* Single instruction + HTML + action
* Single action history + instruction HTML + action

Encoder only models (BERT): text input, categorical output.

Semantic categories as one-hot encoding. Add classification layer at the end to get distribution of over semantic categories

Ideas to augment the dataset:

* Do upper and lower case on some of the characters or tags
* Create synthetic data for some tasks based on the ones received

Task list:

* Need to create snippet function have these requirements:
  + all necessary clicked ids must be present
  + max length 512
  + have as much data fields as possible

For snippet, we only take the tree going deeper

Observations:

In the training data, we observe that a lot of actions are duplicated. We could reprocess this and remove every duplicates for better quality.

ACTION = FUNCTION ( SELECTOR, TEXT )

Function: clicking or typing → Use those as actions instead of all 8.

Selector: ref of element

Text: what to input into the field only if the function with ‘typing’ is activated

⇒ Target to predict should be a string with the format {click/text}{ref}{typed action}

If the text is predicted, can we also do it regarding the ref? Should be possible

Try to show that with augmenting the dataset, we can train smaller models like BERT with the same accuracy as larger ones like T5 models. Maybe the issue is the input size which is bigger for T5 models (to confirm) and thus the input snippet is biased for BERT. But if we can find a method to produce the right attention snippet, BERT can do the actions just fine. Perhaps create new snippets of part of the page where the target ref is not present, and the corresponding label simply says that it is not the right part of the page and that we should look elsewhere.

We observe that the model predicts many times to click on ref=1. This is a mistake as most of the time this does not produce anything. Thus we must clean that and retrain. ⇒ Do experiment a)

Experiments:

1. Use test set and make predictions by varying the inputs such as modifying the previous history and see what changes we get.
2. Try to predict the full action history directly from the start (need to recompute a dataset for that).

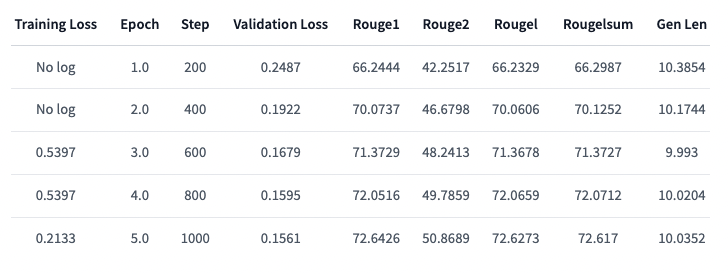
izzeddingur@gmail.com

Questions to the authors:

* How do they perform the addition of history of past actions? Did they include them all from a state, or applied a method to select some of them? What do they think can be done to improve on that? Section 7
* How do you pass the key pressed when doing login etc. Is the target embedded in the action tuples? What about the element reference?
* Do you pass the full string in the ‘type’ action, or do you do it character by character?
* What was the criteria for the snippet selection?

Todo: Need to recompute evaluation for final model, don’t trust the logs

Balanced Dataset Training Log:



| Training Loss | Epoch | Step | Validation Loss | Rouge1 | Rouge2 | Rougel | Rougelsum | Gen Len |

|:-------------:|:-----:|:----:|:---------------:|:-------:|:-------:|:-------:|:---------:|:-------:|

| No log | 1.0 | 200 | 0.2487 | 66.2444 | 42.2517 | 66.2329 | 66.2987 | 10.3854 |

| No log | 2.0 | 400 | 0.1922 | 70.0737 | 46.6798 | 70.0606 | 70.1252 | 10.1744 |

| 0.5397 | 3.0 | 600 | 0.1679 | 71.3729 | 48.2413 | 71.3678 | 71.3727 | 9.993 |

| 0.5397 | 4.0 | 800 | 0.1595 | 72.0516 | 49.7859 | 72.0659 | 72.0712 | 10.0204 |

| 0.2133 | 5.0 | 1000 | 0.1561 | 72.6426 | 50.8689 | 72.6273 | 72.617 | 10.0352 |

