Versions of Approaches

PtrGNCMsg

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	Generating Commit Messages from Diffs using Pointer-Generator Network	Original approach	N/A	Reference: BLEU - 40.79 ROUGE 1/2/L - 40.5/26.89/40.04
1	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Different Dataset for training Name Change	Trained in ATOM proposed Dataset Named Ptr-Net in the paper	Metric results for version: BLEU 1/2/3/4 - 5.80/1.72/0.73/0.45 ROUGE-L - 7.61 METEOR - 4.98
2.1	Context-aware Retrieval-based Deep Commit Message Generation Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Different Dataset for training	Trained in NNGen and CoRec top 1,000 dataset	Metric results for version: BLEU-4 - 12.31 Mod. ngram precision (1/2/3/4) - 27.8/15.3/11.8/10.5 ROUGE-L - 26.76 METEOR - 11.94 From second paper: BLEU 1/2/3/4 - 20.23/9.99/7.17/5.47
2.2	Context-aware Retrieval-based Deep Commit Message Generation	Different Dataset for training	Trained in CoRec top 10,000 dataset	Metric results for version: BLEU-4 - 24.67 Mod. ngram precision (1/2/3/4) - 42.3/27.6/24.4/23.5 ROUGE-L - 37.24 METEOR - 18.64
2.3	Context-aware Retrieval-based Deep Commit Message Generation	Different Dataset for training	Trained in CoRec top 10,000 dataset split by project	Metric results for version: BLEU-4 - 15.34

PtrGNCMsg

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
				Mod. ngram precision (1/2/3/4) - 31.3/16.1/12.8/11.9 ROUGE-L - 29.93 METEOR - 14.25
3	CoreGen: Contextualized Code Representation Learning for Commit Message Generation	Different Dataset for training	Trained in NNGen top 1,000 dataset cleansed by Liu et al. Also used in CoreGen approach. Slightly different dataset from version 2.1	Metric results for version: BLEU-4 - 9.78 ROUGE-1/2/L - 23.66/9.61/23.67 METEOR - 11.41
4	RACE: Retrieval-Augmented Commit Message Generation	Different dataset	Trained in MCMD dataset for fine tuning purposes to compare against RACE	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 17.71/15.98/14.06/15.89/20.78 ROUGE-L - 24.32/21.16/20.17/23.49/27.87 METEOR - 11.33/10.18/9.63/11.36/14.52 CIDEr - 0.99/0.83/0.63/0.76/1.29
5.1	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in original Jiang et al. dataset	Metric results for version: BLEU-MOSES - 35.41 BLEU-B-Norm - 29.86 BLEU-CC - 24.82 ROUGE-1 - 34.99 ROUGE-2 - 23.96 ROUGE-L - 34.77 METEOR - 31.79
5.2		Different dataset	Trained in NNGen dataset	Metric results for version: BLEU-MOSES - 9.69

PtrGNCMsg

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	A large-scale empirical study of commit message generation: models, datasets and evaluation			BLEU-B-Norm - 18.96 BLEU-CC - 10.39 ROUGE-1 - 23.47 ROUGE-2 - 9.41 ROUGE-L - 23.12 METEOR - 19.19
5.3	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in CoDiSum dataset	Metric results for version: BLEU-MOSES - 0.81 BLEU-B-Norm - 12.71 BLEU-CC - 4.77 ROUGE-1 - 16.43 ROUGE-2 - 3.58 ROUGE-L - 16.06 METEOR - 12.16
5.4	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset	Metric results for version(C++/C#/Java/Python/JS): BLEU-MOSES - 5.91/18.92/8.25/8.62/15.15 BLEU-B-Norm - 13.07/19.72/15.33/15.99/19.58 BLEU-CC - 9.97/17.18/11.70/11.87/16.50 ROUGE-1 - 17.53/22.32/19.09/21.39/25.11 ROUGE-2 - 7.69/14.30/8.32/9.91/14.45 ROUGE-L - 17.08/21.99/18.64/20.76/24.60 METEOR - 16.86/22.33/19.13/21.18/24.61
5.5		Different dataset	Trained in MCMD dataset split by project	

PtrGNCMsg

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	A large-scale empirical study of commit message generation: models, datasets and evaluation			Metric results for version (Average): BLEU-B-Norm - 8.61 ROUGE-1 - 12.03 ROUGE-2 - 4.36 ROUGE-L - 11.68 METEOR - 10.75
5.6	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset split by timestamp	Metric results for version (Average): BLEU-B-Norm - 12.98 ROUGE-1 - 16.66 ROUGE-2 - 6.95 ROUGE-L - 16.18 METEOR - 17.61
6.1	Revisiting Learning-based Commit Message Generation	Different dataset	Trained in CoDiSum dataset and evaluated in both training set and test set. Different results than previous trained in CoDiSum dataset	Metric results for version (Training set/Test set): BLEU-4 - 17.41/16.36 ROUGE-L - 21.17/19.82 METEOR - 14.09/13.01
6.2	Revisiting Learning-based Commit Message Generation	Input Change Different dataset	Trained in CoDiSum dataset feeding as input only the code change marks instead of whole commit entry Trained in CoDiSum dataset with different results	Metric results for version: BLEU-4 - 13.17 ROUGE-L - 15.45 METEOR - 9.88
6.3	Revisiting Learning-based Commit Message Generation	Ablation of attention mechanism Different dataset	Ablation study performed deactivating the attention mechanism of the architecture Trained in CoDiSum dataset with different results	Metric results for version: BLEU-4 - 15.81 ROUGE-L - 18.81 METEOR - 12.13
6.4	Revisiting Learning-based Commit Message Generation	Ablation of copy mechanism	Ablation study performed deactivating the copy mechanism for OOV out of the architecture	Metric results for version: BLEU-4 - 15.97

PtrGNCMsg

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
		Different dataset	Trained in CoDiSum dataset with different results	ROUGE-L - 18.97 METEOR - 12.31
7	Revisiting Learning-based Commit Message Generation	Different dataset	Trained in original NMT dataset and with different results than previous trained in same dataset	Metric results for version: BLEU-4 - 12.31 METEOR - 11.94 ROUGE-L - 24.45 CIDEr - 1.10 SPICE - 17.38

ATOM

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Orgiginal approach	N/A	Metric results for version: BLEU 1/2/3/4 - 23.88/15.61/12.17/ 10.51 ROUGE-L - 22.02 METEOR - 18.51
0.1	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Ablation of Generation Module	Ablation deactivating Generation Module, only Retrieval working	Metric results for version: BLEU 1/2/3/4 - 17.74/12.65/10.55/8.52 ROUGE-L - 15.93 METEOR - 14.35
0.2	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Ablation of Retrieval Module	Ablation deactivating Retrieval Module, only Generation working	Metric results for version: BLEU 1/2/3/4 - 15.97/10.70/8.83/7.35 ROUGE-L - 14.80 METEOR - 11.82
1.1	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Different dataset Trained in random partition	Trained on CoMeG dataset Trained in partition of the presented dataset using random criteria	Metric results for version: BLEU-4 - 10.56 ROUGE-L - 14.67 METEOR - 7.14
1.2	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Different dataset Trained in timestamp partition	Trained on CoMeG dataset Trained in partition of the presented dataset using timestamp criteria	Metric results for version: BLEU-4 - 10.04 ROUGE-L - 13.66 METEOR - 6.37
1.3		Different dataset	Trained on CoMeG dataset	Metric results for version: BLEU-4 - 9.43

