Reproduction project

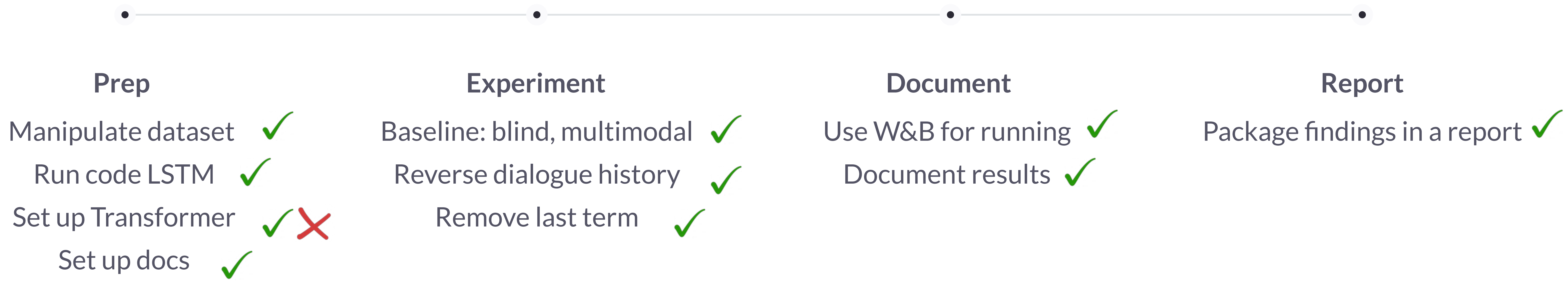
Ryazanskaya, Verma, Atanasoska

Agenda

- **Tamara** - overview of the project scope, what we achieved, what is left to do, conclusions
- **Bhuvanesh** - our contributions in the main research repo, summary of the results and our interpretations
- **Galina** unfortunately cannot join us today

Paper: *Greco, C., Testoni, A., & Bernardi, R. (2020). Which Turn do Neural Models Exploit the Most to Solve GuessWhat? Diving into the Dialogue History Encoding in Transformers and LSTMs. NL4AI@AI*IA.*

