Multi-task Learning with AdapterFusion

Presented by:

Anjali Grover

28.09.2021

Outline


1. Objective
2. Related work
3. Data Source & Distribution
4. Baselines
5. Evaluation & Error Analysis
6. Conclusion & Future work

> *Imagine if a doctor can get all the information she needs about a patient in 2 minutes and spend the next 13 minutes of a 15-minute office visit talking with the patient, instead of spending 13 minutes looking for information and 2 minutes talking with the patient.*

- Lynda Chin, a renowned cancer genomic scientist

Clinical Outcome Prediction Tasks

Present Illness: 58yo man w/ hx of hypertension, AFib on coumadin[...]

Family History: Mother had stroke at age 82. Father unknown

Physical Exam: Vitals: P: 92 R: 13 BP: 151/72, SaO2: 99% intubated. GCS E: 3 V:2 M:5 HEENT: atraumatic, normocephalic Pupils: 4-3mm [...]

Social History: Lives with wife. 25py. No EtOH

Clinical Note at Admission (MIMIC III)

Deep Neural Networks

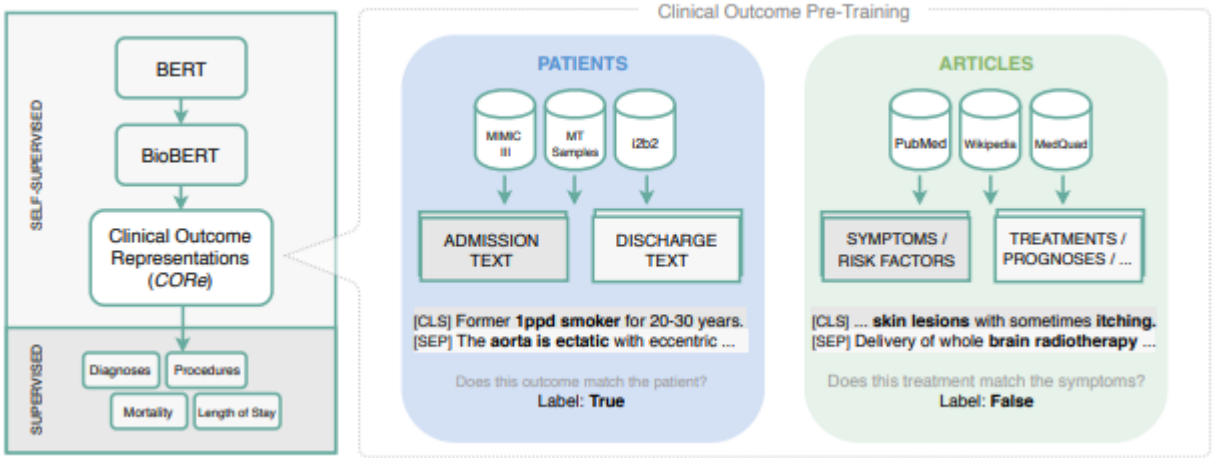
Diagnosis	
300	Neurotic Disorders
401	Hypertension

Procedures	
431	Gastrostomy
311	Tracheostomy

In-Hospital Mortality
Not deceased

Length of Stay
>14 days

The CORe Approach

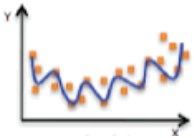


Schematic demonstration of Clinical Outcome Pre-Training

The CORe Approach : Problems

More Resources

High Training Time

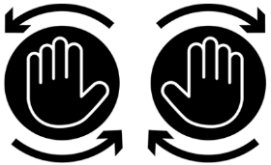


Overfitting

Unclear Relation between Tasks

Multi-task Learning : Humans

When humans learn new tasks, we take advantage of the knowledge gained from related tasks