ATOM

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Trained in project partition	Trained in partition of the presented dataset using project criteria	ROUGE-L - 12.62 METEOR - 5.89
2	FIRA: Fine-Grained Graph-Based Code Change Representation for Automated Commit Message Generation	Different dataset	Trained on CoDiSum dataset used in FIRA experiment	Metric results for version: B-Norm BLEU - 8.35 Penalty BLEU - 7.42 ROUGE-L - 10.17 METEOR - 8.73
3	RACE: Retrieval-Augmented Commit Message Generation	Different dataset	Trained in MCMD dataset for fine tuning purposes to compare against RACE. As per ATOM configuration, only Java subset used	Metric results for version: BLEU-4 - 16.42 ROUGE-L - 22.67 METEOR - 11.66 CIDEr - 0.91
4	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset Java subsection for empirical study	Metric results for version: BLEU-MOSES - 7.47 BLEU-B-Norm - 16.42 BLEU-CC - 9.29 ROUGE-1 - 19.10 ROUGE-2 - 9.58 ROUGE-L - 18.60 METEOR - 20.80

NMT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	Automatically Generating Commit Messages from Diffs using NMT	Original approach	N/A	Metric results for version: BLEU-4 - 31.92 M. Ngram Prec. 1/2/3/4 - 38.1/31.1/29.5/29.7
0.1	Automatically Generating Commit Messages from Diffs using NMT	Different training set	Trained in different dataset, VDO filter from original dataset removed and v0.1 retrained in that dataset	Metric results for version: BLEU-4 on NMT0 dataset - 40.1 BLEU-4 on NMT0.1 dataset (its own) - 30.2 M. Ngram Prec. 1/2/3/4 - 40.1/34.0/33.4/34.3 M. Ngram Prec on new test set - 30.2/23.3/20.7/19.6
		Different test set	New test set used to obtain results. Tested in both NMT v0 and NMT v0.1 test sets	
		VDO filter deactivated	Ablation study performed deactivating the VDO filter	
1.1	Generating Commit Messages from Diffs using Pointer-Generator Network	Different Dataset for training	Trained on top 1000 dataset without lowercasing values (see top1000 on PtrGNCMsg)	BLEU - 33.05 ROUGE 1/2/L - 35.54/25.75/35.38
1.2	Generating Commit Messages from Diffs using Pointer-Generator Network	Different Dataset for training	Trained on top 1000 lowercased dataset values (see on PtrGNCMsg)	BLEU - 37 ROUGE 1/2/L - 36.35/26.05/36.16
1.3	Generating Commit Messages from Diffs using Pointer-Generator Network	Different Dataset for training	Trained on top 2000 dataset without lowercasing values (see on PtrGNCMsg)	BLEU - 37.32 ROUGE 1/2/L - 35.9/24.55/35.71
1.4	Generating Commit Messages from Diffs using Pointer-Generator Network	Different Dataset for training	Trained on top 2000 lowercased dataset values (see on PtrGNCMsg)	BLEU - 37.78 ROUGE 1/2/L - 35.73/24.47/35.48
2.1	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Different Dataset for training	Trained in ATOM proposed Dataset	Metric results for version: BLEU 1/2/3/4 - 13.12/8.01/6.11/5.23 ROUGE-L - 12.73 METEOR - 10.37
		Luong Attention Mechanism	Changed Bahdanau attention mechanism for Luong	

NMT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
2.2	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Different Dataset for training	Mantained Bahdanau attention mechanism but was trained on dataset proposed in ATOM paper	Metric results for version: BLEU 1/2/3/4 - 12.78/7.66/5.72/4.81 ROUGE-L - 11.95 METEOR - 9.87
2.3	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Different Dataset for training Luong Attention Mechanism	Trained in ATOM dataset DIVIDED BY PROJECT Changed Bahdanau attention mechanism for Luong	Metric results for version: BLEU 1/2/3/4 - 4.48/0.98/0.00/0.00 ROUGE-L - 0.04 METEOR - 0.03
2.4	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Different Dataset for training Luong Attention Mechanism	Trained in ATOM dataset DIVIDED BY TIMESTAMP Changed Bahdanau attention mechanism for Luong	Metric results for version: BLEU 1/2/3/4 - 6.69/2.70/1.25/1.47 ROUGE-L - 9.45 METEOR - 6.11
3	Commit Message Generation for Source Code Changes	Different Dataset for training	Trained on dataset proposed for CoDiSum	Metric results for version: BLEU-4 - 0.87 METEOR - 4.81 Recall (%) - 2.98
4.1	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Different dataset Trained in random partition	Trained on CoMeG dataset Trained in partition of the presented dataset using random criteria	Metric results for version: BLEU-4 - 12.82 ROUGE-L - 18.30 METEOR - 9.30
4.2	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Different dataset Trained in timestamp partition	Trained on CoMeG dataset Trained in partition of the presented dataset using timestamp criteria	Metric results for version: BLEU-4 - 11.74 ROUGE-L - 16.85 METEOR - 8.08

NMT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
4.3	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Different dataset Trained in project partition	Trained on CoMeG dataset Trained in partition of the presented dataset using project criteria	Metric results for version: BLEU-4 - 9.20 ROUGE-L - 12.37 METEOR - 5.83
5	Neural-Machine-Translation-Based Commit Message Generation: How Far Are We? CoreGen: Contextualized Code Representation Learning for Commit Message Generation Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Different dataset	Trained on cleansed dataset proposed by NNGen authors	Metric results for version: BLEU-4 - 14.19 Mod. ngram precision (1/2/3/4) - 24.8/14.6/11.4/9.9 From second paper: ROUGE-1/2/L - 20.37/10.44/19.20 METEOR - 9.57 From third paper: BLEU 1/2/3/4 - 14.67/7.65/ 5.42/4.33
6.1	CoreGen: Contextualized Code Representation Learning for Commit Message Generation	Different dataset	Trained on NNGen dataset without the duplicates eliminated by authors, around 15% of content of validation and test sets	Metric results for version: BLEU-4 - 10.54 ROUGE-1/2/L - 20.37/10.44/19.20 METEOR - 9.57
6.2	CoreGen: Contextualized Code Representation Learning for Commit Message Generation	Different dataset CoreGen code contextualization mechanism	Trained on NNGen dataset Mechanism to learn code contextualization from stage I of CoreGen added to the base NMT approach	Metric results for version: BLEU-4 - 17.96 ROUGE-1/2/L - 24.99/14.07/23.70 METEOR - 14.28

NMT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
7.1	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in original Jiang et al. dataset	Metric results for version: BLEU-MOSES - 32.09 BLEU-B-Norm - 26.66 BLEU-CC - 21.51 ROUGE-1 - 28.46 ROUGE-2 - 20.22 ROUGE-L - 28.30 METEOR - 27.87
7.2	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in NNGen dataset	Metric results for version: BLEU-MOSES - 7.46 BLEU-B-Norm - 13.82 BLEU-CC - 8.28 ROUGE-1 - 15.92 ROUGE-2 - 7.04 ROUGE-L - 15.74 METEOR - 15.43
7.3	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in CoDiSum dataset	Metric results for version: BLEU-MOSES - 1.32 BLEU-B-Norm - 9.93 BLEU-CC - 3.81 ROUGE-1 - 11.24 ROUGE-2 - 2.26 ROUGE-L - 11.11 METEOR - 9.61
7.4		Different dataset	Trained in MCMD dataset	

