An Analysis of two Post-Hoc Interpretability Methods in light of Faithfulness

Raziye Sari

28.02.2023

University of Heidelberg
Institute for Computational Linguistics
BA-Thesis Presentation

Table of Contents

Introduction

Motivation

Dr. Rajiv Raman (Retina Surgeon at Sankara Nethralaya)

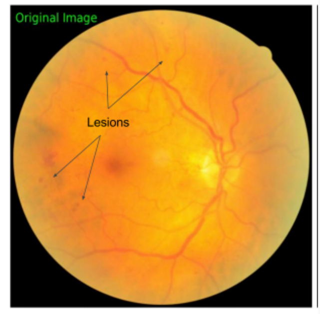


Figure 1: Retinal fundus image[1]

[1] Axiomatic Attribution for Deep Networks, Sundarajan et al., 2017

Motivation

Dr. Rajiv Raman (Retina Surgeon at Sankara Nethralaya)

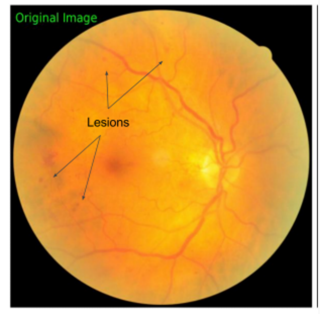


Figure 1: Retinal fundus image[1]

Automated Retinal Disease Assessment (ARDA)

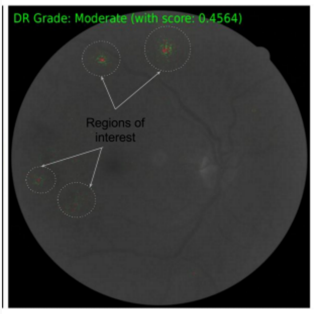


Figure 2: ARDA attributions for prediction[1]

[1] Axiomatic Attribution for Deep Networks, Sundarajan et al., 2017

Motivation

Explainable AI (XAI): **Why** did the model produce this output ?

The problem "Black box" neural systems aren't easily interpretable

Motivation

Explainable AI (XAI): **Why** did the model produce this output ?

The problem "Black box" neural systems aren't easily interpretable

Need for Diverse interpretability methods

Motivation

Danilevsky et al.[1] differentiate between explanations:

Global vs. Local Explaining the model as a whole predictor **vs.** single predictions

[1] A Survey of the State of Explainable AI for Natural Language Processing, 2020

Motivation

Danilevsky et al.[1] differentiate between explanations:

Global vs. Local Explaining the model as a whole predictor **vs.** single predictions

Self-explaining vs. Post-hoc Using the model as an explainer for itself **vs.** Utilizing additional methods

[1] A Survey of the State of Explainable AI for Natural Language Processing, 2020

Post-hoc Interpretability Methods

Integrated Gradients

Introduced[1] as a follow-up on previous method:

Gradients Given a baseline (all-zero vector) and the input feature, calculate the gradients of the feature vector at the given input.

[1] Sundarajan et al., 2017

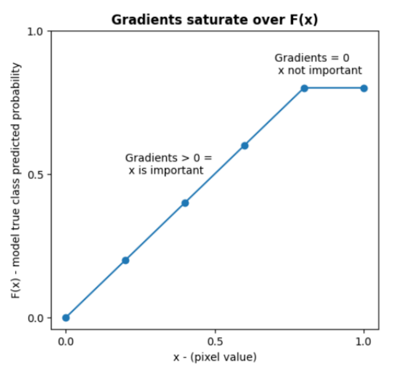
Integrated Gradients

Introduced[1] as a follow-up on previous method:

Gradients Given a baseline (all-zero vector) and the input feature, calculate the gradients of the feature vector at the given input.

$$M(x) = 1 - max(0, 1 - x)$$
$$G(M, x) = max(0, sign(1 - x)) \cdot x$$

$$G(M, 0) = 1 \cdot 0 = 0$$
$$G(M, 2) = 0 \cdot 2 = 0$$



[1] Sundarajan et al., 2017

Integrated Gradients

Integrated Gradients Given a baseline and the input feature, accumulate the gradients along steps of the straight-line-path.

