Are Pre-trained Transformers Robust in Intent Classification? A Missing Ingredient in Evaluation of Out-of-Scope Intent Detection

Anonymous ACL submission

CLINC-Single-Domain-OOS	K	Train	Dev.	Test
In-scope	10	500	500	500
ID-OOS	-	-	400	350
OOD-OOS	-	-	200	1000
BANKING77-OOS	K	Train	Dev.	Test
In-scope	50	5905	1506	2000
ID-OOS	-	-	530	1080
OOD-OOS	-	-	200	1000

Table 1: Statistics of CLINC-Single-Domain-OOS and BANKING77-OOS dataset.

Appendix

A Dataset Construction

For each domain, the original CLINC dataset has 100, 20, and 30 examples for each in-scope intent, and 100, 100, and 1000 OOD-OOS examples for the train, development, and test sets, respectively. To incorporate the ID-OOS intents, we reorganize the original dataset. For each in-scope intent in the training set, we keep 50 examples as a new training set, and move the rest 30 examples and 20 examples to the development and test sets through random sampling. For the examples of each ID-OOS intent in the training set, we randomly sample 60 examples, add them to the development set, and add the rest of the 40 examples to the test set. We move the unused OOD-OOS examples of the training set to the validation set and keep the OOD-OOS test set unchanged. For the BANKING77 dataset, we move the training/validation/test examples of the selected 27 intents to the ID-OOS training/validation/test examples, and we copy the OOD-OOS examples of CLINC as the OOD-OOS examples of BANKING77.

We name the two new datasets as `CLINC-Single-Domain-OOS` and `BANKING77-OOS`, respectively. Table 1 shows the statistics of these two new datasets. Table 2 and Table 3 show which intent labels are treated as ID-OOS for the CLINC dataset and BANKING77 dataset, respectively.

B Experimental Setting

For each component related to the five pre-trained models, we use their base configurations. we use the `roberta-base` configuration for RoBERTa; `bert-base-uncased` for BERT; `albert-base-v2` for ALBERT; `electra-base-discriminator` for ELECTRA; `tod-bert-jnt-v1` for ToDBERT. All the model parameters are updated during the fine-tuning process. where we use the AdamW (Hendrycks et al., 2020) optimizer with a weight decay coefficient of 0.01 for all the non-bias parameters. We use a gradient clipping technique (Pascanu et al., 2013) with a clipping value of 1.0, and also use a linear warmup learning-rate scheduling with a proportion of 0.1 w.r.t. to the maximum number of training epochs.

For each model, we perform hyper-parameters searches for learning rate values $\in \{1e-4, 2e-5, 5e-5\}$, and the number of the training epochs $\in \{8, 15, 25, 35\}$. We set the batch size to 10 and 50 for CLINC and BANKING77, respectively. We take the hyper-parameter sets for each experiment and train the model ten times for each hyper-parameter set to select the best threshold δ (introduced in Section **??**) on the development set. We then select the best hyper-parameter set along with the corresponding threshold. Finally, we apply the model and the threshold to the test set. Experiments were conducted on single NVIDIA Tesla V100 GPU with 32GB memory.

C More Results

Figure 1 shows the model confidence level on the development set of the "Credit cards" domain. We can see that RoBERTa is relatively more robust with limited data. Figure 2 shows the confusion matrices of RoBERTa w.r.t. the "Credit cards" domain. The model is confused to identify ID-OOS intents.

Domain	IN-OOS	In-scope
Banking	balance, bill_due, min_payment, freeze_account, transfer	account_blocked, bill_balance, interest_rate, order_checks, pay_bill, pin_change, report_fraud, routing, spending_history, transactions
Credit cards	report_lost_card, improve_credit_score, rewards_balance, application_status, replacement_card_duration	credit_score, credit_limit, new_card, card_declined, international_fees, apr, redeem_rewards, credit_limit change, damaged_card expiration_date

Table 2: Data split of the ID-OOS and in-scope intents for the CLINC dataset.