NMT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	A large-scale empirical study of commit message generation: models, datasets and evaluation			Metric results for version(C++/C#/Java/Python/JS): BLEU-MOSES - 8.37/20.18/9.17/7.64/15.00 BLEU-B-Norm - 11.56/17.32/13.39/11.53/17.08 BLEU-CC - 9.63/16.35/10.24/9.56/14.77 ROUGE-1 - 14.11/20.11/15.47/14.56/20.78 ROUGE-2 - 8.17/14.74/7.78/8.22/13.07 ROUGE-L - 14.04/20.02/15.33/14.41/20.54 METEOR - 14.75/19.80/16.00/16.4/21.13
7.5	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset split by project	Metric results for version (Average): BLEU-B-Norm - 6.46 ROUGE-1 - 9.40 ROUGE-2 - 6.12 ROUGE-L - 9.36 METEOR - 8.98
7.6	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset split by timestamp	Metric results for version (Average): BLEU-B-Norm - 8.41 ROUGE-1 - 10.66 ROUGE-2 - 5.17 ROUGE-L - 10.60 METEOR - 11.88

NMT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
7.1	Revisiting Learning-based Commit Message Generation	Different dataset	Trained in CoDiSum dataset and evaluated in both training set and test set. Different results than previous trained in CoDiSum dataset	Metric results for version (Training set/Test set): BLEU-4 - 16.94/14.63 ROUGE-L - 20.41/17.31 METEOR - 13.34/10.91
7.2	Revisiting Learning-based Commit Message Generation	Input Change	Trained in CoDiSum dataset feeding as input only the code change marks instead of whole commit entry	Metric results for version: BLEU-4 - 12.80 ROUGE-L - 15.49 METEOR - 8.62
		Different dataset	Trained in CoDiSum dataset with different results	
7.3	Revisiting Learning-based Commit Message Generation	Ablation of attention mechanism	Ablation study performed deactivating the attention mechanism of the architecture	Metric results for version: BLEU-4 - 13.66 ROUGE-L - 16.24 METEOR - 10.09
		Different dataset	Trained in CoDiSum dataset with different results	
8.1	Revisiting Learning-based Commit Message Generation	Different dataset	Trained in Java dataset proposed by authors of NMT reproduction paper	Metric results for version (Training set/Test set): BLEU-4 - 5.33 ROUGE-1 - 23.60 ROUGE-2 - 10.87 ROUGE-L - 26.52 ROUGE-W - 19.35
		Hyperparameter change	SGD optimizer with initial learning rate = 0.1 / Learning rate - 0.1 if no improvement over 10 epochs / early stopping if no validation improvement in 20 epochs	
8.2		Different dataset		

NMT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	Revisiting Learning-based Commit Message Generation	Hyperparameter change	Trained in C# dataset proposed by authors of NMT reproduction paper SGD optimizer with initial learning rate = 0.1 / Learning rate - 0.1 if no improvement over 10 epochs / early stopping if no validation improvement in 20 epochs	Metric results for version (Training set/Test set): BLEU-4 - 7.31 ROUGE-1 - 26.84 ROUGE-2 - 13.16 ROUGE-L - 29.85 ROUGE-W - 22.08
8.3	Revisiting Learning-based Commit Message Generation	Different dataset Hyperparameter change	Trained in original dataset proposed by authors of NMT but with different results in metrics SGD optimizer with initial learning rate = 0.1 / Learning rate - 0.1 if no improvement over 10 epochs / early stopping if no validation improvement in 20 epochs	Metric results for version (Training set/Test set): BLEU-4 - 33.63 ROUGE-1 - 37.20 ROUGE-2 - 23.22 ROUGE-L - 40.01 ROUGE-W - 30.10
8.4	Revisiting Learning-based Commit Message Generation	Different dataset Hyperparameter change	Trained in a processed version of the dataset use in original approach, curated by the authors from origin of dataset from proposal paper SGD optimizer with initial learning rate = 0.1 / Learning rate - 0.1 if no improvement over 10 epochs / early stopping if no validation improvement in 20 epochs	Metric results for version (Training set/Test set): BLEU-4 - 3.19 ROUGE-1 - 20.26 ROUGE-2 - 7.93 ROUGE-L - 23.05 ROUGE-W - 16.37
9.1		Different dataset		

NMT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	Boosting Neural Commit Message Generation with Code Semantic Analysis	Hyperparameter change	Trained in proposed dataset of 18 Java projects preprocessed by ChangeScribe Batch size from original approach changed to 40	Metric results for version: BLEU-4 - 1.10 Modified N-Gram (1/2/3/4) - 4.7/1.7/0.5/0.4
9.2	Boosting Neural Commit Message Generation with Code Semantic Analysis	Different dataset Hyperparameter change	Trained in proposed dataset of 18 Java projects preprocessed by ChangeScribe Batch size from original approach changed to 15	Metric results for version: BLEU-4 - 0.44 Modified N-Gram (1/2/3/4) - 9.1/3.2/0.1/0.0
9.3	Boosting Neural Commit Message Generation with Code Semantic Analysis	Different dataset Hyperparameter change	Trained in proposed dataset of 18 Java projects with raw commit information Batch size from original approach changed to 40	Metric results for version: BLEU-4 - 0.41 Modified N-Gram (1/2/3/4) - 3.9/0.9/0.1/0.1
10	Correlating Automated and Human Evaluation of Code Documentation Generation Quality	Different dataset Resullts evaluation	Trained in original dataset, results differ from previous versions over same dataset Not re-trained, reported results on original code repository scored over metrics	Metric results for version: BLEU-4 - 14.19 METEOR - 12.99 ROUGE-L - 23.66 CIDEr - 1.06 SPICE - 18.07
11	Evaluating Commit Message Generation: To BLEU Or Not To BLEU?	Different and new metric	Performance measured over MCMD dataset (already used) but evaluated in newly proposed metric	Metric results for version (C++/C#/Java/Python/JS): Log-MNEXT - 11.69/17.96/10.7/13.58/9.34

CoDiSum

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	Commit Message Generation for Source Code Changes	Original approach	N/A	Metric results for version: BLEU-4 - 2.19 METEOR - 7.46 Recall (%) - 33.96
0.1	Commit Message Generation for Source Code Changes	Removes copying mechanism Structure and semantics together	Copying mechanism deleted from original approach Jointly models code semantics and structure instead of separatedly as original	Metric results for version: BLEU-4 - 2.06 METEOR - 7.04 Recall (%) - 30.21
0.2	Commit Message Generation for Source Code Changes	Delete code semantics Replace identifiers with placeholders	Not capturing code semantics with bi-GRU layer, only captures structure Not same mechanism used by original approach with CopyNet	Metric results for version: BLEU-4 - 1.97 METEOR - 6.86 Recall (%) - 29.63
0.3	Commit Message Generation for Source Code Changes	Concatenate semantics and structure	Directly concatenates structure and semantics instead of treating them separatedly	Metric results for version: BLEU-4 - 2.05 METEOR - 7.14 Recall (%) - 33.87
1	ATOM: Commit Message Generation Based on Abstract Syntax Tree and Hybrid Ranking	Trained in different dataset	Trained in ATOM dataset	Metric results for version: BLEU 1/2/3/4 - 7.82/3.61/2.22/1.75 ROUGE-L - 9.87 METEOR - 8.35
2.1	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Different dataset Trained in random partition	Trained on CoMeG dataset Trained in partition of the presented dataset using random criteria	Metric results for version: BLEU-4 - 12.50 ROUGE-L - 17.72 METEOR - 9.09