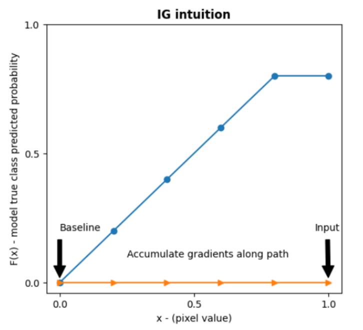
$$IG_i(M, x, x') = (x_i - x_i') \cdot \sum_{k=1}^{m} \frac{\partial M(x' + \frac{k}{m} \cdot (x - x'))}{\partial x_i} \cdot \frac{1}{m}$$

[1] https://www.tensorflow.org/tutorials/interpretability/integrated_gradients

Integrated Gradients

Integrated Gradients Given a baseline and the input feature, accumulate the gradients along steps of the straight-line-path.

$$IG_i(M, x, x') = (x_i - x_i') \cdot \sum_{k=1}^{m} \frac{\partial M(x' + \frac{k}{m} \cdot (x - x'))}{\partial x_i} \cdot \frac{1}{m}$$



Integrated Gradients Intutition[1]

$$(2-0) \cdot \sum \begin{pmatrix} max(0, sign(1-0.0) \\ max(0, sign(1-0.2) \\ max(0, sign(1-0.4) \\ \vdots \\ max(0, sign(1-2.0) \end{pmatrix} \cdot \frac{1}{m}$$

$$IG(M, 2, 0) \approx 1$$

[1] https://www.tensorflow.org/tutorials/interpretability/integrated_gradients

SHAP

Introduced by Lunderg et al.[1] as SHapley Additive exPlanations and based on:

Shapley values Given a coalition c of members $m \in c$ that produce final value v. **How much** did each m contribute to v ?

[1] A Unified Approach to Interpreting Model Predictions, 2017

SHAP

Introduced by Lunderg et al.[1] as SHapley Additive exPlanations and based on:

Shapley values Given a coalition c of members $m \in c$ that produce final value v. **How much** did each m contribute to v ?

1. Sample all coalition pairs $c_{i_m}, c_{j_{\setminus m}} \quad \forall i, j \in C$ such that only member of interest is missing

[1] A Unified Approach to Interpreting Model Predictions, 2017

SHAP

Introduced by Lunderg et al.[1] as SHapley Additive exPlanations and based on:

Shapley values Given a coalition c of members $m \in c$ that produce final value v. **How much** did each m contribute to v ?

1. Sample all coalition pairs $c_{i_m}, c_{j_{\setminus m}} \quad \forall i, j \in C$ such that only member of interest is missing
2. Calculate all marginal contributions of m as $v_i - v_j$ $\forall i, j \in V(C)$

[1] A Unified Approach to Interpreting Model Predictions, 2017

SHAP

Introduced by Lunderg et al.[1] as SHapley Additive exPlanations and based on:

Shapley values Given a coalition c of members $m \in c$ that produce final value v. **How much** did each m contribute to v ?

1. Sample all coalition pairs $c_{i_m}, c_{j_{\setminus m}} \quad \forall i, j \in C$ such that only member of interest is missing
2. Calculate all marginal contributions of m as $v_i - v_j$ $\forall i, j \in V(C)$
3. Average over all marginal contributions of m is the final Shapley value

[1] A Unified Approach to Interpreting Model Predictions, 2017

SHapley Additive exPlanations

Reformulates Shapley values as a linear regression problem:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

where $z' \in \{0, 1\}^M$ and $\phi_j \in \mathbb{R}$ and $M =$ maximum coalition size

SHapley Additive exPlanations

Reformulates Shapley values as a linear regression problem:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

where $z' \in \{0,1\}^M$ and $\phi_j \in \mathbb{R}$ and $M =$ maximum coalition size

Problem 4 features $\rightarrow$ 64 coalitions
32 features $\rightarrow$ 17.1B coalitions

Solution SHAP Kernel: Approximates Shapley values through *permutated* samples and *weighted* linear regression