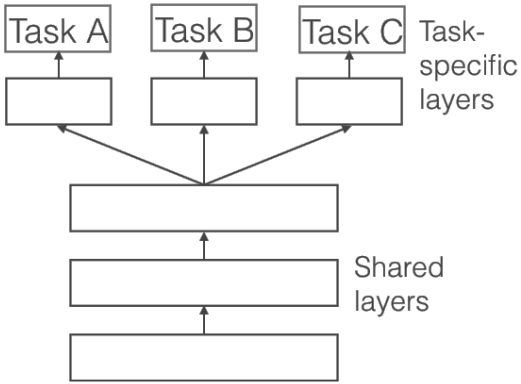


Related Task : Floor Sanding

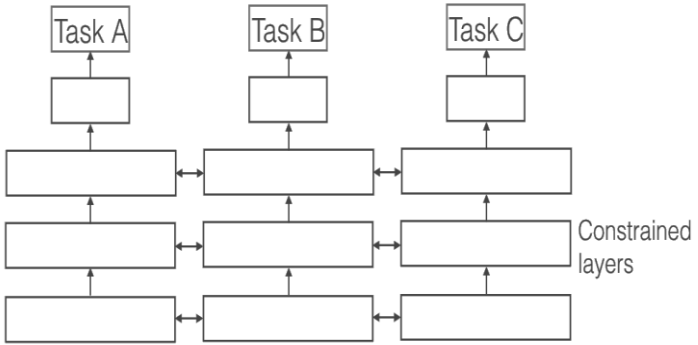
Target Task : Karate

Multi-task Learning : Machine Learning

Multi-task learning enables sharing representations between related tasks, making the model to generalize better on the original task