CoDiSum

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
2.2	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Different dataset Trained in timestamp partition	Trained on CoMeG dataset Trained in partition of the presented dataset using timestamp criteria	Metric results for version: BLEU-4 - 11.46 ROUGE-L - 16.20 METEOR - 7.83
2.3	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Different dataset Trained in project partition	Trained on CoMeG dataset Trained in partition of the presented dataset using project criteria	Metric results for version: BLEU-4 - 10.23 ROUGE-L - 13.83 METEOR - 6.72
3	FIRA: Fine-Grained Graph-Based Code Change Representation for Automated Commit Message Generation	Potential reimplementation	Reimplementation from FIRA authors according to code as code is said not to be available	Metric results for version: B-Norm BLEU - 16.55 Penalty BLEU - 12.07 ROUGE-L - 19.73 METEOR - 12.83
4	RACE: Retrieval-Augmented Commit Message Generation	Different dataset	Trained in MCMD dataset for fine tuning purposes to compare against RACE	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 13.97/12.71/12.44/14.61/11.22 ROUGE-L - 16.12/14.40/14.39/17.02/13.26 METEOR - 6.02/5.56/6.00/8.59/5.32 CIDEr - 0.39/0.36/0.42/0.42/0.28
5	Mucha: Multi-channel based Code Change Representation Learning for Commit Message Generation	Different dataset	Trained in variation of Tufano et al. used for Mucha training and measurement	Metric results for version: BLEU-4 - 3.56 ROUGE-L - 5.64 METEOR - 1.30
6.1	A large-scale empirical study of commit message	Different dataset	Trained in original Jiang et al. dataset	Metric results for version: BLEU-MOSES - 0.00

CoDiSum

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	generation: models, datasets and evaluation			BLEU-B-Norm - 6.88 BLEU-CC - 0.49 ROUGE-1 - 11.64 ROUGE-2 - 0.03 ROUGE-L - 11.64 METEOR - 4.57
6.2	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in NNGen dataset	Metric results for version: BLEU-MOSES - 0.00 BLEU-B-Norm - 8.03 BLEU-CC - 0.86 ROUGE-1 - 11.32 ROUGE-2 - 0.45 ROUGE-L - 11.32 METEOR - 4.82
6.3	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in CoDiSum dataset measured with different metrics	Metric results for version: BLEU-MOSES - 1.74 BLEU-B-Norm - 15.45 BLEU-CC - 5.72 ROUGE-1 - 19.00 ROUGE-2 - 4.66 ROUGE-L - 18.62 METEOR - 12.30
6.4	A large-scale empirical study of commit message	Different dataset	Trained in MCMD dataset	Metric results for version(C++/C#/Java/Python/JS):

CoDiSum

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	generation: models, datasets and evaluation			BLEU-MOSES - 3.99/1.78/2.00/2.63/1.51 BLEU-B-Norm - 12.46/12.73/14.00/14.63/11.24 BLEU-CC - 5.71/4.74/5.37/5.74/4.31 ROUGE-1 - 15.17/15.07/16.84/19.45/13.89 ROUGE-2 - 4.54/3.50/4.43/5.94/2.94 ROUGE-L - 14.94/14.78/16.51/18.91/13.53 METEOR - 10.20/10.01/11.34/12.08/9.51
6.5	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset split by project	Metric results for version (Average): BLEU-B-Norm - 7.76 ROUGE-1 - 9.81 ROUGE-2 - 2.64 ROUGE-L - 9.63 METEOR - 6.45
6.6	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset split by timestamp	Metric results for version (Average): BLEU-B-Norm - 11.17 ROUGE-1 - 12.99 ROUGE-2 - 3.13 ROUGE-L - 12.72 METEOR - 8.33
6.1		Different dataset		Metric results for version (Training set/Test set):

CoDiSum

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	Revisiting Learning-based Commit Message Generation		Trained in CoDiSum dataset and evaluated in both training set and test set. Different results than previous trained in CoDiSum dataset	BLEU-4 - 18.03/16.57 ROUGE-L - 21.81/ 19.84 METEOR - 14.69/ 13.08
6.2	Revisiting Learning-based Commit Message Generation	Input Change	Trained in CoDiSum dataset feeding as input only the code change marks instead of whole commit entry	Metric results for version: BLEU-4 - 14.24 ROUGE-L - 16.46 METEOR - 10.46
		Different dataset	Trained in CoDiSum dataset with different results	
6.3	Revisiting Learning-based Commit Message Generation	Ablation of attention mechanism	Ablation study performed deactivating the attention mechanism of the architecture	Metric results for version: BLEU-4 - 15.93 ROUGE-L - 18.88 METEOR - 12.30
		Different dataset	Trained in CoDiSum dataset with different results	
6.4	Revisiting Learning-based Commit Message Generation	Ablation of copy mechanism	Ablation study performed deactivating the copy mechanism for OOV out of the architecture	Metric results for version: BLEU-4 - 16.29 ROUGE-L - 19.42 METEOR - 12.79
		Different dataset	Trained in CoDiSum dataset with different results	
6.5	Revisiting Learning-based Commit Message Generation	Ablation of anonymization mechanism	Ablation study performed deactivating the anonymization mechanism out of the architecture	Metric results for version: BLEU-4 - 16.03 ROUGE-L - 19.38 METEOR - 12.20
		Different dataset	Trained in CoDiSum dataset with different results	

CommitBERT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	CommitBERT: Commit Message Generation Using Pre-Trained Programming Language Model	Original approach	N/A	Metric results for version: BLEU-4 (Python PHP JavaScript Java Go Ruby) 12.93/14.30/11.49/9.81/12.76/10.56 Dev PPL (Python PHP JavaScript Java Go Ruby) 49.29/47.89/75.53/77.80/64.43/82.82
0.1	CommitBERT: Commit Message Generation Using Pre-Trained Programming Language Model	Random weight initialization	Initialization of the proposal on random weights instead of pretrained	Metric results for version: BLEU-4 (Python PHP JavaScript Java Go Ruby) 7.95/7.01/8.41/7.60/10.38/7.17 Dev PPL (Python PHP JavaScript Java Go Ruby) 144.60/138.39/195.98/275.84/257.29/207.68
0.2	CommitBERT: Commit Message Generation Using Pre-Trained Programming Language Model	RoBERTa initialization	Initialization on RoBERTa's weights	Metric results for version: BLEU-4 (Python PHP JavaScript Java Go Ruby) 10.94/9.71/9.50/6.40/10.21/8.95 Dev PPL (Python PHP JavaScript Java Go Ruby) 76.02/81.97/103.48/164.32/122.70/104.68
0.3	CommitBERT: Commit Message Generation Using Pre-Trained Programming Language Model	CodeBERT initialization	Initialization on CodeBERT's weights	Metric results for version: BLEU-4 (Python PHP JavaScript Java Go Ruby) 12.05/13.06/10.47/8.91/11.19/10.33 Dev PPL (Python PHP JavaScript Java Go Ruby) 68.18/63.90/94.62/116.50/109.43/91.50
1.1	COME: Commit Message Generation with Modification Embedding RACE: Retrieval-Augmented Commit Message Generation	Potential reimplementation Different dataset	Not specified in paper which approaches are reimplemented or which took the source code Trained in MCMD Dataset	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 22.32/20.67/16.16/17.29/23.40 ROUGE-L - 28.03/25.76/19.90/22.36/30.51 METEOR - 12.63/12.31/10.05/11.31/15.64 CIDEr - 1.42/1.25/0.94/1.01/1.54

CommitBERT

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
1.2	COME: Commit Message Generation with Modification Embedding	Potential reimplementation	Not specified in paper which approaches are reimplemented or which took the source code	Metric results for version : BLEU-4 - 17.44 ROUGE-L - 21.64 METEOR - 9.10 CIDEr - 0.64
		Different dataset	Trained in CoDiSum Dataset	
2	RACE: Retrieval-Augmented Commit Message Generation	Reimplementation	Reimplemented approach making use of code diff encoder trained from RACE approach to obtain high dimensional semantic vectors	Metric results for version (C++/C#/Java/Python/JS): BLEU - 27,69/25,29/19,18/20,83/29,29
		Different dataset	Trained in MCMD Dataset	