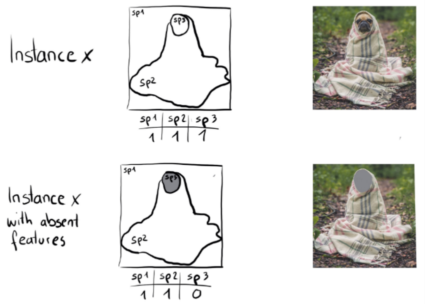
SHAP Kernel

1. For a given datapoint z: Sample coalitions of type: $z' \in \{0, 1\}^M$ and permutate 0's from Background data B
2. Take average of model output y over all synthetic datapoints of z as: $\bar{y} = \mathbb{E}[y_{f_1, f_2, f_i, .., f_M}] \ \forall i \in B$

[1] https://christophm.github.io/interpretable-ml-book/shap.html

SHAP Kernel

1. For a given datapoint z: Sample coalitions of type: $z' \in \{0,1\}^M$ and permutate 0's from Background data B
2. Take average of model output y over all synthetic datapoints of z as: $\bar{y} = \mathbb{E}[y_{f_1,f_2,f_i,..,f_M}]\ \forall i \in B$

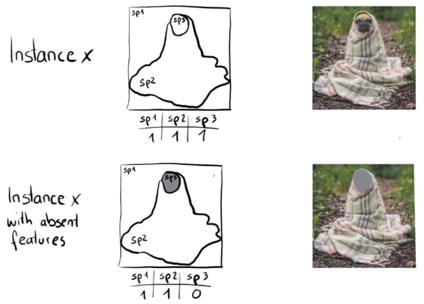


Shapley value estimation[1]

[1] https://christophm.github.io/interpretable-ml-book/shap.html

SHAP Kernel

1. For a given datapoint z: Sample coalitions of type: $z' \in \{0, 1\}^M$ and permutate 0's from Background data B
2. Take average of model output y over all synthetic datapoints of z as: $\bar{y} = \mathbb{E}[y_{f_1, f_2, f_i, .., f_M}] \ \forall i \in B$



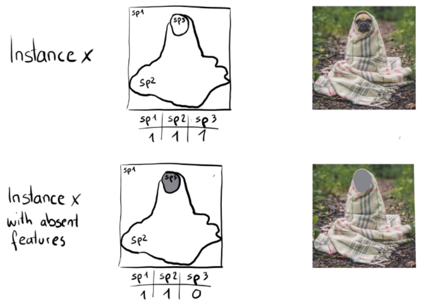
Shapley value estimation[1]

- Coalitions z': [1,1,0], [1,0,0], [0,1,0] ...

[1] https://christophm.github.io/interpretable-ml-book/shap.html

SHAP Kernel

1. For a given datapoint z: Sample coalitions of type: $z' \in \{0,1\}^M$ and permutate 0's from Background data B
2. Take average of model output y over all synthetic datapoints of z as: $\bar{y} = \mathbb{E}[y_{f_1,f_2,f_i,..,f_M}]\ \forall i \in B$



Shapley value estimation[1]

- Coalitions z': [1,1,0], [1,0,0], [0,1,0] ...
- Weigh each z':

$$\pi_z(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)}$$

[1] https://christophm.github.io/interpretable-ml-book/shap.html

Approach

Setup

Task Sequence classification

- Input tokens as features

Data CardioDE corpus from Dieterich Lab (Heidelberg)

- Unit of 500 doctoral letters from the cardiology department
- Each section of the letter belongs to one of 11 labels: *Anrede, Diagnosen, AllergienUnverträglichkeitenRisiken, Anamnese, Medikation, KUBefunde, Befunde, EchoBefunde, Zusammenfassung, Mix, Abschluss*

Details

Model BertForSequenceClassification[1]

Deployment Fine-tuned on 400 letters

- Split into 90% train & 10% development set
- Trained for 2 epochs

Results Tested on 100 held-out letters

- Overall accuracy: 93%
- Best performing labels: *Anrede, Medikation, KUBefunde* (98% and above)
- Worst performing labels: *Mix, EchoBefunde* (79% and below)

[1] https://huggingface.co/bert-base-german-cased

Experiments

Explanations

ferret[1] unifies state-of-the-art local post-hoc interpretability methods under a benchmarking suite.

Feature importance attributions for:

SHAP are the coefficients of weighted linear regression model.

IG are the accumulated gradients along straight-line-path.

