Sofia Katsaki, Christos Aivazidis, Nikolaos Mylonas, Ioannis Mollas, Grigorios Tsoumakas

On the Adaptability of Attention-Based Interpretability in Different Transformer Architectures for Multi-Class Classification Tasks

Aristotle University of Thessaloniki

Transformers Interpretability

Model Agnostic

- LIME
- SHAP

Neural Specific

- Integrated Gradients (IG)
- Layer-wise Relevance Propagation (LRP)

Transformer Specific

- Attention Scores
- LRP for Transformers
- Attention Rollout – Attention Flow
- BertViz (Visualization)

Interpretability Evaluation

Ground Truth / Rationale-based

- Human-annotated

AIMLAI	workshop	is	awesome
0	0	0	1

- Compared with feature importance interpretations usually with metrics like AUPRC, F1-token
- May contain bias and noise

Faithfulness-based

- Emulates human user by removing/altering the elements of the input
- Known metrics:
  - Faithfulness
  - Truthfulness
  - Faithfulness Violation Test

Optimus Prime

Attention Scores

- Self-attention layer receives a S × E matrix
  - S: sequence length,
  - E: embedding size.
- Three linear layers produce Q, K, V of S x E dimensions from the input matrix
- Dot product of Q and K is calculated, and divided by the square root of the embedding size
- The attention mask is added
- Those operations result in a matrix of dimensions S × S which contains both negative and positive values, namely the attention scores
- Attention scores are normalized using softmax function

	[CLS]	You	Need	Attention	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
You	0.06	0.23	0.30	0.39	0.01
Need	0.05	0.08	0.68	0.18	0.00
Attention	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30

Example: Attention Matrix

$$A = softmax\left(\frac{Q \cdot K^T}{\sqrt{E}} + mask\right)$$

Optimus Prime

Interpretation Extraction

	[CLS]	You	Need	Attention	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
You	0.06	0.23	0.30	0.39	0.01
Need	0.05	0.08	0.68	0.18	0.00
Attention	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30
	0.17	**0.14**	**0.32**	**0.35**	0.01

From [CLS]

	[CLS]	You	Need	Attention	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
You	0.06	0.23	0.30	0.39	0.01
Need	0.05	0.08	0.68	0.18	0.00
Attention	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30
	0.17	**0.06**	**0.05**	**0.08**	0.18

To [CLS]

	[CLS]	You	Need	Attention	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
You	0.06	0.23	0.30	0.39	0.01
Need	0.05	0.08	0.68	0.18	0.00
Attention	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30
	0.11	**0.14**	**0.33**	**0.36**	0.07