ID-OOS	"pin_blocked", "top_up_by_cash_or_cheque" "top_up_by_card_charge", "verify_source_of_funds", "transfer_into_account", "exchange_rate", "card_delivery_estimate", "card_not_working", "top_up_by_bank_transfer_charge", "age_limit", "terminate_account", "get_physical_card", "passcode_forgotten", "verify_my_identity", "topping_up_by_card", "unable_to_verify_identity", "getting_virtual_card", "top_up_limits", "get_disposable_virtual_card", "receiving_money", "atm_support", "compromised_card", "lost_or_stolen_card", "card_swallowed", "card_acceptance", "virtual_card_not_working", "contactless_not_working"

Table 3: Data split of the ID-OOS intents for the BANKING77 dataset. Where 27 intents are randomly selected as ID-OOS intents and the rest are treated as in-scope intents.

Figure 3 shows the tSNE visualizations for ID-OOS intents in the "Banking" domain. The models struggle to classify the ID-OOS intents even with more data.

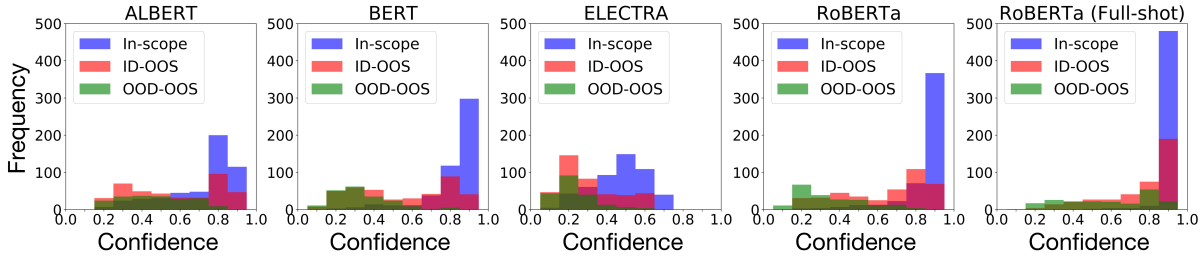

Figure 1: Model confidence on the development set of the "Credit cards" domain in CLINC-Single-Domain-OOS dataset under 5-shot setting. Darker colors indicate overlaps.