[1] https://github.com/g8a9/ferret

ferret[1] unifies state-of-the-art local post-hoc interpretability methods under a benchmarking suite.

Feature importance attributions for:

SHAP are the coefficients of weighted linear regression model.

IG are the accumulated gradients along straight-line-path.

	Per	##ip	##here	Ö	##dem	##e	:	moder	##ate	US	-	Ö.1	##dem.1	##e.1
Partition SHAP	0.14	0.02	-0.03	0.13	0.03	-0.04	0.19	0.11	0.01	-0.03	-0.01	-0.00	0.04	0.22
LIME	0.19	0.12	0.03	0.10	0.11	0.02	0.15	0.03	-0.00	-0.03	0.01	0.05	0.10	0.06
Gradient	0.05	0.07	0.08	0.07	0.06	0.05	0.13	0.10	0.07	0.09	0.03	0.05	0.04	0.03
Gradient (x Input)	-0.03	0.02	-0.03	-0.08	0.04	0.12	-0.10	-0.10	-0.15	0.04	0.01	0.00	0.14	0.06
Integrated Gradient	-0.01	-0.01	-0.01	0.25	-0.16	-0.11	0.07	0.13	0.01	0.03	0.02	0.03	-0.07	-0.02
Integrated Gradient (x Input)	0.08	0.06	0.06	0.12	0.14	0.13	0.14	0.06	0.02	0.01	-0.00	0.05	0.06	0.07

Figure 3: *ferret* explanations for a sample sentence from *KUBefunde*

[1] https://github.com/g8a9/ferret

Evaluating interpretability methods

Faithfulness

"Measures how accurate the explanation reflects the inner-workings of the model"[0]

[0] Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?, Jacovi and Goldberg, 2020

Faithfulness

"Measures how accurate the explanation reflects the inner-workings of the model"[0]

Selects most important tokens (r) per explanation and measures:

1. **Comprehensiveness** $f(x)_j - f(x \setminus r_j)_j$
2. **Sufficiency** $f(x)_j - f(r_j)_j$

[0] Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?, Jacovi and Goldberg, 2020

Faithfulness

"Measures how accurate the explanation reflects the inner-workings of the model"[0]

Selects most important tokens (r) per explanation and measures:

1. **Comprehensiveness** $f(x)_j - f(x \setminus r_j)_j$
2. **Sufficiency** $f(x)_j - f(r_j)_j$

$\rightarrow$ Records change in prediction once sentence omits[1]/only keeps[2] tokens in r

[0] Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?, Jacovi and Goldberg, 2020

Faithfulness

Process of measuring Faithfulness:

1. Filter out tokens with negative contribution to prediction

Faithfulness

Process of measuring Faithfulness:

1. Filter out tokens with negative contribution to prediction
2. With steps of 10: Choose k % highest contributing tokens

Faithfulness

Process of measuring Faithfulness:

1. Filter out tokens with negative contribution to prediction
2. With steps of 10: Choose k % highest contributing tokens
3. For each step: Calculate Comprehensiveness/Sufficiency score

Faithfulness

Process of measuring Faithfulness:

1. Filter out tokens with negative contribution to prediction
2. With steps of 10: Choose k % highest contributing tokens
3. For each step: Calculate Comprehensiveness/Sufficiency score
4. Finally: Take the mean of the 10 scores

Faithfulness

Process of measuring Faithfulness:

1. Filter out tokens with negative contribution to prediction
2. With steps of 10: Choose k % highest contributing tokens
3. For each step: Calculate Comprehensiveness/Sufficiency score
4. Finally: Take the mean of the 10 scores

Specific to *ferret*:

- When omitting tokens $\in r$ from the sentence, they prefer deleting (instead of masking out)

Comprehensiveness

SHAP's best-scoring label: $Anrede = j$

"**über Ihren Patienten** B-SA**L**UTE B-PER **I**-PER geboren am ⟨[Pseudo] 24/06/1977⟩ **wohn**haft in **B**-PLZ B-LOC **I**-ADDR **I**-ADDR der sich vom bis in **unserer** stationären Behandlung **befand**."