Our Focus : Hard Parameter Sharing



Soft Parameter Sharing

Multi-task Learning : Clinical Outcome Prediction

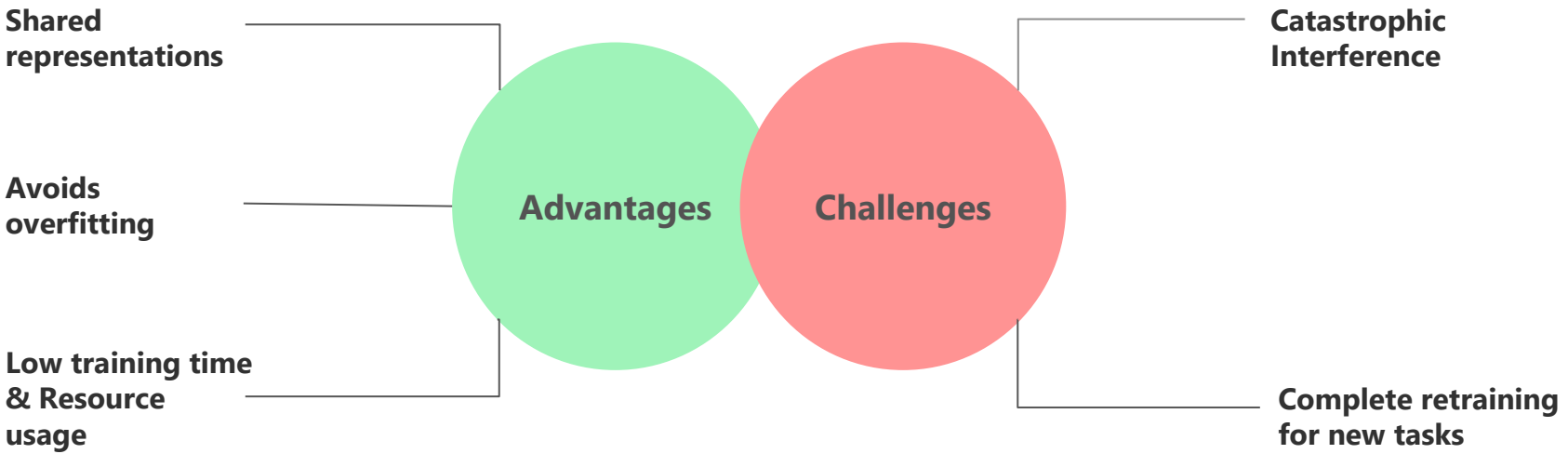
Heart Failure

High Mortality

Diabetes

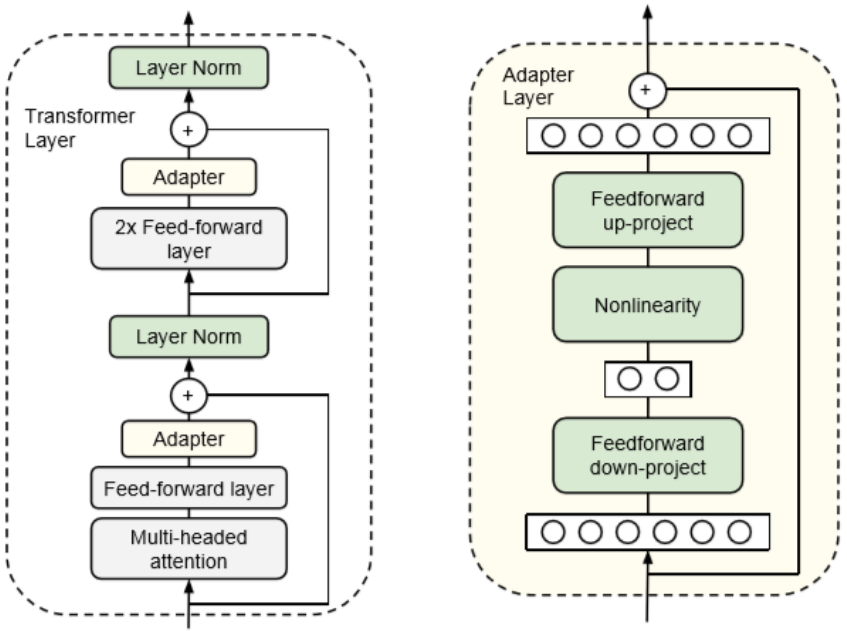
Increased Hospital stays

Multi-task Learning : Pros & Cons



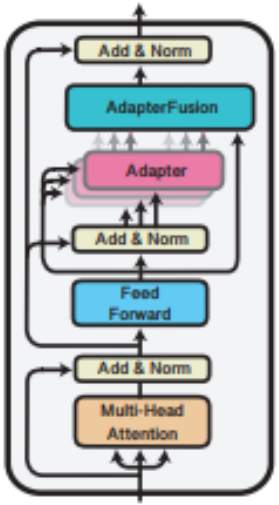
Adapters

- Only use a few task specific parameters
- No change required in the underlying model
- Yields Compact & Extensible models



Adapter Architecture

Single Task AdapterFusion



AdapterFusion Architecture

- **Knowledge Extraction**: Train the adapters for each of the N tasks independently to retrieve *Single-Task Adapters*
- **Knowledge Combination:** Combine N adapters using AdapterFusion

$$\Psi_m \leftarrow \underset{\Psi}{argmin}[L_m(D_m; \Theta, \Phi_1, ..., \Phi_n, \Psi)]$$

Ψ_m AdapterFusion parameters for target task m

Θ Parameters across all tasks

Φ_n Task specific parameters

Hypothesis

1. *Multi-task learning suffers from Catastrophic Interference and does not surpass the CORe approach.*

2. *AdapterFusion mitigates the Catastrophic Interference problem and surpasses the CORe approach.*

3. *Additionally, we expect AdapterFusion to perform better than other approaches in terms of training time and resource usage.*

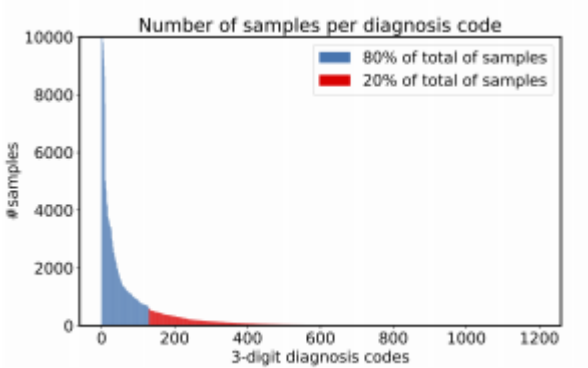
Data pre-processing

Data Source: MIMIC III (Medical Information Mart for Intensive Care)

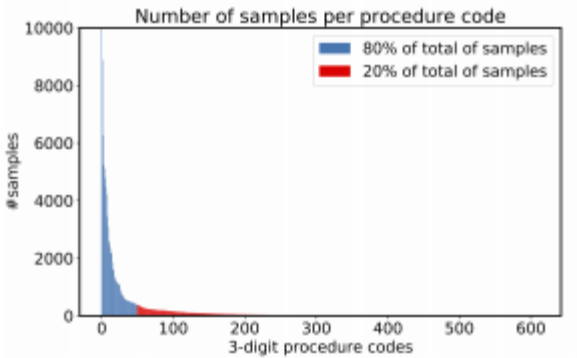
TEXT	Diagnosis	Procedures	Mortality	Length of Stay
CHIEF COMPLAINT: Knee pain.				
ALLERGIES: Patient recorded as having No Known Allergies to Drugs	038, 025	389, 399, 887	0	More than 14 days

Sample format of MIMIC III data after pre-processing. TEXT column includes the Admission notes for the patients. Diagnosis, Procedures, Mortality, and Length of Stay represent the labels for the four clinical outcome tasks.