Our plan before starting



Prep

Manipulate dataset ✓

Run code LSTM ✓

Set up Transformer ✓ ✗

Set up docs ✓

Experiment

Baseline: blind, multimodal ✓

Reverse dialogue history ✓

Remove last term ✓

Document

Use W&B for running ✓

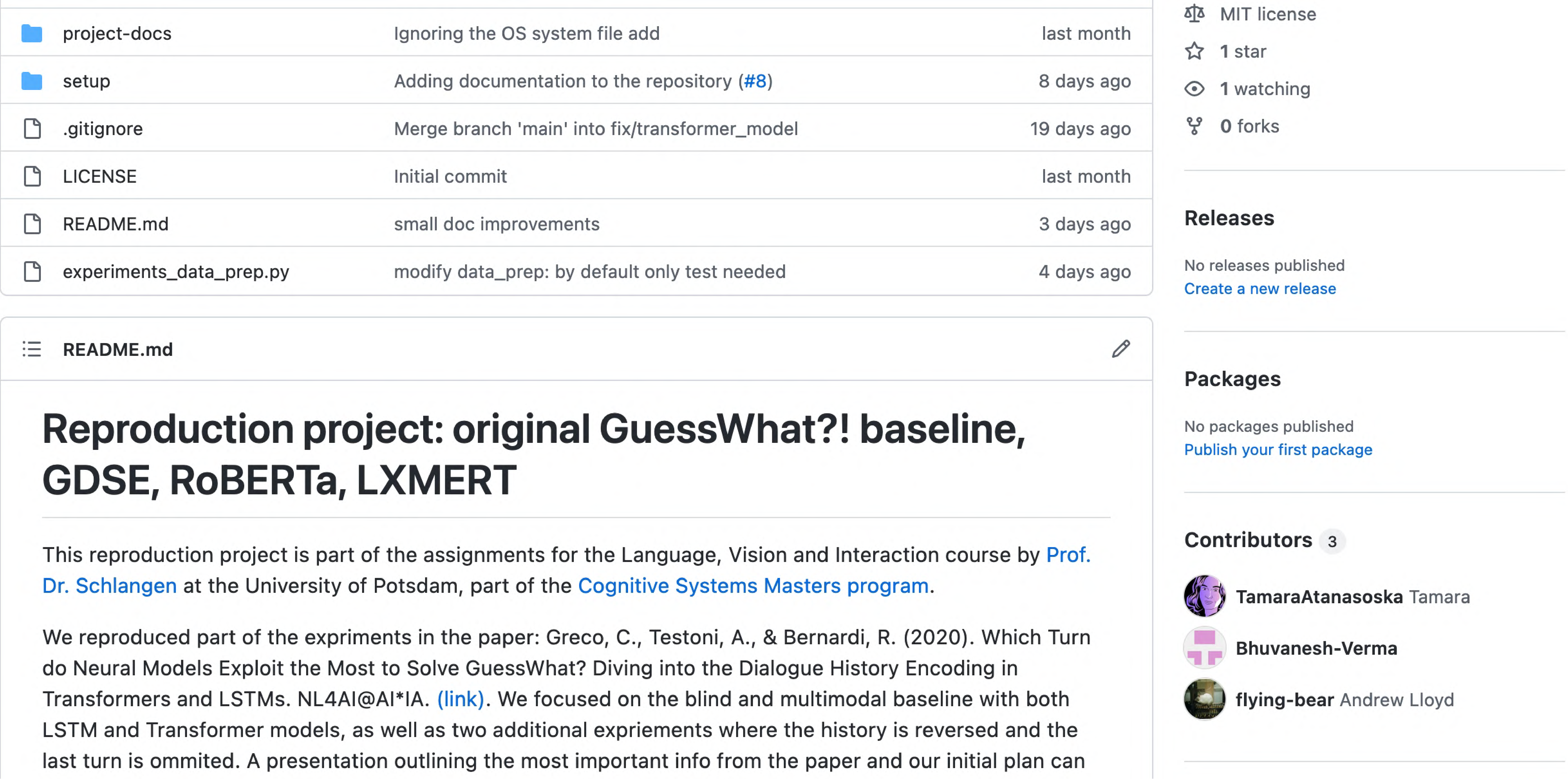
Document results ✓

Report

Package findings in a report ✓

What we achieved

- Obtained results from 4 models, across 2 repositories (original GuessWhat?! model and adapted original, GDSE, RoBERTa)
- Contributed back to the original GuessWhat?! repo (PRs: link, link, link) and it was merged ✨
- Kept an open conversation with the paper authors and exchanged knowledge and reasoning not present in the papers + code files and resources not available anywhere online
- Created an updated, comprehensive and detailed documentation with easily reproducible environments and complete running guide to help others reproducing after us
- Integrated the W&B platform to log our runs and get meaningful results



project-docs	Ignoring the OS system file add	last month
setup	Adding documentation to the repository (#8)	8 days ago
.gitignore	Merge branch 'main' into fix/transformer_model	19 days ago
LICENSE	Initial commit	last month
README.md	small doc improvements	3 days ago
experiments_data_prep.py	modify data_prep: by default only test needed	4 days ago

README.md

Reproduction project: original GuessWhat?! baseline, GDSE, RoBERTa, LXMERT

This reproduction project is part of the assignments for the Language, Vision and Interaction course by Prof. Dr. Schlangen at the University of Potsdam, part of the Cognitive Systems Masters program.