Comprehensiveness

SHAP's best-scoring label: $Anrede = j$

"**über Ihren Patienten** B-SA**L**UTE B-PER **I**-PER geboren am ⟨[Pseudo] 24/06/1977⟩ **wohn**haft in **B**-PLZ B-LOC **I**-ADDR **I**-ADDR der sich vom bis in **unserer** stationären Behandlung **befand**."

$$Compr = f(x)_j - \boldsymbol{f(x \setminus r_j)_j} = 1.0$$

Comprehensiveness

SHAP's best-scoring label: $Anrede = j$

"**über Ihren Patienten** B-SA**L**UTE B-PER **I**-PER geboren am ⟨[Pseudo] 24/06/1977⟩ **wohn**haft in **B**-PLZ B-LOC **I**-ADDR **I**-ADDR der sich vom bis in **unserer** stationären Behandlung **befand**."

$$Compr = f(x)_j - \boldsymbol{f(x \setminus r_j)_j} = 1.0$$

$$Compr^{\textbf{SHAP}}_{k=10} = 1.0 - \mathbf{0.02} = \underline{0.98}$$

where $r_k = [\texttt{über}_{0.27}, \texttt{Ihren}_{0.16}, \texttt{Patienten}_{0.09}, \texttt{wohn}_{0.03}, \texttt{unserer}_{0.04}, \texttt{befand}_{0.04}]$

Comprehensiveness

SHAP's best-scoring label: $Anrede = j$

"**über Ihren Patienten** B-SA**L**UTE B-PER **I**-PER geboren am ⟨[Pseudo] 24/06/1977⟩ **wohn**haft in **B**-PLZ B-LOC **I**-ADDR **I**-ADDR der sich vom bis in **unserer** stationären Behandlung **befand**."

$$Compr = f(x)_j - \boldsymbol{f(x \setminus r_j)_j} = 1.0$$

$$Compr^{\textbf{SHAP}}_{k=10} = 1.0 - \mathbf{0.02} = \underline{0.98}$$

where $r_k = [\texttt{über}_{0.27}, \texttt{Ihren}_{0.16}, \texttt{Patienten}_{0.09}, \texttt{wohn}_{0.03}, \texttt{unserer}_{0.04}, \texttt{befand}_{0.04}]$

$$Compr^{\textbf{IG}}_{k=10} = 1.0 - \mathbf{0.82} = 0.18$$

where $r_k = [\texttt{über}_{0.37}, \texttt{Ihren}_{0.04}]$

Comprehensiveness - Results

Mean Comprehensiveness scores over 10 samples per label:

Label	Mean Scores SHAP	Mean Scores IG	F1-Score
Anrede	**1.0**	0.4	**1.0**
Mix	0.86	**0.64**	0.79
AllergienUnverträglichkeitenRisiken	0.85	0.31	0.96
KUBefunde	0.77	0.3	0.98
Diagnosen	0.75	0.31	0.96
Zusammenfassung	0.75	0.14	0.9
Befunde	0.67	0.2	0.9
EchoBefunde	0.66	0.25	0.73
Anamnese	0.64	0.02	0.85
Medikation	0.64	0.13	0.98
Abschluss	0.61	0.3	0.96

Table 1: Comprehensiveness mean scores for SHAP & IG with F1-scores per label

Comprehensiveness

SHAP's worst-scoring label: $Abschluss = j$

"**I-P**ER / **I.**"

$$Compr = f(x)_j - f(x \setminus r_j)_j = 0.86$$

Comprehensiveness

SHAP's worst-scoring label: $\mathit{Abschluss} = j$

"**I**-**P**ER / **I.**"

$$\mathit{Compr} = f(x)_j - f(x \setminus r_j)_j = 0.86$$

$\mathit{Compr}^{\textbf{SHAP}}_{k=20} = 1.0 - \mathbf{0.7} = 0.3$ where $r_k = [\texttt{P}]$

$\mathit{Compr}^{\textbf{SHAP}}_{k=30} = 1.0 - \mathbf{0.0} = \underline{1.0}$ where $r_k = [\texttt{P, I}]$

$\vdots$

$\mathit{Compr}^{\textbf{SHAP}}_{k=100} = 1.0 - \mathbf{0.0} = \underline{1.0}$ where $r_k = [\texttt{I}_{0.26}\texttt{, -}_{0.03}\texttt{, P}_{0.44}\texttt{, I}_{0.07}\texttt{, .}_{0.04}]$