Training Longer Epochs

| Training Loss | Epoch | Step | Validation Loss | Rouge1 | Rouge2 | Rougel | Rougelsum | Gen Len |

|:-------------:|:-----:|:----:|:---------------:|:-------:|:-------:|:-------:|:---------:|:-------:|

| No log | 1.0 | 200 | 0.2444 | 66.7817 | 42.6651 | 66.7626 | 66.7919 | 10.3601 |

| No log | 2.0 | 400 | 0.1814 | 70.0918 | 46.5921 | 70.1243 | 70.1434 | 10.2082 |

| 0.5272 | 3.0 | 600 | 0.1547 | 72.4743 | 50.6706 | 72.5164 | 72.4974 | 10.0366 |

| 0.5272 | 4.0 | 800 | 0.1390 | 73.8367 | 53.2101 | 73.8745 | 73.904 | 10.0985 |

| 0.1909 | 5.0 | 1000 | 0.1283 | 73.9474 | 53.5947 | 73.9864 | 74.0176 | 10.1153 |

| 0.1909 | 6.0 | 1200 | 0.1237 | 74.629 | 54.6033 | 74.6188 | 74.6671 | 10.0429 |

| 0.1909 | 7.0 | 1400 | 0.1191 | 75.0201 | 55.5809 | 75.0455 | 75.0941 | 10.0359 |

| 0.1637 | 8.0 | 1600 | 0.1170 | 75.0738 | 55.6533 | 75.0934 | 75.1241 | 10.052 |

| 0.1637 | 9.0 | 1800 | 0.1155 | 75.4808 | 56.466 | 75.4776 | 75.515 | 10.0183 |

| 0.1525 | 10.0 | 2000 | 0.1153 | 75.5284 | 56.5291 | 75.5377 | 75.5656 | 10.0239 |

| Training Loss | Epoch | Step | Validation Loss | Rouge1 | Rouge2 | Rougel | Rougelsum | Gen Len |

|:-------------:|:-----:|:----:|:---------------:|:-------:|:-------:|:-------:|:---------:|:-------:|

| No log | 1.0 | 200 | 0.1090 | 76.0748 | 57.3974 | 75.9994 | 76.0099 | 10.019 |

| No log | 2.0 | 400 | 0.1049 | 76.6283 | 58.6213 | 76.6019 | 76.6361 | 10.0056 |

| 0.1404 | 3.0 | 600 | 0.0998 | 77.0808 | 59.506 | 77.0629 | 77.0568 | 9.9951 |

| 0.1404 | 4.0 | 800 | 0.0973 | 77.4654 | 60.2912 | 77.4984 | 77.4749 | 9.9782 |

| 0.1225 | 5.0 | 1000 | 0.0940 | 77.7612 | 60.593 | 77.8149 | 77.8074 | 9.9944 |

| 0.1225 | 6.0 | 1200 | 0.0928 | 78.2849 | 61.6607 | 78.2716 | 78.2756 | 9.9796 |

| 0.1225 | 7.0 | 1400 | 0.0908 | 78.1167 | 61.2447 | 78.1233 | 78.0901 | 9.9831 |

| 0.1146 | 8.0 | 1600 | 0.0891 | 78.271 | 61.493 | 78.3078 | 78.2791 | 9.9677 |

| 0.1146 | 9.0 | 1800 | 0.0873 | 78.4701 | 61.9206 | 78.458 | 78.4623 | 9.9782 |

| 0.109 | 10.0 | 2000 | 0.0863 | 78.8487 | 62.6432 | 78.8643 | 78.8634 | 9.9627 |

| 0.109 | 11.0 | 2200 | 0.0846 | 78.9234 | 62.7477 | 78.9508 | 78.935 | 9.9712 |

| 0.109 | 12.0 | 2400 | 0.0834 | 79.3395 | 63.5346 | 79.3691 | 79.3624 | 9.9698 |

| 0.1053 | 13.0 | 2600 | 0.0830 | 79.3797 | 63.6269 | 79.3946 | 79.3873 | 9.9817 |

| 0.1053 | 14.0 | 2800 | 0.0821 | 79.6115 | 64.0536 | 79.6639 | 79.6606 | 9.9944 |

| 0.1037 | 15.0 | 3000 | 0.0817 | 79.4473 | 63.6415 | 79.4805 | 79.4898 | 9.9979 |

| 0.1037 | 16.0 | 3200 | 0.0809 | 79.8574 | 64.4337 | 79.8826 | 79.8735 | 10.0007 |