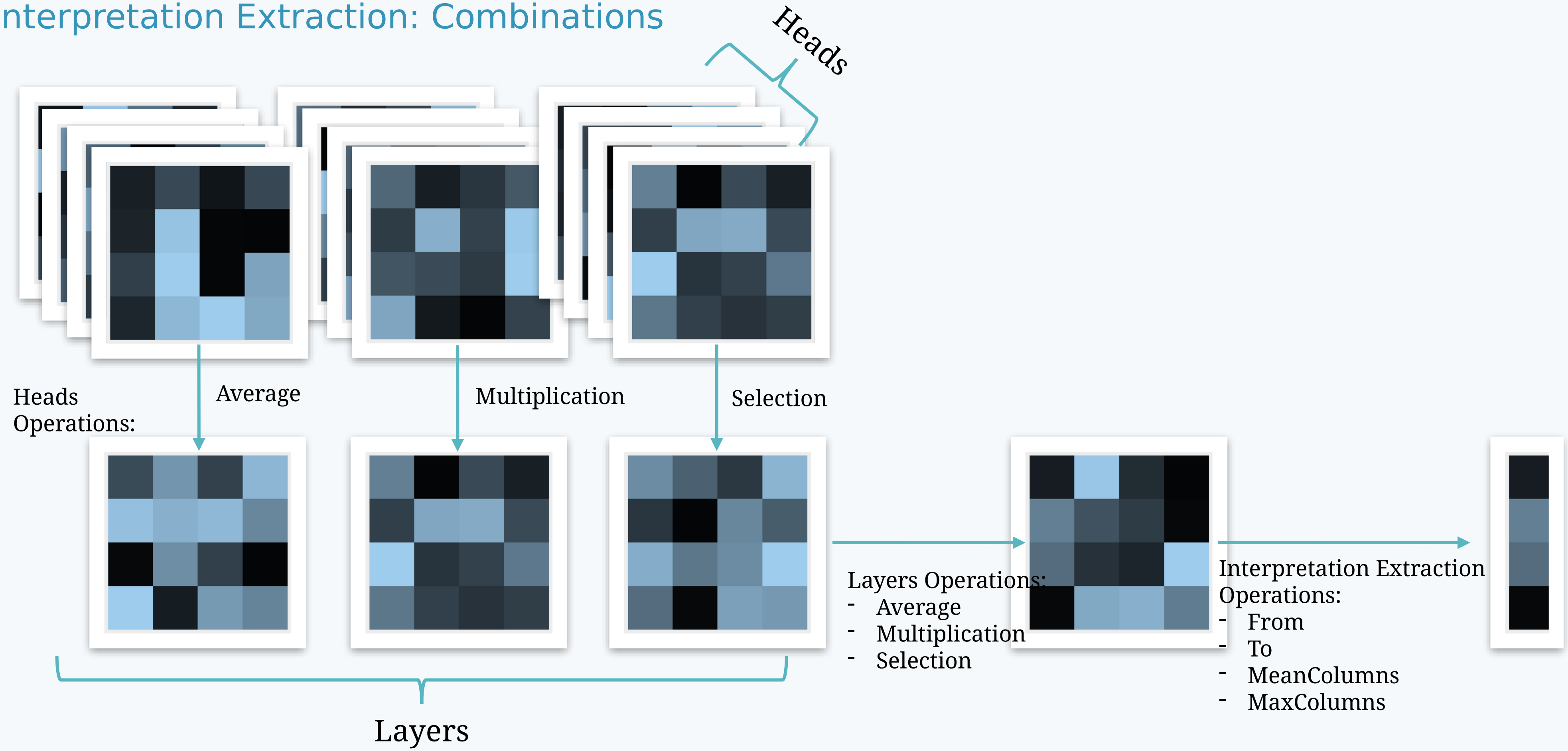
Mean Columns

	[CLS]	You	Need	Attention	[SEP]
[CLS]	0.17	0.14	0.32	0.35	0.01
You	0.06	0.23	0.30	0.39	0.01
Need	0.05	0.08	0.68	0.18	0.00
Attention	0.08	0.07	0.15	0.69	0.01
[SEP]	0.18	0.17	0.19	0.17	0.30
	0.18	**0.23**	**0.68**	**0.69**	0.30

Max Columns

Optimus Prime

Interpretation Extraction: Combinations



Optimus Prime

Selecting most Faithful Interpretation

$$RFT(x,z)=\frac{1}{S}\sum_{i=1}^{S}\frac{u(x,z,i)}{r(t_i)}$$

$$u(x,z,i)=\begin{cases} f_p(x)-f_p(x^{-1}), If\ w_i>0 \\ f_p(x^{(-1)})-f_p(x), If\ w_i<0 \\ -\left|f_p(x)-f_p(x^{-1})\right|, If\ w_i=0 \end{cases}$$

Select a Faithfulness-based metric (such as Ranked Faithful Truthfulness)

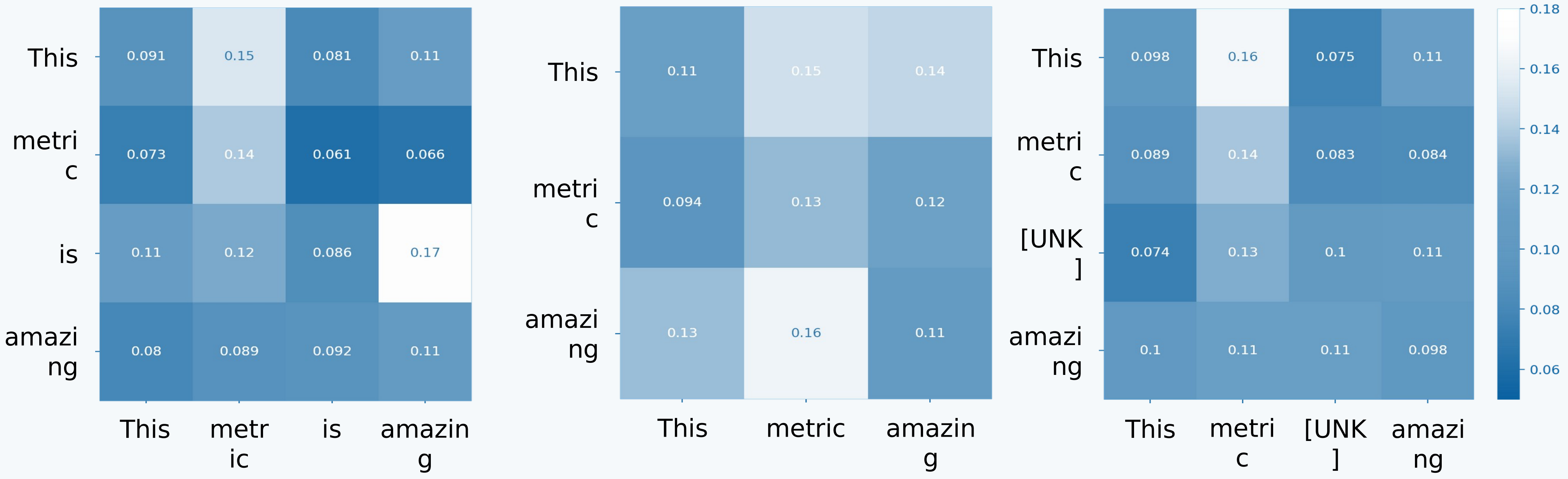
Among the calculated operations, choose the most faithful one

Two variations:

- Optimus Class: best per class
- Optimus Batch: choose the combination that performs better in a validation set