Comprehensiveness

SHAP's worst-scoring label: $Abschluss = j$

"**I**-**P**ER / **I.**"

$$Compr = f(x)_j - f(x \setminus r_j)_j = 0.86$$

$Compr^{\textbf{SHAP}}_{k=20} = 1.0 - \mathbf{0.7} = 0.3$ where $r_k = $ `[P]`

$Compr^{\textbf{SHAP}}_{k=30} = 1.0 - \mathbf{0.0} = \underline{1.0}$ where $r_k = $ `[P, I]`

$\vdots$

$Compr^{\textbf{SHAP}}_{k=100} = 1.0 - \mathbf{0.0} = \underline{1.0}$ where $r_k = [\mathtt{I}_{0.26}, \mathtt{-}_{0.03}, \mathtt{P}_{0.44}, \mathtt{I}_{0.07}, \cdot_{0.04}]$

Important note:

1. *ferret* leaves out scores that do not contain any changes to prior state of r, e.g.: $k = 10, 40, 50, 80, 90$

Comprehensiveness

"I-**PER** / **I**."

$$Compr = f(x)_j - f(x \setminus r_j)_j = 0.06$$

$Compr^{\mathbf{IG}}_{k=20} = 1.0 - \mathbf{0.99} = 0.01$ where $r_k =$ [ER]

Comprehensiveness

"I-**PER** / **I**."

$$Compr = f(x)_j - f(x \setminus r_j)_j = 0.06$$

$Compr^{\textbf{IG}}_{k=20} = 1.0 - \mathbf{0.99} = 0.01$ where $r_k = $ [ER]

$\vdots$

$Compr^{\textbf{IG}}_{k=100} = 1.0 - \mathbf{0.88} = 0.12$ where $r_k = [\mathrm{P}_{0.04},\ \mathrm{ER}_{0.52},\ \mathrm{I}_{0.09}]$

Comprehensiveness

"I-**PER** / **I**."

$$Compr = f(x)_j - f(x \setminus r_j)_j = 0.06$$

$$Compr^{\textbf{IG}}_{k=20} = 1.0 - \mathbf{0.99} = 0.01 \text{ where } r_k = [\texttt{ER}]$$

$\vdots$

$$Compr^{\textbf{IG}}_{k=100} = 1.0 - \mathbf{0.88} = 0.12 \text{ where } r_k = [\texttt{P}_{0.04}, \ \texttt{ER}_{0.52}, \ \texttt{I}_{0.09}]$$

While **SHAP** ascribed negative attribution to ER, **IG** quite contrarily marks it as most important. Coincidence ?

Interim Conclusion

We saw throughout, that ...

- SHAP's choice of tokens effected sentence score more than IG's choice.
    - What about same size r of tokens ?

Interim Conclusion

We saw throughout, that ...

- SHAP's choice of tokens effected sentence score more than IG's choice.
  - What about same size r of tokens ?
- IG may attribute highest importance to a token, which received negative attribution with SHAP.
  - Also in Sufficiency the case ?

Sufficiency

IG's best-scoring label: $EchoBefunde = j$

"**Untersuchung am Bett auf** Kardio-Intensiv. Vorbekannt deutlich reduzierte **Schallbedingungen**, v.a. **von** parasternal."

$$Suff = f(x)_j - \boldsymbol{f(r_j)_j} = 0.14_{SHAP} | 0.34_{IG}$$

Sufficiency

IG's best-scoring label: $EchoBefunde = j$

"**Untersuchung am Bett auf** Kardio-Intensiv. Vorbekannt deutlich reduzierte **Schallbedingungen**, v.a. **von** parasternal."

$$Suff = f(x)_j - \boldsymbol{f(r_j)_j} = 0.14_{SHAP}|0.34_{IG}$$

$Suff^{\textbf{IG}}_{k=10} = 0.98 - \mathbf{0.05} = 0.93$ where $r_k =$ `[Untersuchung]`

$Suff^{\textbf{SHAP}}_{k=10} = 0.98 - \mathbf{0.25} = \underline{0.73}$ where $r_k =$ `[am, Schall]`

