AI-Based War Crime Detection

By Egle Duobaite

Table of contents

01	Context	Russia's War Against Ukraine
02	Methodology	Telegram and Transformers
03	Video Processing	Frame selection and manual captioning
04	Baseline Model	Vision Transformer - GPT2 Model Results
05	Fine-tuned Model	Result comparison of 2 models
06	Performance evaluation	Rogue scores
07	Conclusions	
08	Limitations	

Context

01 Status Quo






Trial International: 'war crimes do not engage State responsibility but individual criminal responsibility.'

02 'War Feed'

'Telegram - a rapidly evolving weapon of psychological warfare'

03 Automating OSINT

NLP

Computer Vision

Forecasting and Prediction

Question answering;
Images of weapons, damage to infrastructure, location, criminal acts;
Prediction of fatalities, regime change.

Methodology

Data

OSINTUkraine – Ukrainian and Russian Telegram archive

Video Processing

Reducing frames with OpenCV

Sampling

5 random frames

Baseline model

Hugging Face: pre-trained Vision Transformer – GPT2 Captioning model

Fine-tuned model

Baseline model trained with 2 images and human-written descriptions

Evaluation

Rogue score: recall, precision, F1.
Human evaluation

DISCLAIMER

The following slides include media, displaying acts of torture and inhumane treatment of war prisoners.

Video Processing, Sampling and Captioning





Image 1 Caption 1: 'a man laying on the ground with his hands tied in the back, another man is kneeling next to him, two other blindfolded men are holding onto a person in front'
Image 1 Caption 2: "four men on a forest road, one man is laying on the ground, another man is kneeling, two other men are standing, three of them are in military uniforms'

Transformer

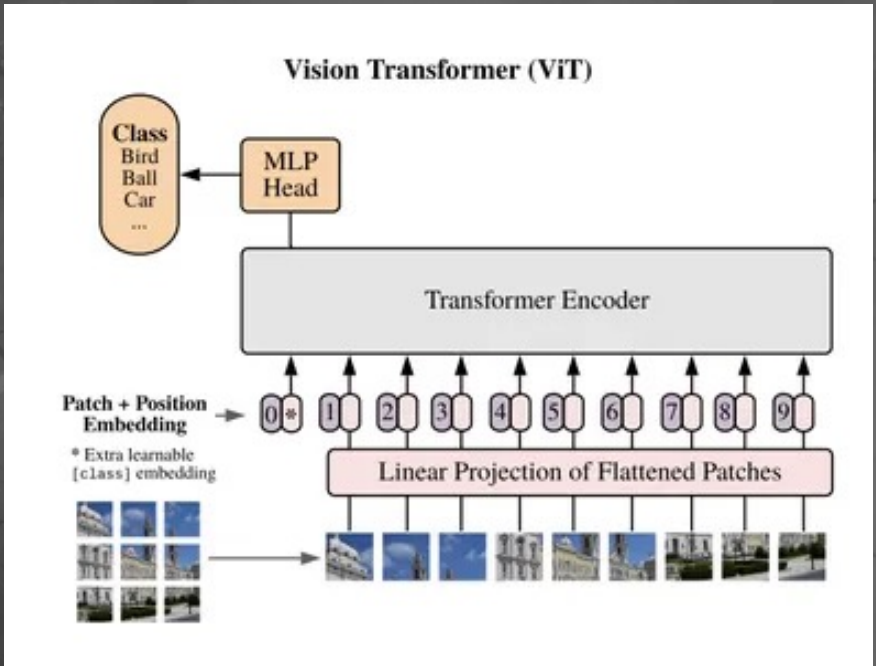


Image Embedding: Converts image patches into embeddings using linear projections.

Attention Mechanism: Applies self-attention across these embeddings to capture spatial relationships.

Feature Extraction: Generates hierarchical representations of images, enabling tasks like classification or object detection.

Baseline Model Results

'a man bending over to take a picture of another man'

‘two men in fatigues with a dog in the back’

‘a large group of animals laying on top of a blanket’

‘number of people standing on top of a dirt road’

‘men riding on the back of a motorcycle

Training (fine-tuning) the model

‘a handcuffed man walking next to man in a military uniform‘

‘a man in a military uniform walking next to a man in a military uniform‘

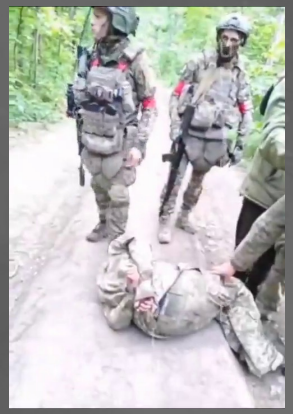
‘a man with one bandaged eye is surrendering to a man in a military uniform‘

‘ a man in military uniform with bandaged eye and bandaged eye‘

Comparison of the results

'a man **bending over** to take a picture of another man'

‘a man with a **bandaged eye** is **being held** by a man in a’



‘two men in **fatigues** with a dog in the back’

‘a man in a **military uniform** with a man in a **military uniform’**

‘a large group of animals laying on top of a blanket’

’ a man with a bandaged knee is holding a gun’