COME

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	COME: Commit Message Generation with Modification Embedding	Original approach	N/A	Metric results for CoDiSum: BLEU-4 - 19.64 ROUGE-L - 24.56 METEOR - 10.70 CIDEr - 0.82 Metric results for MCMD (C++/C#/Java/Python/JS): BLEU-4 - 27.17/ 27.29/20.80/23.17/26.91 ROUGE-L - 34.59/33.33/27.01/30.48/34.44 METEOR - 16.91/17.77/14.55/16.46/17.84 CIDEr - 1.90/1.91/1.25/1.50/1.92
0.1	COME: Commit Message Generation with Modification Embedding	Removal of Code contextualization	Code contextualization task is removed from embedding module to perform ablation study	Metric results for CoDiSum: BLEU-4 - 17.70 ROUGE-L - 22.01 METEOR - 9.30 CIDEr - 0.66 Metric results for MCMD (C++/C#/Java/Python/JS): BLEU-4 - 22.76/22.21/16.73/17.99/22.87/ ROUGE-L - 30.23/29.08/22.86/25.27/29.81 METEOR - 14.57/14.51/11.69/12.74/15.12 CIDEr - 1.43/1.33/0.85/0.96/1.50
		Removal of modification embedding	Modification embeddings are not performed in embedding module to perform ablation study	
0.2	COME: Commit Message Generation with Modification Embedding	Removal of Code contextualization	Code contextualization task is removed from embedding module to perform ablation study	Metric results for CoDiSum: BLEU-4 - 18.40 ROUGE-L - 22.96 METEOR - 9.88 CIDEr - 0.71 Metric results for MCMD

COME

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
				(C++/C#/Java/Python/JS): BLEU-4 - 24.02/23.48/18.77/20.13/24.22 ROUGE-L - 31.07/29.92/24.16/27.09/31.47 METEOR - 14.66/15.11/13.27/14.47/16.14 CIDEr - 1.59/1.48/1.06/1.16/1.69

CoMeG

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0.0.1	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Original approach	N/A	Metric results for version: BLEU-4 - 14.60 ROUGE-L - 21.48 METEOR - 11.19
		Trained in random partition	Trained in partition of the presented dataset using random criteria	
0.0.2	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Original approach	N/A	Metric results for version: BLEU-4 - 13.47 ROUGE-L - 19.98 METEOR - 10.01
		Trained in timestamp partition	Trained in partition of the presented dataset using timestamp criteria	
0.0.3	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Original approach	N/A	Metric results for version: BLEU-4 - 10.69 ROUGE-L - 14.49 METEOR - 6.98
		Trained in project partition	Trained in partition of the presented dataset using project criteria	
0.1.1	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Removed diff modelling	For ablation study purposes, diff tokenization and modelling is removed	Metric results for version: BLEU-4 - 12.43 ROUGE-L - 17.66 METEOR - 8.92
		Trained in random partition	Trained in partition of the presented dataset using random criteria	
0.1.2	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Removed diff modelling	For ablation study purposes, diff tokenization and modelling is removed	Metric results for version: BLEU-4 - 11.25 ROUGE-L - 15.89 METEOR - 7.70
		Trained in timestamp partition	Trained in partition of the presented dataset using timestamp criteria	
0.1.3	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Removed diff modelling	For ablation study purposes, diff tokenization and modelling is removed	Metric results for version: BLEU-4 - 10.20 ROUGE-L - 13.85 METEOR - 6.64

CoMeG

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
		Trained in project partition	Trained in partition of the presented dataset using project criteria	
0.2.1	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Removed context ASTs and Pointer-Networks	For ablation study purposes, the processing of the before and after ASTs and the pointer networks that interpreted the output are removed	Metric results for version: BLEU-4 - 12.96 ROUGE-L - 18.87 METEOR - 9.68
		Trained in random partition	Trained in partition of the presented dataset using random criteria	
0.2.2	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Removed context ASTs and Pointer-Networks	For ablation study purposes, the processing of the before and after ASTs and the pointer networks that interpreted the output are removed	Metric results for version: BLEU-4 - 12.12 ROUGE-L - 17.65 METEOR - 8.75
		Trained in timestamp partition	Trained in partition of the presented dataset using timestamp criteria	
0.2.3	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Removed context ASTs and Pointer-Networks	For ablation study purposes, the processing of the before and after ASTs and the pointer networks that interpreted the output are removed	Metric results for version: BLEU-4 - 9.82 ROUGE-L - 13.10 METEOR - 6.27
		Trained in project partition	Trained in partition of the presented dataset using project criteria	
0.3.1	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Remove path embedding	For ablation study purposes, path embedding at input is removed, detecting less code regularities	Metric results for version: BLEU-4 - 14.66 ROUGE-L - 21.59 METEOR - 11.21
		Trained in random partition	Trained in partition of the presented dataset using random criteria	
0.3.2	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Remove path embedding	For ablation study purposes, path embedding at input is removed, detecting less code regularities	Metric results for version: BLEU-4 - 13.36 ROUGE-L - 19.85 METEOR - 9.93
		Trained in timestamp partition	Trained in partition of the presented dataset using timestamp criteria	
0.3.3	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Remove path embedding	For ablation study purposes, path embedding at input is removed, detecting less code regularities	Metric results for version: BLEU-4 - 10.62

CoMeG

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
		Trained in project partition	Trained in partition of the presented dataset using project criteria	ROUGE-L - 14.32 METEOR - 6.88
0.4.1	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Remove AST diff encoder Trained in random partition	For ablation study purposes, the AST diff encoder and its function is removed Trained in partition of the presented dataset using random criteria	Metric results for version: BLEU-4 - 13.71 ROUGE-L - 20.01 METEOR - 10.42
0.4.2	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Remove AST diff encoder Trained in timestamp partition	For ablation study purposes, the AST diff encoder and its function is removed Trained in partition of the presented dataset using timestamp criteria	Metric results for version: BLEU-4 - 12.81 ROUGE-L - 19.01 METEOR - 9.47
0.4.3	Combining Code Context and Fine-grained Code Difference for Commit Message Generation	Remove AST diff encoder Trained in project partition	For ablation study purposes, the AST diff encoder and its function is removed Trained in partition of the presented dataset using project criteria	Metric results for version: BLEU-4 - 9.71 ROUGE-L - 12.91 METEOR - 6.20

CoRec

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0.0.1	Context-aware Retrieval-based Deep Commit Message Generation	Original approach	N/A	Metric results for version: BLEU-4 - 19.80 Mod. ngram precision (1/2/3/4) - 33.1/22.2/19.0/17.9 ROUGE-L - 31.13 METEOR - 15.69
		Top 1,000 dataset	Version of the approach trained in the top 1,000 dataset with 100k epochs in training	
0.0.2	Context-aware Retrieval-based Deep Commit Message Generation	Original approach	N/A	Metric results for version: BLEU-4 - 41.26 Mod. ngram precision (1/2/3/4) - 55.1/44.9/44.3/45.4 ROUGE-L - 47.20 METEOR - 26.87
		Top 10,000 dataset	Version of the approach trained in the top 10,000 dataset with 400k epochs in training	
0.1.1	Context-aware Retrieval-based Deep Commit Message Generation	Only output from retrieval module	Output only from retrieval module without passing through the context-aware mechanism	Metric results for version: BLEU-4 - 18.01 Mod. ngram precision (1/2/3/4) - 29.1/19.0/15.6/13.9 ROUGE-L - 29.14 METEOR - 14.62
		Top 1,000 dataset	Version of the approach trained in the top 1,000 dataset with 100k epochs in training	
0.1.2	Context-aware Retrieval-based Deep Commit Message Generation	Only output from retrieval module	Output only from retrieval module without passing through the context-aware mechanism	Metric results for version: BLEU-4 - 36.24 Mod. ngram precision (1/2/3/4) - 46.1/36.2/33.6/32.3
		Top 10,000 dataset		