Sufficiency

IG's best-scoring label: $EchoBefunde = j$

"**Untersuchung am Bett auf** Kardio-Intensiv. Vorbekannt deutlich reduzierte **Schallbedingungen**, v.a. **von** parasternal."

$$Suff = f(x)_j - \boldsymbol{f(r_j)_j} = 0.14_{SHAP}|0.34_{IG}$$

$Suff^{\mathbf{IG}}_{k=10} = 0.98 - \mathbf{0.05} = 0.93 \quad \text{where } r_k = [\texttt{Untersuchung}]$

$Suff^{\mathbf{SHAP}}_{k=10} = 0.98 - \mathbf{0.25} = \underline{0.73} \quad \text{where } r_k = [\texttt{am, Schall}]$

What about **equal length $\boldsymbol{r_k}$**?

$Suff^{\mathbf{IG}}_{k=30} = 0.98 - \mathbf{0.57} = 0.41 \quad \text{where } r_k = [\texttt{Untersuchung}_{0.09}\texttt{, Bett}_{0.09}\texttt{, auf}_{0.05}\texttt{, von}_{0.05}]$

Sufficiency

IG's best-scoring label: *EchoBefunde* $= j$

"**Untersuchung am Bett auf** Kardio-Intensiv. Vorbekannt deutlich reduzierte **Schallbedingungen**, v.a. **von** parasternal."

$$Suff = f(x)_j - \boldsymbol{f(r_j)_j} = 0.14_{SHAP}|0.34_{IG}$$

$Suff^{\textbf{IG}}_{k=10} = 0.98 - \mathbf{0.05} = 0.93$ where $r_k =$ `[Untersuchung]`

$Suff^{\textbf{SHAP}}_{k=10} = 0.98 - \mathbf{0.25} = \underline{0.73}$ where $r_k =$ `[am, Schall]`

What about **equal length $\boldsymbol{r_k}$**?

$Suff^{\textbf{IG}}_{k=30} = 0.98 - \mathbf{0.57} = 0.41$ where $r_k = [\texttt{Untersuchung}_{0.09}, \texttt{Bett}_{0.09}, \texttt{auf}_{0.05}, \texttt{von}_{0.05}]$

$Suff^{\textbf{SHAP}}_{k=20} = 0.98 - \mathbf{0.69} = \underline{0.29}$ where $r_k = [\texttt{am}_{0.19}, \texttt{Bett}_{0.17}, \texttt{Schall}_{0.20}, \texttt{bedingungen}_{0.17}]$

Sufficiency - Results

Mean Sufficiency scores over 10 samples per label:

Label	Mean Scores SHAP	Mean Scores IG	F1-Score
Zusammenfassung	**0.0**	0.44	0.9
Befunde	0.02	0.41	0.9
Anamnese	0.03	0.46	0.85
EchoBefunde	0.04	**0.37**	0.73
AllergienUnverträglichkeitenRisiken	0.06	0.6	0.96
Medikation	0.07	0.44	0.98
Anrede	0.1	0.71	**1.0**
Abschluss	0.15	0.45	0.96
Diagnosen	0.19	0.66	0.96
KUBefunde	0.25	0.77	0.98
Mix	0.37	0.8	0.79

Table 2: Sufficiency Mean scores for SHAP & IG with overall F1-scores for each label

Sufficiency

SHAP & IG's worst-scoring label: $Mix = j$

"- **Kost**aufbau nach **Ernährungskon**s**il**"

$$f(x)_j - f(r_j)_j = 0.58_{SHAP} | 0.93_{IG}$$

Sufficiency

SHAP & IG's worst-scoring label: $Mix = j$

"- **Kost**aufbau nach **Ernährungskon**s**il**"

$$f(x)_j - f(r_j)_j = 0.58_{SHAP}|0.93_{IG}$$

$Suff^{\textbf{SHAP}}_{k=60} = 0.94 - \mathbf{0.03} = 0.91 \quad r_k =$ `[Kost, Ernährung, skon, il]`

Sufficiency

SHAP & IG's worst-scoring label: $Mix = j$

"- **Kost**aufbau nach **Ernährungskonsil**"

$$f(x)_j - f(r_j)_j = 0.58_{SHAP} | 0.93_{IG}$$

$Suff_{k=60}^{\textbf{SHAP}} = 0.94 - \mathbf{0.03} = 0.91 \quad r_k = [$`Kost, Ernährung, skon, il`$]$