We reproduced part of the expriments in the paper: Greco, C., Testoni, A., & Bernardi, R. (2020). Which Turn do Neural Models Exploit the Most to Solve GuessWhat? Diving into the Dialogue History Encoding in Transformers and LSTMs. NL4AI@AI*IA. (link). We focused on the blind and multimodal baseline with both LSTM and Transformer models, as well as two additional exprieriments where the history is reversed and the last turn is ommited. A presentation outlining the most important info from the paper and our initial plan can

MIT license

1 star

1 watching

0 forks

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Contributors 3

TamaraAtanasoska Tamara

Bhuvanesh-Verma

flying-bear Andrew Lloyd

https://github.com/TamaraAtanasoska/dialogue-history

Contact with Authors

- Provided access to Github repository
- Provided multiple missing files
  - QGen and QGenImgCap scripts
  - MSCOCO bottom up features for LXMERT
- Help understand code structure

Roadblocks

- Setting up training environment
  - no requirement.txt
  - GPU machine incompatible
- Bugs in train pipeline
  - missing parameters in configs
  - mismatch in keys to access data from GuessWhat jsons
- Inconsistency in features created using the feature scripts

How it finally worked

- Train script requires data directory containing
  - GuessWhat data
  - N2N data files (if not available then created using GuessWhat data)
  - ResNet features
  - Vocabulary file (if not available then created using GuessWhat data)
- It also requires a config which contains parameters for various modules like optimizer, models, data paths etc
- N2N data files
  - manipulated guesswhat dataset based on parameters like *successful_only*, *max_no_qs* etc
  - parameters are provided in config file

New Changes to Repository

- Added checkpoint loading
- Integrated W&B experiment tracking framework, more at https://wandb.ai/we/lv
- Add train script for blind LSTM model
- Add Test related scripts
  - extract features for test data : ResNet image and object features
  - test LSTM and BERT based model
- Changes in config
  - number of epochs to 30 for LSTM models and 20 for BERT based models
  - batch size for training LSTM models to 32
- More is coming ...

Replication results: Task Success

	LSTM	V-LSTM	RoBERTa
original	64.7	64.5	67.9
replication	65.3	65.0	68.7

- Overall replication accuracy closely matches the original
- RoBERTa model is the best-performing one
- Blind and multimodal LSTMs perform similarly

Differences:

- batch size
- possibly, different random seed handling dependent on PyTorch versions

Replication results: No Last Turn

	LSTM	**V-LSTM**	**RoBERTa**
original	46.2 (18.5)	49.8 (14.7)	44.7 (23.2)
replication	47.3 (18)	47.5 (**17.5**)	**52.0** (**16.7**)

- LSTM and V-LSTM replication accuracy on the no-last-turn set is similar to the original
- RoBERTa model, unlike the original results, is the best-performing one
- Blind and multimodal LSTMs perform similarly (more so than in the original)

Differences:

- Results reported for across all turns, but in the original reported results for 3, 5, 8 turn dialogues
- possible differences in the code versions of LSTM between PyTorch and Tensorflow versions

Replication results: Reversed History

	LSTM	**V-LSTM**	**RoBERTa**
original	56 (8.7)	51.3 (13.2)	66.5 (1.4)
replication	**49.2** (**16.1**)	53.2 (11.8)	67.2 (1.5)

- RoBERTa and V-LSTM results are quite close to the original
- RoBERTa model is the best-performing one
- LSTM in the replication does is not as robust to changes in dialogue history

Differences:

- possible differences in the code versions of LSTM between PyTorch and Tensorflow versions

Conclusions

- The replication experiment was successful
- We were able to make the provided code run
- We were able to replicate the general findings in the scope that we selected for the project:
  - We observed that Transformer-based models (RoBERTa in our case) outperformed RNN-based ones
  - We observed that blind (LSTM) and multimodal (V-LSTM) models performed very similarly
  - We confirmed that the largest difference between the models was observed on no-last-turn
  - We confirmed that RoBERTa was the most robust to changes in dialogue history, thus being most able to identify salient information
    - In our project, it was so across the experiments, while in the original RoBERTa was not the best model in the no-last-turn experiment

What we didn't manage, future plans

- We obtained the features needed to run LXMERT, however we lacked computational resources
- Primary goal was to reproduce : we keep a list of small improvements to make

Future plans:

- Remove the raw category from the dataset and run all the experiments again (authors suggestion)
- Run LXMERT (we have strategies in mind to compensate)
- Contribute back to the authors
- Implement some of the small improvements, as time allows