CoRec

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
			Version of the approach trained in the top 10,000 dataset with 400k epochs in training	ROUGE-L - 42.84 METEOR - 24.11
0.2.1	Context-aware Retrieval-based Deep Commit Message Generation	Basic model	Basic version for ablation study without sample decay mechanism and retrieval module	Metric results for version: BLEU-4 - 18.01 Mod. ngram precision (1/2/3/4) - 29.2/18.9/16.1/15.1 ROUGE-L - 28.50 METEOR - 14.15
		Top 1,000 dataset	Version of the approach trained in the top 1,000 dataset with 100k epochs in training	
0.2.2	Context-aware Retrieval-based Deep Commit Message Generation	Basic model	Basic version for ablation study without sample decay mechanism and retrieval module	Metric results for version: BLEU-4 - 37.27 Mod. ngram precision (1/2/3/4) - 52.0/41.7/41.3/43.0 ROUGE-L - 43.14 METEOR - 24.18
		Top 10,000 dataset	Version of the approach trained in the top 10,000 dataset with 400k epochs in training	
0.3.1	Context-aware Retrieval-based Deep Commit Message Generation	Removed retrieval module	Approach for ablation study without the retrieval module action on the proposal	Metric results for version: BLEU-4 - 19.08 Mod. ngram precision (1/2/3/4) - 33.8/22.9/20.2/19.9 ROUGE-L - 29.75 METEOR - 14.87
		Top 1,000 dataset	Version of the approach trained in the top 1,000 dataset with 100k epochs in training	
0.3.2	Context-aware Retrieval-based Deep Commit Message Generation	Removed retrieval module	Approach for ablation study without the retrieval module action on the proposal	Metric results for version: BLEU-4 - 40.25 Mod. ngram precision (1/2/3/4) -

CoRec

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
		Top 10,000 dataset	Version of the approach trained in the top 10,000 dataset with 400k epochs in training	54.8/44.8/44.6/46.2 ROUGE-L - 45.81 METEOR - 26.09
0.4.1	Context-aware Retrieval-based Deep Commit Message Generation	Removed sample decay mechanism	Proposal for the ablation study without the sample decay mechanism in the proposal	Metric results for version: BLEU-4 - 19.23 Mod. ngram precision (1/2/3/4) - 29.9/19.5/16.4/15.0 ROUGE-L - 30.22 METEOR - 15.19
		Top 1,000 dataset	Version of the approach trained in the top 1,000 dataset with 100k epochs in training	
0.4.2	Context-aware Retrieval-based Deep Commit Message Generation	Removed sample decay mechanism	Proposal for the ablation study without the sample decay mechanism in the proposal	Metric results for version: BLEU-4 - 40.14 Mod. ngram precision (1/2/3/4) - 52.1/42.0/41.0/41.8 ROUGE-L - 45.66 METEOR - 26.12
		Top 10,000 dataset	Version of the approach trained in the top 10,000 dataset with 400k epochs in training	
0.5	Context-aware Retrieval-based Deep Commit Message Generation	Original approach	Orignal approach	Metric results for version: BLEU-4 - 32.87 Mod. ngram precision (1/2/3/4) - 44.8/35.7/35.7/37.6 ROUGE-L - 39.24 METEOR - 21.66
		Dataset split by project	Version trained on top 10,000 datset split by project	
1.1	COME: Commit Message Generation with Modification Embedding	Potential reimplementation	Not specified in paper which approaches are reimplemented or which took the source code	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 18.51/18.41/14.02/15.09/21.30 ROUGE-L -

CoRec

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
		Different dataset	Trained in MCMD Dataset	24.78/23.73/20.10/22.35/27.53 METEOR - 11.26/11.70/8.63/9.60/13.84 CIDEr - 1.13/1.12/0.72/0.80/1.40
1.2	COME: Commit Message Generation with Modification Embedding	Potential reimplementation	Not specified in paper which approaches are reimplemented or which took the source code	Metric results for version : BLEU-4 - 13.06 ROUGE-L - 15.37 METEOR - 6.44 CIDEr - 0.38
		Different dataset	Trained in CoDiSum Dataset	
2	FIRA: Fine-Grained Graph-Based Code Change Representation for Automated Commit Message Generation	Different Dataset	Trained on CoDiSum dataset (same as version 1.2). Different version number assigned because 1.2 has a vague specification on wether it might be a reimplementation while FIRA specifies is not reimplemented.	Metric results for version: B-Norm BLEU - 13.03 Penalty BLEU - 10.49 ROUGE-L - 15.47 METEOR - 12.04
3.1	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in original Jiang et al. dataset	Metric results for version: BLEU-MOSES - 41.14 BLEU-B-Norm - 36.93 BLEU-CC - 30.74 ROUGE-1 - 40.40 ROUGE-2 - 29.91 ROUGE-L - 40.12 METEOR - 38.54
3.2	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in NNGen dataset	Metric results for version: BLEU-MOSES - 19.44 BLEU-B-Norm - 25.38 BLEU-CC - 18.29 ROUGE-1 - 29.11 ROUGE-2 - 17.20 ROUGE-L - 28.71 METEOR - 27.53
3.3		Different dataset	Trained in CoDiSum dataset	

CoRec

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	A large-scale empirical study of commit message generation: models, datasets and evaluation			Metric results for version: BLEU-MOSES - 3.00 BLEU-B-Norm - 10.62 BLEU-CC - 5.15 ROUGE-1 - 12.52 ROUGE-2 - 3.33 ROUGE-L - 12.27 METEOR - 11.30
3.4	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset	Metric results for version(C++/C#/Java/Python/JS): BLEU-MOSES - 10.14/24.31/11.39/10.03/16.81 BLEU-B-Norm - 13.80/22.23/16.09/15.13/19.84 BLEU-CC - 11.58/20.62/20.62/12.68/12.09/16.91 ROUGE-1 - 16.80/25.07/18.95/19.14/23.68 ROUGE-2 - 9.79/18.56/10.19/10.61/15.32 ROUGE-L - 16.62/24.87/18.66/18.81/23.36 METEOR - 17.42/25.38/19.58/20.29/23.84
3.5	A large-scale empirical study of commit message generation: models, datasets and evaluation	Different dataset	Trained in MCMD dataset split by project	Metric results for version (Average): BLEU-B-Norm - 7.19 ROUGE-1 - 9.17 ROUGE-2 - 4.52 ROUGE-L - 9.06 METEOR - 9.07
3.6		Different dataset	Trained in MCMD dataset split by timestamp	

CoRec

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	A large-scale empirical study of commit message generation: models, datasets and evaluation			Metric results for version (Average): BLEU-B-Norm - 12.17 ROUGE-1 - 14.48 ROUGE-2 - 7.54 ROUGE-L - 14.29 METEOR - 16.12

