototype c_k.

$$Pr(y_q = k|x_q) = \frac{\exp(-||q - c_k||_2)}{\sum_{k'} \exp(-||q - c_{k'}||_2)}$$

3.2 Combining BERT

In the Prototypical Network formulation f_ϕ is used to embed samples to a M dimensional representation. In our case we need a network to encode our variable length sentences to fixed size representation. We have many design choices here, traditionally a LSTM would have been be a good choice to encode sentences. But it is difficult to learn a good LSTM based network for our purpose with little amount of data. BERT is a far better choice as it showed it performed really well when fine-tuned on downstream tasks, also as BERT is already trained on a large corpus so we have a fairly good starting representation for sentences. Thus we use a pre-trained BERT-Base model to initialize f_ϕ, afterwards we keep fine-tuning it. Specifically we use the hidden state of the final layer of BERT corresponding to the [CLS] token as $f_\phi(sentence)$

4 Experiment

We perform our experiment on two data-set for intent classification and slot filling **ATIS** (Guo et al., 2014) and **SNIPS** (Coucke et al., 2018). We used the same splits as done by (Goo et al., 2018) in their experiments.

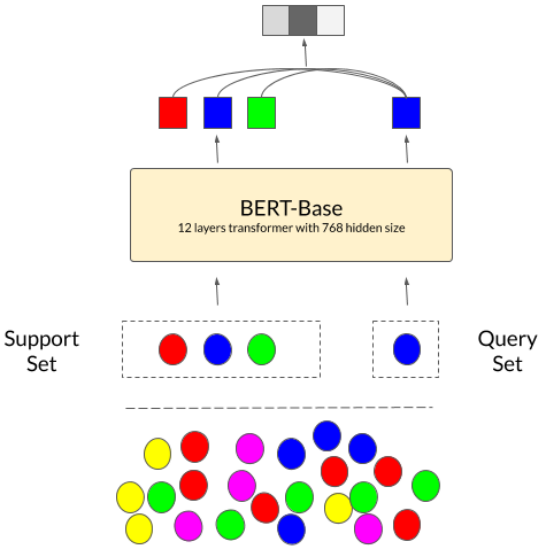

Figure 1: BERT with Prototypical Network

4.1 Data

ATIS data-set is mainly used for NLU benchmarks, it consists of spoken utterances of flight booking instructions. The data-set contains 4478 training, 500 validation and 893 test examples. There are 21 unique intent labels, every sentence is assigned one intent among them. Further we omitted examples with intent *atis_meal, atis_ground_service#atis_ground_fare, atis_aircraft#atis_flight#atis_flight_no, atis_airline #atis_flight_no, atis_restriction, atis_cheapest* as it contained less than or equal to six examples, which is not sufficient to construct our support set and query set simultaneously . SNIPS data-set is user utterances collected from the SNIPS voice assistant platform. It contains 13,084 training, 700 validation and 700 test examples from 7 intent classes.

4.2 Setup

We used pre-trained BERT-base-uncased model for our experiment which have 12-layer, 768-hidden, 12-heads and 110M parameters. We trained our network on ATIS and while testing we test our network on SNIPS. We followed the procedure provided by (Snell et al., 2017), all our experiments consisted of 10-way while training and 5-way while testing. We also kept the number of shots equal for training and testing. While training for each iteration we uniformly sampled 10 classes, then we sampled either 1 or 5 example (depending on 1-shot or 5-shot) from each of those

Experiment	Accuracy
Fully Supervised (Chen et al., 2019)	98.6
1-shot	73.3
5-shot	89.1

Table 2: Experimental Results (1) Fully supervised training on SNIPS and testing on SNIPS, (2)(3) Our model few-shot training on ATIS testing on SNIPS

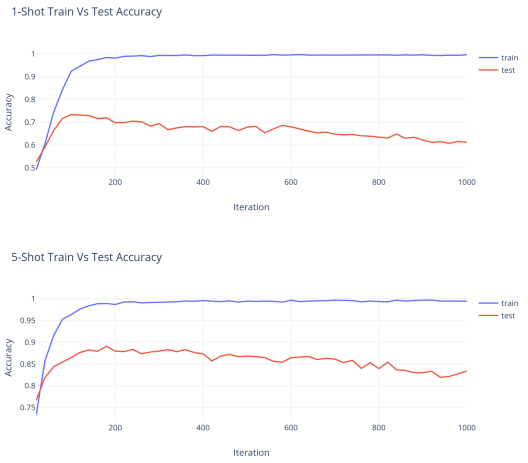

Figure 2: Train Vs Test Accuracy for 1-shot and 5-shot intent classification

classes and constructed our support set. Later from each of the sampled 10 classes we sample another 1 sample which is not part of the previously sampled examples and constructed our query set. Both support-set and query-set are constructed as batch. We passed both of the batches separately through BERT and take the hidden state corresponding to the [CLS] from the final layer. Following (Snell et al., 2017) testing strategy for prototypical networks we generate 100 random batches from our test data and compute the average accuracy over it.

5 Results

We compare our results to the state of the art model for joint slot filling and intent classification with BERT (Chen et al., 2019), which is trained in a fully supervised manner on SNIPS data. Whereas our model have never seen examples from SNIPS while training. We ran each of our experiments for 1000 iterations over 10 independent runs. Figure 2 shows the training and testing accuracy averaged over 10 independent runs. While table 2 shows the average accuracy we get for our 1-shot and 5-shot experiment.

While previous few-shot classifiers showed that it is able to generalize to new labels, we show here that our method is able to generalize over new labels coming from completely new domain. Using BERT as f_ϕ actually gives it the power to generalize across domain.

6 Conclusion

We proposed a few-shot learning framework using BERT and Prototypical Network. We showed that it can achieve comparable accuracy to a fully supervised trained state of the art intent classifier just using 5 examples from the new domain. We show that its is possible to train a sentence classifier where the use case demands extending the label space to new domain with very small number of samples. Further we are looking into whether such metric learning objective can actually help language modeling and improve BERT. We are also looking to how we can improve upon Prototypical Network and adapt the notion of metric learning better for NLP tasks like entailment.

References

Katherine Bailey and Sunny Chopra. 2018. Few-shot text classification with pre-trained word embeddings and a human in the loop. *CoRR*, abs/1804.02063.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. 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, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ruiying Geng, Binhua Li, Yongbin Li, Yuxiao Ye, Ping Jian, and Jian Sun. 2019. Few-shot text classification with induction network. *CoRR*, abs/1902.10482.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 753–757, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Guo, Gökhan Tür, Wen tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 554–559.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Suman V. Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In *INTERSPEECH*.

Sara Sabour, Nicholas Frosst, and Geoffrey Hinton. 2017. Dynamic routing between capsules.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4077–4087. Curran Associates, Inc.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2017. Learning to compare: Relation network for few-shot learning. *CoRR*, abs/1711.06025.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. *CoRR*, abs/1606.04080.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. *CoRR*, abs/1809.00385.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215, New Orleans, Louisiana. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626.