Data distribution



Distribution of ICD-9 Diagnosis codes



Distribution of ICD-9 Procedures codes

Mortality		Length of Stay (in days)			
0	1	≤ 3	> 3 & ≤ 7	> 7 & ≤ 14	> 14
43,609	5,136	5,596	16,134	13,391	8,488

Distribution of Mortality and Length of Stay labels

Baselines

	Diagnosis (1266 classes)	*Procedures* (711 classes)	*In-Hospital Mortality* (2 classes)	*Length-of-Stay* (4 classes)
BERT Base	82.08	85.84	81.13	70.40
BioBERT Base	82.81	86.36	82.55	71.59
CORe	**83.54**	**87.65**	**84.04**	**72.53**

Baseline models, BioBERT and CORe on clinical outcome prediction tasks in macro-averaged AUROC. The CORe approach surpasses BioBERT Base model in all four tasks.

Evaluation – Test Dataset

	Experiment Setting	*Diagnosis* (1266 classes)	*Procedures* (711 classes)	*In-Hospital Mortality* (2 classes)	*Length-of-Stay* (4 classes)
Baselines	BioBERT Base	82.81	86.36	82.55	71.59
	CORe	83.54	87.65	**84.04**	72.53
MTL	Diagnosis - Procedures	82.45	**90.50**	-	
	Diagnosis - MP	79.70	-	77.53	-
	Diagnosis - LOS	76.55	-		57.07
	Procedures - MP	-	86.27	74.81	-
	Procedures - LOS	-	87.71	-	63.74
	Diagnosis - Procedures - MP	75.86	85.34	78.46	-
	Diagnosis - Procedures - LOS	77.52	87.12	-	57.51
	Diagnosis - Procedures - MP - LOS	69.54	77.79	75.64	62.98
Adapters	ST-A	**83.98**	87.21	83.15	**75.45**
	AdapterFusion	77.15	86.91	79.30	75.15

Performance of Baseline models (BioBERT and CORe), Traditional Multitask learning (**MTL**), Single-task Adapters **(ST-A)**, and AdapterFusion on clinical outcome prediction tasks in macro-averaged % AUROC on the test dataset.

Evaluation – Validation Dataset

	Experiment Setting	*Diagnosis* (1266 classes)	*Procedures* (711 classes)	*In-Hospital Mortality* (2 classes)	*Length-of-Stay* (4 classes)
MTL	Diagnosis - Procedures	82.13	**90.09**	-	
	Diagnosis - MP	80.27	-	76.35	-
	Diagnosis - LOS	75.90	-		56.52
	Procedures - MP	-	87.09	74.49	-
	Procedures - LOS	-	88.70	-	63.42
	Diagnosis - Procedures - MP	76.30	86.18	76.37	-
	Diagnosis - Procedures - LOS	76.90	87.71	-	57.19
	Diagnosis - Procedures - MP - LOS	70.80	77.55	75.71	61.67
Adapters	ST-A	84.30	87.22	82.30	**74.50**
	AdapterFusion	**85.26**	89.82	**82.95**	74.45

Performance of Traditional Multitask learning (**MTL**), Single-task Adapters **(ST-A)**, and AdapterFusion on clinical outcome prediction tasks in macro-averaged % AUROC on the validation dataset.

Quantitative Error Analysis

Diagnosis

- ST-As are more precise at predicting sample Diagnosis ICD-9 codes.
- AdapterFusion has zero prediction probability for 77% of sample Diagnosis ICD-9 codes.