CoreGen

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	CoreGen: Contextualized Code Representation Learning for Commit Message Generation	Original approach	N/A	Metric results for version: BLEU-4 - 21.06 ROUGE-1/2/L - 32.87/20.17/30.85 METEOR - 16.53
0.1	CoreGen: Contextualized Code Representation Learning for Commit Message Generation	Ablation of Stage I	Stage I of CoreGen, which corresponds to the previous training to learn code representations of explicit code changes and implicit changes on binary files. Transformer directly trained from base weights in Stage II instead of trained in learned weights	Metric results for version: BLEU-4 - 18.74 ROUGE-1/2/L - 30.65/18.06/28.86 METEOR - 15.18
0.2	CoreGen: Contextualized Code Representation Learning for Commit Message Generation	Different dataset	Trained in variant of dataset without duplicates from original dataset. Around 15% of entries from validation and test sets are removed because of duplicities.	Metric results for version: BLEU-4 - 15.86 ROUGE-1/2/L - 27.31/15.09/25.46 METEOR - 13.38
0.3	CoreGen: Contextualized Code Representation Learning for Commit Message Generation	Addition of In-Statement Code Structure Modelling (ICSM)	In last experiments, authors include a new task to learn contextualized code changes because code structure is usually more rigid than natural language. This mechanism enhances original approach to predict new token based on the rest of tokens in the same statement	Metric results for version: BLEU-4 - 21.37 ROUGE-1/2/L - 32.88/20.33/32.87 METEOR - 16.72
0.4	CoreGen: Contextualized Code Representation Learning for Commit Message Generation	Change of loss function	Change individual loss functions from training stages I and II and combining them as a hybrid approach which only takes 1 training stage with a loss function being the sum of training stages functions	Metric results for version: BLEU-4 - 15.41 ROUGE-1/2/L - 22.15/11.04/20.71 METEOR - 13.79

CoreGen

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
1.1	COME: Commit Message Generation with Modification Embedding	Potential reimplementation Different dataset	Not specified in paper which approaches are reimplemented or which took the source code Trained in MCMD Dataset	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 21.30/17.08/16.74/17.74/22.53 ROUGE-L - 28.04/22.74/22.83/24.75/29.23 METEOR - 13.17/11.36/11.72/12.22/14.91 CIDEr - 1.31/0.94/0.90/0.97/1.50
1.2	COME: Commit Message Generation with Modification Embedding	Potential reimplementation Different dataset	Not specified in paper which approaches are reimplemented or which took the source code Trained in CoDiSum Dataset	Metric results for version : BLEU-4 - 14.10 ROUGE-L - 17.92 METEOR - 6.79 CIDEr - 0.38
2	FIRA: Fine-Grained Graph-Based Code Change Representation for Automated Commit Message Generation	Different dataset	Trained on CoDiSum dataset (same as version 1.2). Different version number assigned because 1.2 has a vague specification on wether it might be a reimplementation while FIRA specifies is not reimplemented.	Metric results for version: B-Norm BLEU - 14.15 Penalty BLEU - 11.15 ROUGE-L - 18.22 METEOR - 12.90
3	Mucha: Multi-channel based Code Change Representation Learning for Commit Message Generation	Different dataset	Trained in variation of Tufano et al. used for Mucha training and measurement	Metric results for version: BLEU-4 - 6.72 ROUGE-L - 6.59 METEOR - 2.51
4.1	Revisiting Learning-based Commit Message Generation	Different dataset	Trained in CoDiSum dataset and evaluated in both training set and test set. Different results than previous trained in CoDiSum dataset	Metric results for version (Training set/Test set): BLEU-4 - 17.17/13.53 ROUGE-L - 22.54/ 17.31 METEOR - 16.30/11.72

CoreGen

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
4.2	Revisiting Learning-based Commit Message Generation	Input Change	Trained in CoDiSum dataset feeding as input only the code change marks instead of whole commit entry	Metric results for version: BLEU-4 - 10.28 ROUGE-L - 13.35 METEOR - 8.21
		Different dataset	Trained in CoDiSum dataset with different results	

FIRA

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	FIRA: Fine-Grained Graph-Based Code Change Representation for Automated Commit Message Generation	Original approach	N/A	Metric results for version: B-Norm BLEU - 17.67 Penalty BLEU - 13.30 ROUGE-L - 21.58 METEOR - 14.93
0.1	FIRA: Fine-Grained Graph-Based Code Change Representation for Automated Commit Message Generation	Edit operations removed	For ablation study. Edit operation inclusion between graph TOKEN and graph FINAL is removed, so input does not contain information about the edit actions that were performed between the chopped AST (see preprocessing)	Metric results for version: B-Norm BLEU - 17.39 ROUGE-L - 21.15 METEOR - 14.54
0.2	FIRA: Fine-Grained Graph-Based Code Change Representation for Automated Commit Message Generation	Single copy mechanism	For ablation study. Degradation of dual copying mechanism proposed in the paper and substiution for a single copy mechanism that is not able to detect and copy subtokens from integral token nodes of ASTs (see preprocessing)	Metric results for version: B-Norm BLEU - 17.36 ROUGE-L - 20.97 METEOR - 14.09
0.3	FIRA: Fine-Grained Graph-Based Code Change Representation for Automated Commit Message Generation	Edit operations removed	For ablation study. Edit operation inclusion between graph TOKEN and graph FINAL is removed, so input does not contain information about the edit actions that were performed between the chopped AST (see preprocessing)	Metric results for version: B-Norm BLEU - 16.82 ROUGE-L - 20.15 METEOR - 13.42

FIRA

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
		Single copy mechanism	For ablation study. Degradation of dual copying mechanism proposed in the paper and substiution for a single copy mechanism that is not able to detect and copy subtokens from integral token nodes of ASTs (see preprocessing)	
1	COME: Commit Message Generation with Modification Embedding	Potential reimplementation	Not specified in paper which approaches are reimplemented or which took the source code	Metric results for version : BLEU-4 - 17.66 ROUGE-L - 20.90 METEOR - 8.31 CIDEr - 0.56
		Different dataset	Trained in CoDiSum Dataset	
2	CCT5: A Code-Change-Oriented Pre-trained Model	Different dataset	Trained in CoDiSum Dataset as approach is designed for Java language and uses specific Java components, so main dataset MCMD used in CCT5 paper cannot be used.	No results reported in the paper comparison, it is said that results are taken from Dong et al. reported results, but none showed in paper
3.1	Revisiting Learning-based Commit Message Generation	Different dataset	Trained in CoDiSum dataset and evaluated in both training set and test set. Different results than previous trained in CoDiSum dataset	Metric results for version (Training set/Test set): BLEU-4 - 18.76/17.48 ROUGE-L - 22.83/21.19 METEOR - 16.06/14.58
3.2	Revisiting Learning-based Commit Message Generation	Input Change	Trained in CoDiSum dataset feeding as input only the code change marks instead of whole commit entry	Metric results for version: BLEU-4 - 15.42 ROUGE-L - 18.07 METEOR - 11.62
		Different dataset	Trained in CoDiSum dataset with different results	
3.4				

FIRA

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
	Revisiting Learning-based Commit Message Generation	Ablation of copy mechanism	Ablation study performed deactivating the copy mechanism for OOV out of the architecture	Metric results for version: BLEU-4 - 17.13 ROUGE-L - 20.73 METEOR - 14.13
		Different dataset	Trained in CoDiSum dataset with different results	
3.5	Revisiting Learning-based Commit Message Generation	Ablation of anonymization mechanism	Ablation study performed deactivating the anonymization mechanism out of the architecture	Metric results for version: BLEU-4 - 17.19 ROUGE-L - 21.31 METEOR - 14.40
		Different dataset	Trained in CoDiSum dataset with different results	