| 0.1037 | 17.0 | 3400 | 0.0806 | 79.7134 | 64.2018 | 79.7691 | 79.7556 | 9.9887 |

| 0.1012 | 18.0 | 3600 | 0.0801 | 79.9594 | 64.6971 | 80.0013 | 80.0035 | 9.9937 |

| 0.1012 | 19.0 | 3800 | 0.0799 | 79.8193 | 64.4281 | 79.881 | 79.8639 | 9.9902 |

| 0.1003 | 20.0 | 4000 | 0.0798 | 79.9219 | 64.6124 | 79.9912 | 79.9735 | 9.9902 |

### Smaller Learning Rate, Iterations 70-80 T5-small, Dataset Balanced T5

### Training results

| Training Loss | Epoch | Step | Validation Loss | Rouge1 | Rouge2 | Rougel | Rougelsum | Gen Len |

|:-------------:|:-----:|:----:|:---------------:|:-------:|:-------:|:-------:|:---------:|:-------:|

| No log | 1.0 | 200 | 0.0499 | 86.15 | 76.4371 | 86.0738 | 86.1072 | 9.9972 |

| No log | 2.0 | 400 | 0.0495 | 86.3096 | 76.8752 | 86.2507 | 86.3009 | 9.9923 |

| 0.0696 | 3.0 | 600 | 0.0494 | 86.6349 | 77.4927 | 86.5773 | 86.6223 | 10.0 |

| 0.0696 | 4.0 | 800 | 0.0495 | 86.5155 | 77.1226 | 86.4466 | 86.4721 | 9.9909 |

| 0.0681 | 5.0 | 1000 | 0.0496 | 86.3998 | 76.9281 | 86.3245 | 86.3662 | 9.9902 |

| 0.0681 | 6.0 | 1200 | 0.0495 | 86.4666 | 77.1672 | 86.3928 | 86.4415 | 10.0 |

| 0.0681 | 7.0 | 1400 | 0.0493 | 86.386 | 76.9386 | 86.3181 | 86.3695 | 9.9958 |

| 0.0676 | 8.0 | 1600 | 0.0492 | 86.5671 | 77.302 | 86.5115 | 86.5403 | 9.9944 |

| 0.0676 | 9.0 | 1800 | 0.0492 | 86.4669 | 77.1537 | 86.392 | 86.4282 | 10.0021 |

| 0.0671 | 10.0 | 2000 | 0.0493 | 86.5054 | 77.2256 | 86.4378 | 86.4681 | 10.0007 |

The following must be added manually:

[2465/8000 22:06 < 49:40, 1.86 it/s, Epoch 12.32/40]

| **Epoch** | **Training Loss** | **Validation Loss** | **Rouge1** | **Rouge2** | **Rougel** | **Rougelsum** | **Gen Len** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | No log | 0.053614 | 85.215900 | 74.721100 | 85.231300 | 85.173700 | 9.997200 |
| 2 | No log | 0.053548 | 85.374600 | 74.931900 | 85.437100 | 85.347900 | 9.978900 |
| 3 | 0.067000 | 0.053249 | 85.395000 | 75.084500 | 85.410700 | 85.367800 | 10.004900 |
| 4 | 0.067000 | 0.053110 | 85.628900 | 75.602700 | 85.670500 | 85.595000 | 10.000700 |
| 5 | 0.066200 | 0.053818 | 85.282300 | 74.866900 | 85.311800 | 85.252400 | 10.004900 |
| 6 | 0.066200 | 0.053545 | 85.400700 | 75.075900 | 85.424100 | 85.355400 | 9.989500 |
| 7 | 0.066200 | 0.053645 | 85.287600 | 74.885900 | 85.293400 | 85.245000 | 10.007000 |
| 8 | 0.064800 | 0.053600 | 85.228300 | 74.705500 | 85.239500 | 85.186500 | 10.004200 |
| 9 | 0.064800 | 0.053405 | 85.231100 | 74.702800 | 85.224000 | 85.183500 | 10.004900 |
| 10 | 0.065100 | 0.052978 | 85.530800 | 75.301400 | 85.550300 | 85.496400 | 10.002100 |
| 11 | 0.065100 | 0.053386 | 85.397800 | 75.073400 | 85.418500 | 85.360000 | 9.990200 |
| 12 | 0.065100 | 0.053126 | 85.497400 | 75.289300 | 85.526700 | 85.470900 | 10.001400 |