Procedures

- Multi-task learning performs best across the sample Procedure ICD-9 codes

Mortality

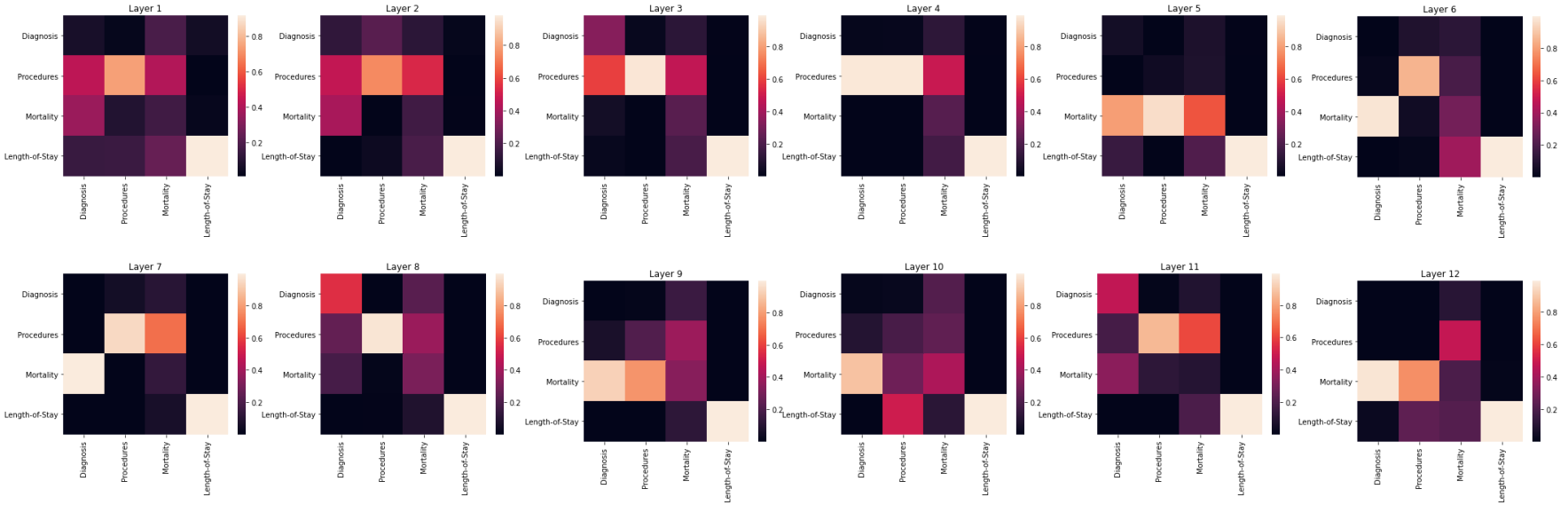
- ST-A has higher precision in correctly predicting the Mortality of a patient

Length of Stay

- ST-A and AdapterFusion perform similar in predicting Length of Stay across all class labels

Qualitative Error Analysis

Attention Plots for AdapterFusion



The rows represent the target task, and the columns represent the pre-trained Single-Task Adapters. Higher activation weight represents higher relevance for the corresponding task adapter.

Qualitative Error Analysis

Diagnosis ICD 9 codes

- AdapterFusion provides high relevance to "Hypertension" Diagnosis codes and has a match rate of 5% with MIMIC III labels
- On average, Multi-task Learning predictions (“Diagnosis-Procedures”) has a match rate of 33% with MIMIC III labels
- On average, ST-A predictions has a match rate of 38% with MIMIC III labels

Conclusion & Future Work

Summary

- ❑ Only Procedures task gain information using inter-contextual representations.
- ❑ AdapterFusion doesn't solve catastrophic interference problem in clinical domain.
- ❑ ST-As use the lowest amount of training time and resources, avoids overfitting, and surpasses baselines on Diagnosis and Length of stay tasks

Future Work

Additional
Data Sources

Adapter Drop

Hyperformer

Thank you!

Questions?

References

1. CORe approach https://aclanthology.org/2021.eacl-main.75.pdf
2. Multi-task Learning https://ruder.io/multi-task/
3. Adapter Architecture https://arxiv.org/pdf/1902.00751.pdf
4. AdapterFusion paper https://arxiv.org/pdf/2005.00247.pdf
5. Adapter Drop https://arxiv.org/pdf/2010.11918.pdf
6. Hyperformer https://arxiv.org/pdf/2106.04489.pdf