RACE

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	RACE: Retrieval-Augmented Commit Message Generation	Original approach	N/A	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 25.66/26.33/19.13/21.79/25.55 ROUGE-L - 32.02/31.31/24.52/28.35/31.79 METEOR - 15.46/16.37/12.55/14.68/16.31 CIDEr - 1.76/1.84/1.14/1.40/1.84
0.1	RACE: Retrieval-Augmented Commit Message Generation	Exemplar Guider ablation	For ablation study, exemplar guider component from Generation module removed from system.	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 23.37/21.33/17.43/19.44/23.39 ROUGE-L - 30.01/27.33/22.03/26.4/30.51 METEOR - 13.98/13.56/12.10/13.89/15.64 CIDEr - 1.53/1.31/0.95/1.01/1.54
1.1	COME: Commit Message Generation with Modification Embedding	Potential reimplementation Different dataset	Not specified in paper which approaches are reimplemented or which took the source code Trained in MCMD Dataset	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 25.66/26.33/19.13/21.79/25.55 ROUGE-L - 32.02/31.31/24.52/28.35/31.79 METEOR - 15.46/16.37/12.55/14.68/16.31 CIDEr - 1.76/1.84/1.14/1.40/1.84
1.2	COME: Commit Message Generation with Modification Embedding	Potential reimplementation Different dataset	Not specified in paper which approaches are reimplemented or which took the source code Trained in CoDiSum Dataset	Metric results for version : BLEU-4 - 17.23 ROUGE-L - 20.49 METEOR - 8.57 CIDEr - 0.57

RACE

Version num.	Variation origin paper	Differences	Difference details	Difference in performance

ReGenSD

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	N/A	Metric results for version: BLEU 1/2/3/4 - 23.50/13.52/10.24/8.78
0.1	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	The message paired to the retrieved entry is not used as part of the input for the Guide module	Metric results for version: BLEU 1/2/3/4 - 21.9/11.65/8.21/6.66
0.2	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	All the enchancement mechanisms are removed: the Selection module, the Difference Vectors and the Relation Gates	Metric results for version: BLEU 1/2/3/4 - 21.85/12.07/8.79/7.12
0.3	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	Removed Difference Vectors and Relation Gates. Only Selection module is kept	Metric results for version: BLEU 1/2/3/4 - 22.64/12.49/8.96/7.35
0.4	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	Removed Difference Vectors and Selection module. Only Relation Gate is kept	Metric results for version: BLEU 1/2/3/4 - 23.33/13.29/9.73/7.91
0.5	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	Removed Selection module and Relation Gates. Only Difference Vectors are kept	Metric results for version: BLEU 1/2/3/4 - 22.92/13.23/9.86/8.19
0.6	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	Removed Relation Gate from enhancement mechanism for ablation purposes	Metric results for version: BLEU 1/2/3/4 - 23.23/13.33/9.93/8.24
0.7	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	Removed Difference Vectors from enhancement mechanism for ablation purposes	Metric results for version: BLEU 1/2/3/4 - 23.34/13.43/9.99/8.27
0.8	Retrieve-Guided Commit Message Generation with Semantic Similarity And Disparity	Original approach	Removed Selection module from enhancement mechanism for ablation purposes	Metric results for version: BLEU 1/2/3/4 - 23.10/13.40/10.22/8.8

CommitGen

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	A Neural Architecture for Generating Natural Language Descriptions from Source Code Changes	Original approach	N/A	Metric results for version (average): BLEU-4 - 15.82 METEOR (Val. Acc.) - 49.55(%)
0.1	A Neural Architecture for Generating Natural Language Descriptions from Source Code Changes	Different dataset	Trained in full dataset	Metric results for version (average): BLEU-4 - 12.01 METEOR (Val. Acc.) - 51.2(%)
1.1	COME: Commit Message Generation with Modification Embedding	Potential reimplementation Different dataset Different name	Not specified in paper which approaches are reimplemented or which took the source code Trained in MCMD Dataset Referred to as CommitGen in the paper	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 14.07/ 13.38/ 11.52/11.02/18.67 ROUGE-L - 18.78/17.44/16.75/16.64/24.10 METEOR - 7.52/8.31/6.98/6.43/11.88 CIDEr - 0.66/0.63/0.45/0.42/1.08
1.2	COME: Commit Message Generation with Modification Embedding	Potential reimplementation Different dataset Different name	Not specified in paper which approaches are reimplemented or which took the source code Trained in CoDiSum Dataset Referred to as CommitGen in the paper	Metric results for version : BLEU-4 - 6.28 ROUGE-L - 8.91 METEOR - 3.74 CIDEr - 0.13
2.1	RACE: Retrieval-Augmented Commit Message Generation	Potential reimplementation Different dataset Different name	Not specified in paper which approaches are reimplemented or which took the source code Trained in MCMD Dataset Referred to as NMTGen in the paper, while the othe NMT is referred to as CommitGen (this approach)	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 15.52/12.71/11.57/11.41/18.22 ROUGE-L - 21.13/17.16/17.46/18.43/24.43 METEOR -

CommitGen

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
				8.91/8.11/7.06/7.18/12.07 CIDEr - 0.86/0.62/0.51/0.48/1.12
2.2	RACE: Retrieval-Augmented Commit Message Generation	Reimplementation	Reimplemented approach making use of code diff encoder trained from RACE approach to obtain high dimensional semantic vectors	Metric results for version (C++/C#/Java/Python/JS): BLEU-4 - 18.85/15.51/13.40/13.42/22.27
		Different dataset	Trained in MCMD Dataset	
		Different name	Referred to as NMTGen in the paper, while the othe NMT is referred to as CommitGen (this approach)	
3	Evaluating Commit Message Generation: To BLEU Or Not To BLEU?	Different and new metric	Performance measured over MCMD dataset (already used) but evaluated in newly proposed metric	Metric results for version (C++/C#/Java/Python/JS): Log-MNEXT - 11.94/18.35/9.63/18.11/8.27

Mucha

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	Mucha: Multi-channel based Code Change Representation Learning for Commit Message Generation	Original approach	N/A	Metric results for version: BLEU-4 - 11.37 ROUGE-L - 7.28 METEOR - 8.27
0.1	Mucha: Multi-channel based Code Change Representation Learning for Commit Message Generation	Ablation of AST channel	From 3 level multi-channel (Line, token and AST levels), removed effect of AST channel and processing for ablation study purposes	Metric results for version: BLEU-4 - 9.82 ROUGE-L - 5.32 METEOR - 7.70
0.2	Mucha: Multi-channel based Code Change Representation Learning for Commit Message Generation	Ablation of Line channel	From 3 level multi-channel (Line, token and AST levels), removed effect of Line channel and processing for ablation study purposes	Metric results for version: BLEU-4 - 10.58 ROUGE-L - 6.44 METEOR - 7.85
0.3	Mucha: Multi-channel based Code Change Representation Learning for Commit Message Generation	Ablation of Token channel	From 3 level multi-channel (Line, token and AST levels), removed effect of Token channel and processing for ablation study purposes	Metric results for version: BLEU-4 - 10.14 ROUGE-L - 5.74 METEOR - 7.91
0.4	Mucha: Multi-channel based Code Change Representation Learning for Commit Message Generation	Different initialization of weights and biases	Initialization of pre-trained model for Encoders on CodeBERT weights and biases instead of the main initialization based on UniXcoder parameters	Metric results for version: BLEU-4 - 4.75 ROUGE-L - 3.63 METEOR - 4.95

CCT5

Version num.	Variation origin paper	Differences	Difference details	Difference in performance
0	CCT5: A Code-Change-Oriented Pre-trained Model	Original approach	N/A	Metric results for version (C++/C#/Java/Python/JS): B-Norm BLEU - 17.64/25.53/20.80/21.37/24.94