\_\_\_\_\_\_\_\_\_\_\_\_

Split 2.3

| **Epoch** | **Training Loss** | **Validation Loss** | **Rouge1** | **Rouge2** | **Rougel** | **Rougelsum** | **Gen Len** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0.095100 | 0.071910 | 94.353500 | 89.892400 | 94.336400 | 94.345200 | 9.089900 |
| 2 | 0.093300 | 0.070967 | 94.412200 | 89.966700 | 94.402200 | 94.401000 | 9.086500 |
| 3 | 0.089200 | 0.070065 | 94.453700 | 90.027600 | 94.435400 | 94.434900 | 9.090500 |
| 4 | 0.088400 | 0.069274 | 94.461800 | 90.028400 | 94.445100 | 94.443800 | 9.088200 |
| 5 | 0.085000 | 0.068022 | 94.508200 | 90.125100 | 94.490700 | 94.491900 | 9.091800 |
| 6 | 0.084500 | 0.068234 | 94.410000 | 89.889500 | 94.397400 | 94.394900 | 9.101500 |
| 7 | 0.083400 | 0.067351 | 94.511100 | 90.016100 | 94.499900 | 94.506600 | 9.099500 |
| 8 | 0.079800 | 0.066639 | 94.699200 | 90.364800 | 94.681700 | 94.690600 | 9.108800 |

Split 3

EpochTraining LossValidation LossRouge1Rouge2RougelRougelsumGen Len10.0818000.06968594.08840089.37280094.08010094.0877009.18170020.0810000.06941594.18120089.58850094.18540094.1952009.18270030.0796000.06884294.17210089.53230094.16590094.1641009.18730040.0790000.06871594.11680089.38950094.09900094.1055009.19860050.0784000.06881494.08610089.35330094.08960094.0877009.18470060.0777000.06832994.13660089.44360094.14070094.1374009.18600070.0753000.06845494.12040089.41560094.11520094.1137009.19360080.0786000.06793194.15620089.50220094.15020094.1477009.18500090.0771000.06817294.14000089.44500094.14230094.1347009.192000100.0771000.06783894.23220089.61350094.23430094.2266009.189300

Training T5-based, 5 epochs, epoch 1-5, learning rate 2e-5

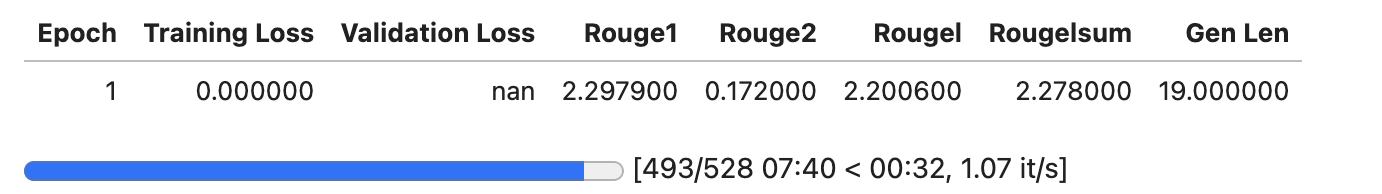
EpochTraining LossValidation LossRouge1Rouge2RougelRougelsumGen Len10.0631000.05851694.48530089.97120094.48530094.4762009.11490020.0633000.05790894.69560090.29810094.70170094.6901009.10860030.0525000.05733694.71930090.35240094.71390094.7149009.11370040.0548000.05793994.70120090.32010094.70090094.6998009.10660050.0575000.06161194.61450090.17420094.61350094.6156009.103800