Appendix

Quantitative Error Analysis

		Precision			Recall			F1		
ICD-9 code	*# Examples*	*MTL*	*ST-A*	*Fusion*	*MTL*	*ST-A*	*Fusion*	*MTL*	*ST-A*	*Fusion*
008	307	0.42	0.82	0.00	0.03	0.05	0.00	0.06	0.09	0.00
0084	282	0.43	0.69	0.00	0.03	0.03	0.00	0.06	0.06	0.00
038	1,251	0.56	0.59	0.51	0.51	0.53	0.43	0.54	0.56	0.47
0381	167	0.26	0.28	0.00	0.08	0.04	0.00	0.13	0.07	0.00
0384	209	0.32	0.44	0.00	0.13	0.02	0.00	0.18	0.04	0.00
0389	742	0.41	0.48	0.44	0.26	0.18	0.12	0.32	0.26	0.19
041	850	0.31	0.45	0.00	0.08	0.02	0.00	0.13	0.03	0.00
0411	238	0.22	0.50	0.00	0.03	0.00	0.00	0.06	0.01	0.00

TABLE 5.4: Precision, Recall, and F1 scores for select Diagnosis ICD-9 codes. AdapterFusion has zero prediction probability for 77% of the analyzed codes. ST-As have higher precision score compared to the Multi-task learning (MTL) experiment. We use "Diagnosis-Procedures" task setting to represent the results for MTL in this table.

Quantitative Error Analysis

		Precision			Recall			F1		
ICD-9 code	*#Examples*	*MTL*	*ST-A*	*Fusion*	*MTL*	*ST-A*	*Fusion*	*MTL*	*ST-A*	*Fusion*
004	399	0.65	0.70	0.68	0.51	0.42	0.29	0.57	0.53	0.41
001	396	0.20	0.00	0.00	0.03	0.00	0.00	0.05	0.00	0.00
0040	305	0.62	0.61	0.59	0.46	0.37	0.18	0.53	0.46	0.28
006	291	0.68	0.71	0.72	0.63	0.55	0.33	0.66	0.62	0.45
0066	260	0.66	0.69	0.70	0.65	0.59	0.30	0.65	0.63	0.42
0045	178	0.55	0.50	1.00	0.31	0.19	0.01	0.40	0.27	0.01
0017	160	0.13	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00
0014	149	0.04	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
015	131	0.71	0.77	0.87	0.64	0.62	0.25	0.67	0.69	0.39
013	126	0.53	0.00	0.00	0.48	0.00	0.00	0.50	0.00	0.00

TABLE 5.5: Precision, Recall, and F1 scores for select Procedures ICD-9 codes.We use "Diagnosis-Procedures" task setting to represent the results for MTL in this table. MTL performs best across all ICD-9 codes with high F1 score compared to Single-task Adapters and AdapterFusion.

Quantitative Error Analysis

		Precision			Recall			F1		
Label	***# Examples***	***MTL***	***ST-A***	***Fusion***	***MTL***	***ST-A***	***Fusion***	***MTL***	***ST-A***	***Fusion***
0	8,797	0.92	0.92	0.95	0.96	0.97	0.82	0.94	0.94	0.88
1	1,033	0.46	0.54	0.28	0.28	0.26	0.61	0.35	0.35	0.38

TABLE 5.6: Precision, Recall, and F1 scores for Mortality task labels. "0" Label represents that the patient did not die during the hospitalization and "1" represents that the patient passed away during the hospitalization. On average, Single-task Adapters have higher precision in predicting if the patient would die during the hospitalization. We use "Diagnosis-Procedures-Mortality" task setting to represent the results for MTL in this table.

Quantitative Error Analysis

		Precision			Recall			F1		
Label	***# Examples***	***MTL***	***ST-A***	***Fusion***	***MTL***	***ST-A***	***Fusion***	***MTL***	***ST-A***	***Fusion***
0	1,121	0.36	0.47	0.46	0.42	0.34	0.35	0.39	0.40	0.40
1	3,328	0.43	0.47	0.48	0.51	0.50	0.53	0.47	0.49	0.50
2	2,692	0.36	0.38	0.37	0.41	0.36	0.39	0.38	0.37	0.38
3	1,656	0.33	0.36	0.39	0.31	0.41	0.36	0.32	0.38	0.37

TABLE 5.7: Precision, Recall, and F1 scores for select Mortality Prediction task labels. Label "0" : <3, "1" : 4-7, "2": 8-14, "3": >14 days at the hospital. Single-task Adapters and AdapterFusion have very similar scores across all labels and perform better than the Multi-task learning approach. We use "Procedures-Length of Stay" task setting to represent the results for MTL in this table.