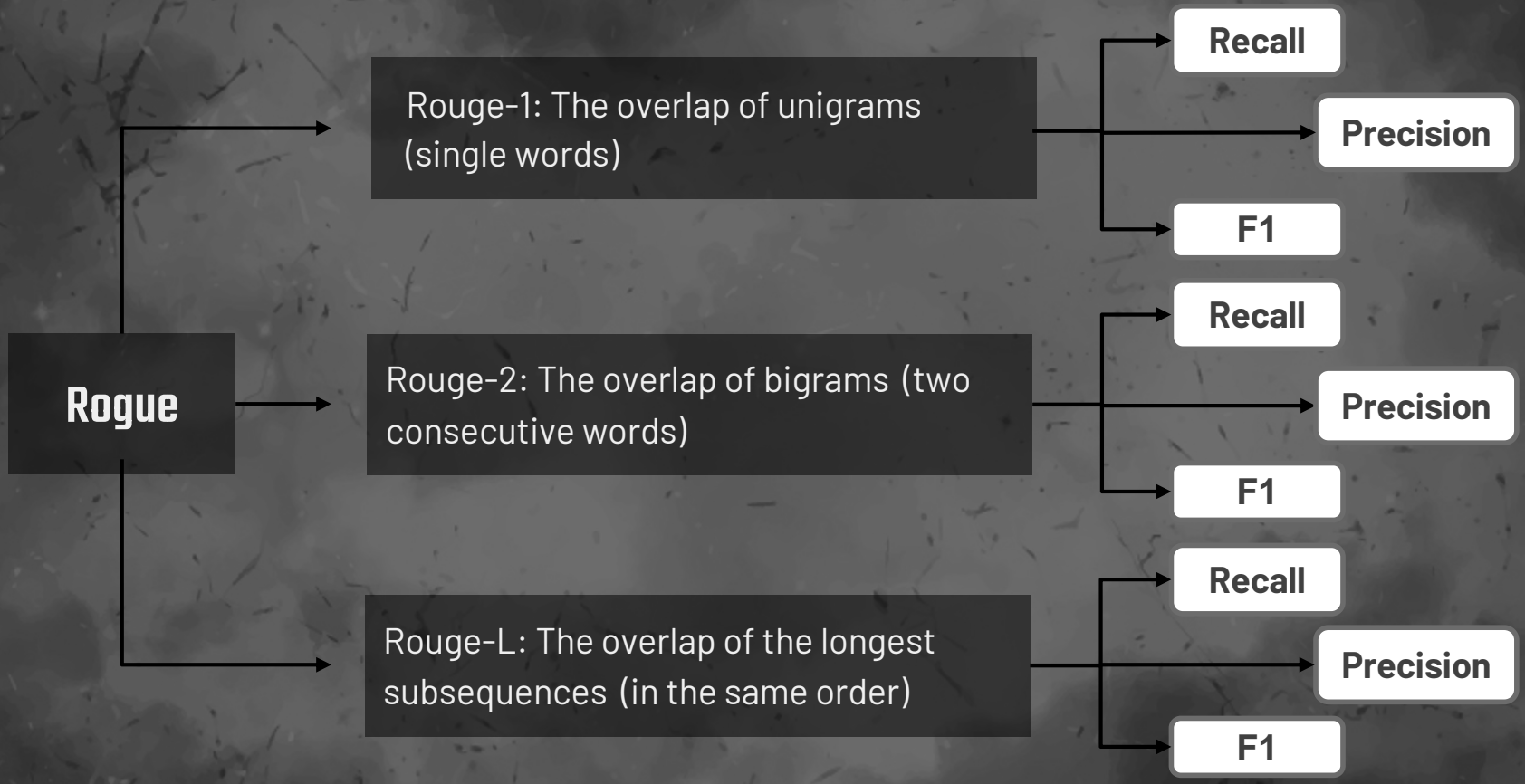
‘number of people standing on top of a dirt road’

’ a man in a **military uniform** is next to a man in a **military uniform**

‘men riding on the back of a motorcycle

’ a man in a **military uniform** with a man in a **military uniform’**

Evaluation


Rogue
Rouge-1: The overlap of unigrams (single words)
Rouge-2: The overlap of bigrams (two consecutive words)
Rouge-L: The overlap of the longest subsequences (in the same order)
Recall
Precision
F1
Recall
Precision
F1
Recall
Precision
F1

Performance evaluation

	Baseline Model	Fine-tuned model
Overall Performance	Lower overall scores	Higher overall scores
Recall vs. Precision	Lower precision	Higher precision
ROUGE-2 Scores	Often zero	Consistently higher
Long-Form Content (ROUGE-L)	Lower ROUGE-L scores	Higher ROUGE-L scores
Consistency Across Metrics	Fluctuating scores	Consistent scores
Overfitting	N/A	All scores are much higher on train set than test

Conclusions

Improved performance

Rogue-score-based evaluation suggests an increased performance across all metrics

Inconsistent results

Both models captured some objects and actions correctly, sometimes did not understand or misunderstood the context and sometimes were completely wrong

Overfitting

Higher rogue scores for train set suggest that the fine-tuned model is overfitting

Human-time consuming

Generating frame-by-frame descriptions and analyzing them separately is time-consuming

Limitations

Computational power:
8 GB RAM

Fine-tuning:
2 images and descriptions

Sampling:
1 video – 5 frames

Evaluation:
Human judgement and Rogue scores

References

Wikipedia,
https://cetas.turing.ac.uk/publications/artificial-intelligence-osint-and-russias-information-landscape,
https://www.analyticsvidhya.com/blog/2023/06/vision-transformers/,
https://reutersinstitute.politics.ox.ac.uk/news/15-tips-investigating-war-crimes-ukraine-and-beyond,
https://trialinternational.org/topics-post/war-crimes/,
https://osintukraine.com/en/archive,
https://www.un.org/en/genocideprevention/war-crimes.shtml,\https://trialinternational.org/topics-post/war-crimes/,
https://journals.sagepub.com/doi/full/10.1177/00027642221144848

Thanks!

Do you have any questions?