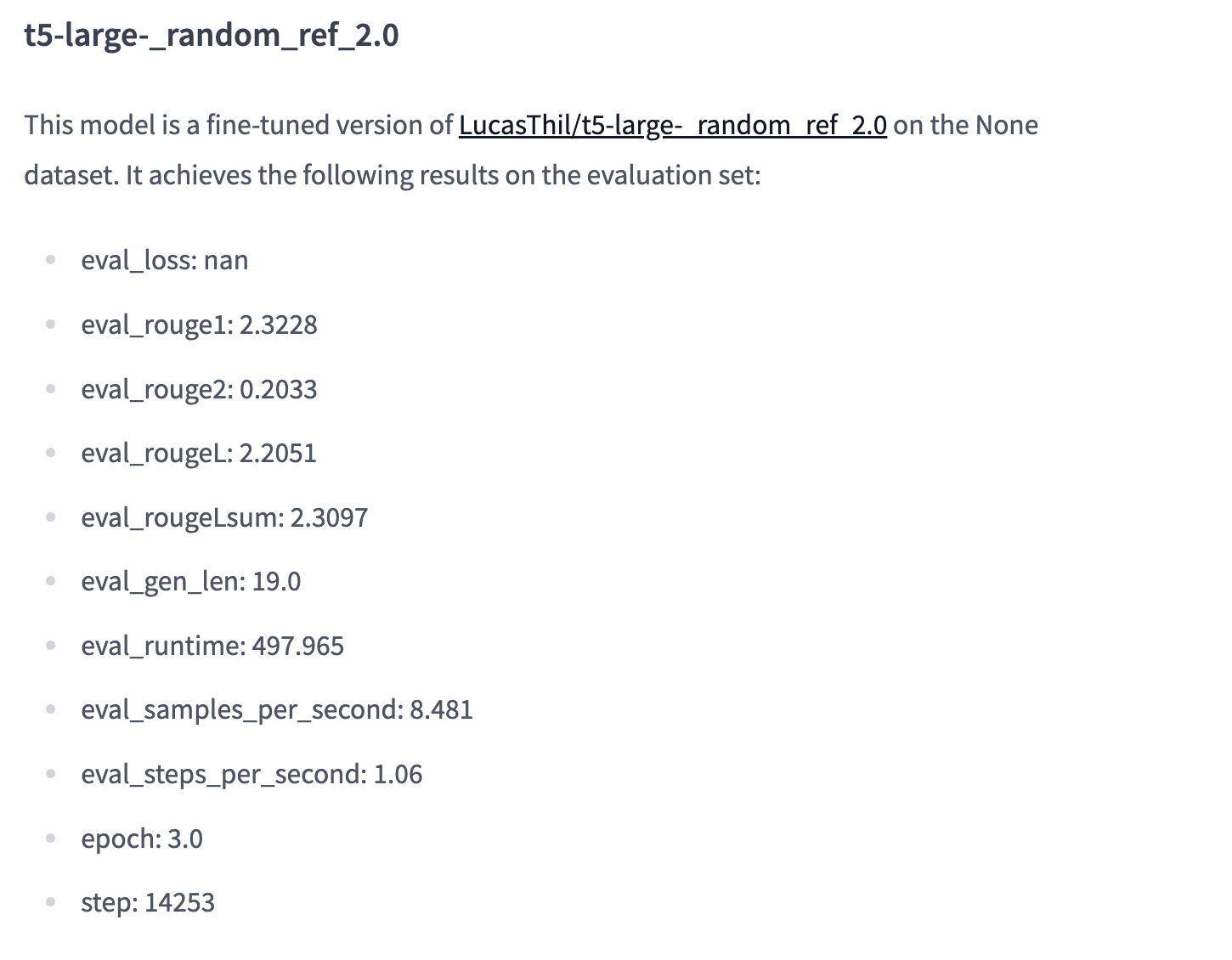
* 3 Epochs with 1e-5 learning rate:

EpochTraining LossValidation LossRouge1Rouge2RougelRougelsumGen Len10.0555000.05144994.81120090.70570094.81440094.8152009.15050020.0544000.05111694.83240090.71860094.83060094.8277009.14620030.0501000.05123594.85170090.74800094.85560094.8510009.148700

512 Tokens T5 base:

T5 Large first epochs, but we’re missing some steps due to page deconnection:





LucasThil/T5\_base\_hierarchy11\_256\_512:

Epoch1:

Training LossEpochStepValidation LossRouge1Rouge2RougelRougelsumGen Len0.07891.029920.08030.77810.50370.77660.776312.2667