Fech Scen Khoo, 11 September 2021
NASA Frontier Development Lab 2021

Earth Science Challenge: Automated Reporting of Natural Events Using Satellite Imagery and Metadata
(Also see: https://frontierdevelopmentlab.org/fdl-2021, then
https://www.calameo.com/read/005503280322eaa428773?page=1)

Experiment I. Hierarchical Earth Science Transformer (ESTR)

In this experiment, we work on a subset of the event-specific dataset, the Volcano events from our Earth Observatory (EO) dataset *provided by NASA*. We hereby propose our Volcanoes-Hierarchical ESTR model. In what follows, we observe how this model fairs alongside a few other baseline GPT-2-based models. For each set of model inputs taken from unseen data, we conditionally generate 3 samples. We report the best performing sample based on a metric, the rouge-L f1 score. Although it is debatable what the best means simply via a metric, it is sufficient as one of the ways to measure the model performance in terms of token similarity with respect to the ground-truth text, and this is precisely how the rouge-L metric functions.

We are interested in the outputs generated by the following 4 models. There are in total 8 inputs that we provide to the model for conditional generation in the final fine-tuning stage on Volcanoes, where 6 of them are metadata of the event: title, short caption, location, sensors, date and topics, in addition to our generated summary and extracted keywords.

Model A: The publicly available GPT-2. It is not trained or fine-tuned on any of our data sources.

Model B: GPT-2 fine-tuned on Volcano events from EO. The fine-tuning is performed with the 12 transformer layers in GPT-2 unfrozen.

Model C: GPT-2 fine-tuned on EO and subsequently on Volcanoes. This is hence an EO-fine-tuned GPT-2 on Volcanoes. GPT-2 is first conditionally fine-tuned on EO with only the 6 metadata, where 12 transformer layers are unfrozen. This is followed by another fine-tuning on Volcanoes with 12 layers unfrozen.

Model D: Adapted GPT-2 (i.e. an ESTR) fine-tuned on EO and subsequently on Volcanoes. This is our Volcanoes-Hierarchical ESTR model. GPT-2 is first adaptively fine-tuned on all our data sources where at each training epoch a random 75% of AGU data is fed to the model for a maximum of 50 epochs. The last 6 transformer layers are unfrozen during this training. Next, the adaptively fine-tuned model is conditionally fine-tuned again on all EO using 6 metadata, under 6 unfrozen last layers. Lastly the model is fine-tuned on Volcanoes with the last 6 layers unfrozen.

Note that the (base) model in Model A is a GPT-2 that has not seen our data sources, while the base models in Models B and C have only seen either partial or the entire EO. In Model D, i.e. our Volcanoes-Hierarchical ESTR model, it has an ESTR transformer as its base. We further remark that Model C has gone through 2 stages of fine-tuning, where it is first trained on the entire EO dataset and then again on Volcanoes which are part of the EO dataset. Meanwhile, our final Volcanoes-Hierarchical ESTR model has undergone the most fine-tuning in stages, from all data sources to EO to Volcanoes. The knowledge of Volcanoes is thrice distilled into the model learning.

For results, we present a generated sample by each model which exhibits the highest rouge-L f1 score, except for Model A where the sample shown here is not the model's highest-scored. These are the inputs provided to the models:

Title: Ever Restless Popocatépetl

Short caption: The volcano has been erupting since January 2005, with near constant venting from fumaroles, punctuated by minor steam, gas, and ash emissions.

Location: Tochimilco, Puebla, Mexico

Sensors: Terra - ASTER

Date: 2020-03-05 11:53:23

Topics: Atmosphere, Land, Volcanoes, Remote Sensing, Volcanoes

Summary: Nasa Terra satellite acquired a false-color image of a volcanic plume venting from the summit crater on February 25, 2020.

Keywords: February 25, 2020

The models' generated texts with their rouge-L f1 scores:

Model A (GPT2):

l Survey <lNASAlSGSlAUSTRALIAIPRAGUEIINEDIEUROPEIYUGOSLAVIA> Geostationary Observatory (GEO) - Landsat-2 - Earth-Moon Interplanetary Observatory, GEO-1, Geology and Geophysics, National Aeronautics and Space Administration (NASA) < lGSOl lSEl].

Rouge-L f1-score: 0.057

Model B (Volcanoes-GPT2):

The image was acquired by *the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)* aboard *NASAs Terra* spacecraft.

Rouge-L f1-score: 0.349 (Model from the epoch checkpoint with best Rouge-L recall score)

Model C (EO-GPT2 fine-tuned on Volcanoes):

The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) on NASAs Terra (March 5, 2010) captured this nighttime view of the eruption.