Optimus Prime

Token Replacement by [UNK]



Optimus Class

ORIGINAL

- Applicable in Binary or Multi-Label tasks through Optimus Prime and Optimus Label
- Applicable in BERT & DistilBERT
- Non-optimized runtime

EXTENSION

- Applicable in Multi-Class tasks through Optimus Class
- Applicable in BERT, DistilBERT, RoBERTa & AlBERT
- Optimized runtime on inference

Optimus Class

Multi-Class Adaptation

Diverse Application:

- Optimus extended from binary to multi-class tasks
- Introduction of Optimus Class (OC) technique

Multi-Class Adaptation:

- RFT metric adjusted for multi-class scenarios
- OC finds optimal attention setup for each class

Optimus Class

RoBERTa & AlBERT

Key Properties for Compatibility:
- Sequence classification capability
- Encoder-based architecture
- Accessible attention matrices per head and layer

Pooling Strategy Restriction:
- Only models using [CLS] token embedding or averaging token embeddings considered
- Excluded models with different pooling strategies like GPT-2

Selection of Compatible Transformers:
- Explored new Transformers: RoBERTa and ALBERT
- Models fulfilling criteria: sequence classification, encoder-based, accessible attention matrices

Adaptations for Consistency:
- Modified Optimus for RoBERTa and ALBERT compatibility
- Replaced [UNK] token with <unk> for RoBERTa
- Adjusted tokenizer in Optimus to match models' tokenizers

Optimus Class

Optimization Actions

Time Response Improvement:

- Initial Optimus implementation had slow token-level interpretation times
- Issue stemmed from continuous model queries during attention setup search

Twin Model Approach:

- Two models introduced for efficiency enhancement. One model generates attention matrices, while the other handles predictions
- Twin model setup accelerates Optimus by obtaining necessary predictions faster

Performance Enhancements:

- Additional implementation improvements incorporated
- Focus on optimizing efficiency, speed, and overall functionality
- Resulted in enhanced runtime performance for Optimus

Experiments - Setup

Datasets

HateXplain
- Token-Level Rationales
- Multi-Class (3 Classes)
- Hate Speech Domain

ESNLI
- Token-level Rationales
- Multi-Class (3 Classes)
- Natural Language Understanding Domain

Experiments

1. Experiments on RoBERTa& AlBERT
   Also BERT and DistilBERT
2. Comparison of Optimus Class with Integrated Gradients & Baseline Attention
3. Time-response analysis

Experiments

Comparison of Optimus with other Techniques based on RFT

Dataset/ Model	IG	B	OB	OC
ESNLI (BERT)	0.456	0.488	0.615	**0.876**
ESNLI (DistilBERT)	0.385	0.481	0.552	**0.706**
ESNLI (RoBERTa)	0.442	0.266	0.597	**0.876**
ESNLI (ALBERT)	0.259	0.612	0.664	**0.863**
HX (BERT)	**0.476**	0.337	0.371	0.458
HX (DistilBERT)	**0.467**	0.357	0.379	0.455
HX (RoBERTa)	0.35	0.35	0.355	**0.422**
HX (ALBERT)	0.314	0.408	0.433	**0.562**

Experiments

Comparison of Optimus with other Techniques based on AUPRC

Dataset/ Model	IG	B	OB	OC
ESNLI (BERT)	0.29	0.514	**0.614**	0.433
ESNLI (DistilBERT)	0.301	0.576	**0.651**	0.498
ESNLI (RoBERTa)	0.316	0.274	**0.593**	0.408
ESNLI (ALBERT)	0.337	0.602	**0.604**	0.438
HX (BERT)	0.508	0.488	**0.541**	0.5
HX (DistilBERT)	0.481	0.498	**0.531**	0.506
HX (RoBERTa)	0.477	0.499	**0.514**	0.489
HX (ALBERT)	**0.464**	0.408	0.422	0.413

Experiments

Computational overhead analysis

	ESNLI				HX			
	BERT	DistilBERT	RoBERTa	ALBERT	BERT	DistilBERT	RoBERTa	ALBERT
IG	0.75	0.5	0.75	0.88	0.85	0.51	0.75	0.83
OC	2.7	1.67	3.17	3.08	3.08	1.75	3.42	3.51

Average time response (seconds) of the examined techniques across different models and datasets

20 %

Reduced runtime compared to the original

Conclusions

Attention can be used as interpretation with appropriate processing

Optimus can be adapted to multiclass classification and for various transformer models

Optimus Class outperforms the competitors

Speed can be greatly improved through twin-models technique

Future Work

01
Different Transformer Models (e.g. encoder-decoder based)

02
Different Down-stream Task (e.g. Token Classification)

03
Faster Runtime

04
Other Datasets and metrics

Thank you!

Sofia Katsaki, Christos Aivazidis, Nikolaos Mylonas, Ioannis Mollas, Grigorios Tsoumakas
Aristotle University of Thessaloniki

On the Adaptability of Attention-Based Interpretability in Different Transformer Architectures for Multi-Class Classification Tasks

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant” (Project Number: 514).