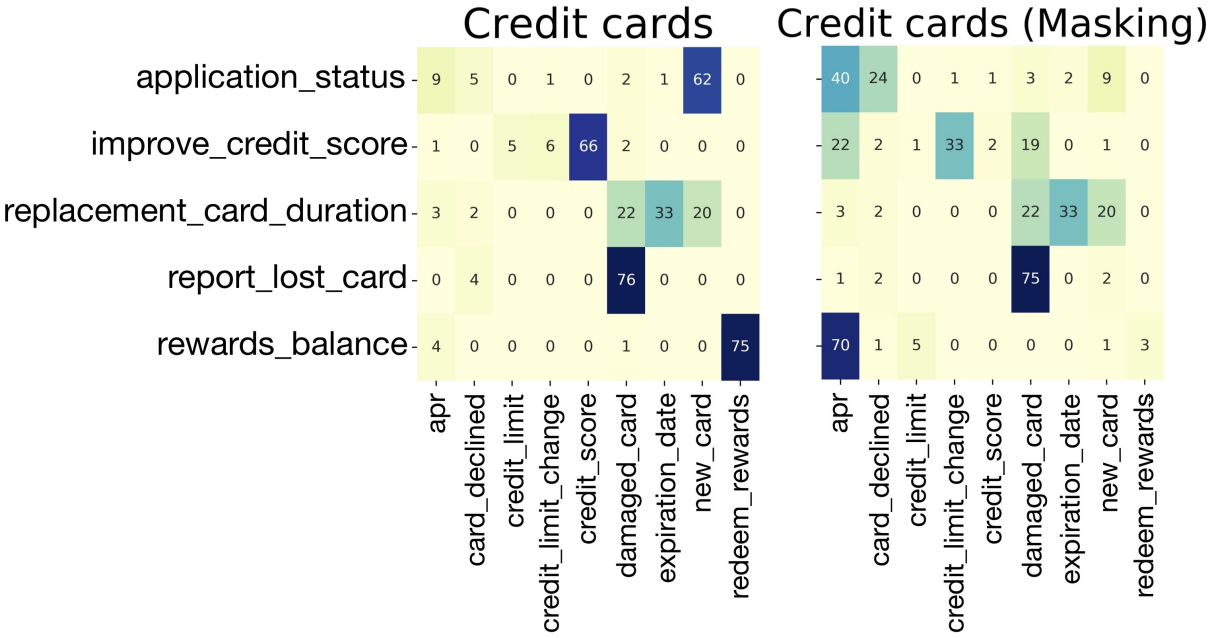

Figure 2: Full-shot confusion matrices on the development set with and without masking ("Credit cards", RoBERTa). Vertical axis: ID-OOS; horizontal axis: in-scope (only predicted intents considered).

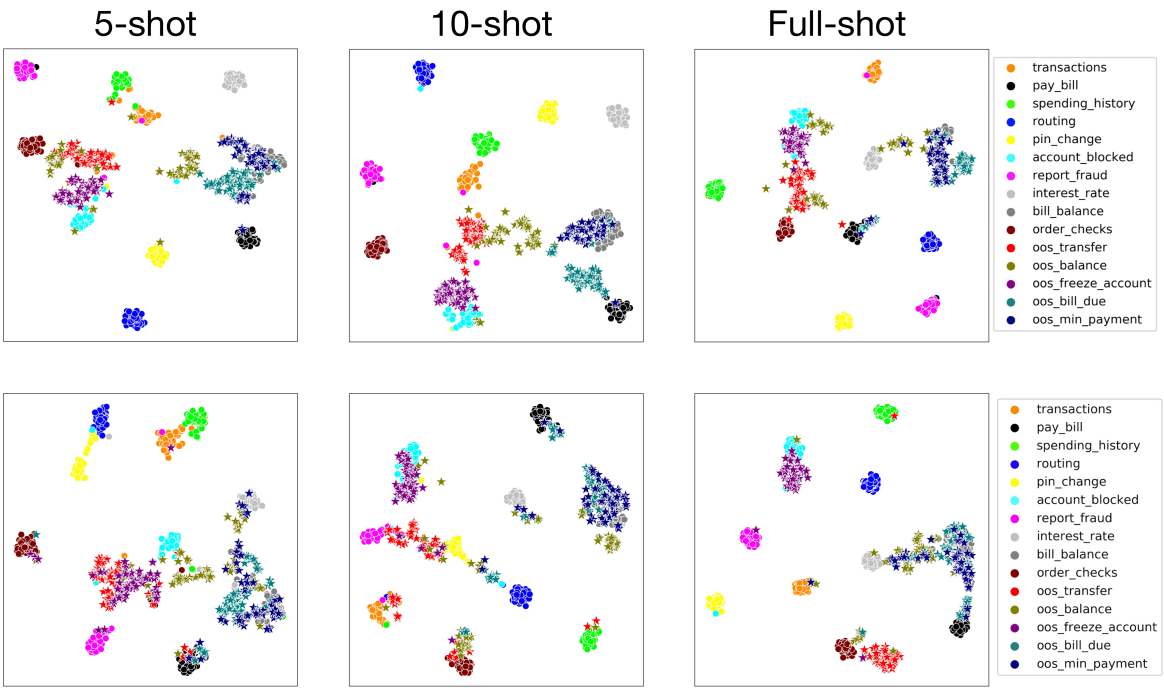

Figure 3: RoBERTa (first row) and ELECTRA (second row) tSNE visualizations on the development set of the "Banking" domain in CLINC-Single-Domain-OOS dataset.

References

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained Transformers Improve Out-of-Distribution Robustness. *arXiv preprint arXiv:2004.06100*.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 1310–1318.