Rouge-L f1-score: 0.418 (Model from the last checkpoint)

Model D (Volcanoes-Hierarchical ESTR):

The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) on NASAs Terra and Aqua *satellites acquired this false color image* (top) of the ongoing eruption. *The image combines* near-*infrared, red, green*, blue, yellow, orange, shortwave-, and green light. *Vegetation* is *red*. Bare ground is brown.

Rouge-L f1-score: 0.511 (Model from the epoch checkpoint with best Rouge-L recall score)

The underlined sequences can be found in the model inputs. In italic are sequences found in the ground truth.

Ground-truth text:

On February 25, 2020, the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) on NASAs Terra satellite acquired this false-color image of a volcanic plume venting from

the summit crater. The image combines infrared, red, and green wavelengths such that the surrounding vegetation appears red.

We notice that the corresponding rouge-L f1 score increases from Model A to Model D. This metric behaviour is a positive indication aligned with our observation on the generated texts that an increasing amount of content is successfully reproduced across the models. Without any fine-tuning, Model A's output barely forms a comprehensible text. The length of the generated text grows with the number of fine-tuning from Models B to D.

Model hallucinations however do occur in the generation. For example, in Model B, a hallucination in terms of sensors occurs. The word "spacecraft" used is incorrect with respect to the ground truth. In Model C, the year "2010" generated is incorrect. From our Volcanoes-Hierarchical ESTR model, information such as "blue, yellow, orange,…" which does not appear in the group truth is generated.

We provide below the overall performance of the models over the Volcanoes test set based on the rouge-L f1 score of the best generated samples. The same increasing metric behaviour is observed from the left to the right model.

GPT2	Volcanoes-GPT2	EO-GPT2 fine-tuned on Volcanoes	Volcanoes-Hierarchical ESTR
0.108 ± 0.045	0.114 ± 0.047	0.125 ± 0.058	0.137 ± 0.058

Table: Mean and standard deviation of Rouge-L f1 score on the Volcanoes test set achieved by the GPT2-based models

In this experiment, we use rouge-L score to measure our text/paragraph regeneration. This score is not suited to justify the model correctness, as the model could have acquired the capability to produce information which although is not contained in the ground truth, but can be valid coherently as part of a narrative.

To summarize, the language model performs better with more fine-tuning. Model A (GPT2) without any fine-tuning performs the worst, whereas our Volcanoes-Hierarchical ESTR with the most fine-tuning shows the best promising results. The average Rouge-L f1 scored by Volcanoes-Hierarchical ESTR is 27% higher than GPT2. Besides the benefit of multi-fine-tuning, our experimental results also substantiate the strength of an adapted base model for text generation, as we note the improvement from Model C to D (Volcanoes-Hierarchical ESTR).

Experiment II. On the number of unfrozen transformer layers in GPT2

Following from the last experiment, in this section, we investigate the effect of unfreezing the last layers in the transformer during training. We report the corresponding rouge-L f1 score averaged over the respective test set.

Case 1:

Model: EO-GPT2
Model inputs: metadata + summary + keywords

6 unfrozen last layers	12 unfrozen layers
0.136 ± 0.035	0.138 ± 0.041

Table: Mean and standard deviation of rouge-L f1 score on the EO test set when the unfrozen layers are 6 and 12 in the EO-GPT2 model

Case 2:

Model: EO-GPT2
Model inputs: metadata

6 unfrozen last layers	12 unfrozen layers
0.138 ± 0.055	0.116 ± 0.055

Table: Mean and standard deviation of rouge-L f1 score on the EO test set when the unfrozen layers are 6 and 12 in the EO-GPT2 model

Case 3:

Model: Storms-GPT2
Model inputs: metadata + summary + keywords

6 unfrozen last layers	12 unfrozen layers
0.131 ± 0.044	0.128 ± 0.048

Table: Mean and standard deviation of rouge-L f1 score on the Storms test set when the unfrozen layers are 6 and 12 in the Storms-GPT2 model

Case 4:

Model: Floods-GPT2
Model inputs: metadata + summary + keywords

6 unfrozen last layers	12 unfrozen layers
0.143 ± 0.049	0.122 ± 0.038

Table: Mean and standard deviation of rouge-L f1 score on the Floods test set when the unfrozen layers are 6 and 12 in the Floods-GPT2 model

In case 1, the mean value of rouge-L f1 score is only slightly higher when all 12 layers in the transformer are unfrozen. While in other cases, the mean value of rouge-L f1 score is comparatively higher for 6 unfrozen last layers. Note that even when fewer inputs are provided to the model as in case 2 compared to case 1, the average value of the metric for the model trained with 6 unfrozen last layers remains approximately 0.14, while a decline of 16% is observed for the model trained with 12 layers unfrozen. Hence, the significance of the number of model inputs seems to surface when the model is trained with 12 unfrozen layers.