$Suff_{k=70}^{\textbf{SHAP}} = 0.94 - \mathbf{0.71} = \underline{0.23} \quad r_k = [$`-`$_{0.09}$, `Kost`$_{0.41}$, `Ernährung`$_{0.36}$, `skon`$_{0.05}$, `il`$_{0.1}]$

SHAP & IG's worst-scoring label: $Mix = j$
"- **Kost**aufbau nach **Ernährungskonsil**"

$$f(x)_j - f(r_j)_j = 0.58_{SHAP}|0.93_{IG}$$

$Suff^{\textbf{SHAP}}_{k=60} = 0.94 - \mathbf{0.03} = 0.91 \quad r_k = [\texttt{Kost, Ernährung, skon, il}]$

$Suff^{\textbf{SHAP}}_{k=70} = 0.94 - \mathbf{0.71} = \underline{0.23} \quad r_k = [\texttt{-}_{0.09}, \texttt{Kost}_{0.41}, \texttt{Ernährung}_{0.36}, \texttt{skon}_{0.05}, \texttt{il}_{0.1}]$

$Suff^{\textbf{IG}}_{k=100} = 0.98 - \mathbf{0.05} = 0.93 \quad r_k = [\texttt{Kost}_{0.28}, \texttt{aufbau}_{0.03}, \texttt{Ernährung}_{0.02}, \texttt{skon}_{0.17}, \texttt{s}_{0.2}, \texttt{il}_{0.0}]$

Data Statistics

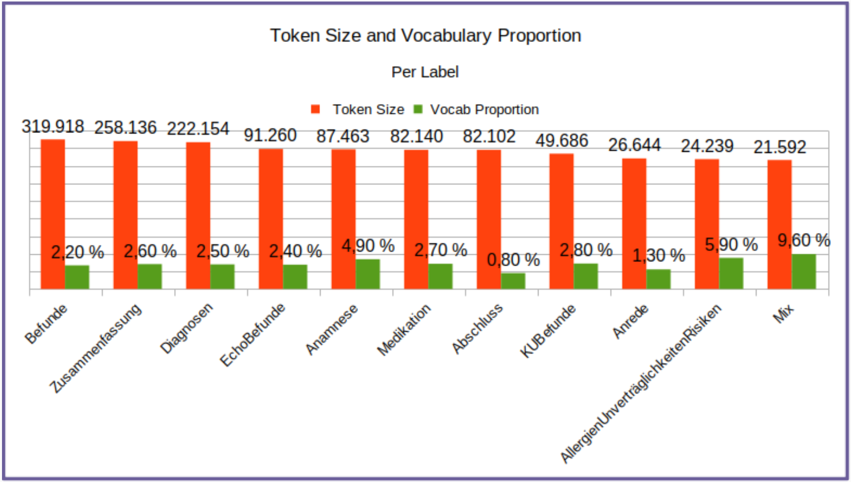


Figure 4: Token Size and Vocabulary Proportion per Label in Training Data

Sufficiency

SHAP's best-scoring label: *Zusammenfassung* $= j$

"Röntgenologisch wurde der V.a. eine Stauungspneumonie gestellt."

$$f(x)_j - f(r_j)_j = \mathbf{0.0}_{SHAP} | 0.99_{IG}$$

Sufficiency

SHAP's best-scoring label: $\mathit{Zusammenfassung} = j$

"Röntgenologisch wurde der V.a. eine Stauungspneumonie gestellt."

$$f(x)_j - f(r_j)_j = \mathbf{0.0}_{SHAP} | 0.99_{IG}$$

$\mathit{Suff}_{k=10}^{\textbf{SHAP}} = 1.0 - \mathbf{1.0} = \underline{0.0} \quad r_k = [\texttt{wurde}_{0.44}]$

Sufficiency

SHAP's best-scoring label: *Zusammenfassung* $= j$

"Röntgenologisch wurde der V.a. eine Stauungspneumonie gestellt."

$$f(x)_j - f(r_j)_j = \mathbf{0.0}_{SHAP} | 0.99_{IG}$$

$Suff^{\textbf{SHAP}}_{k=10} = 1.0 - \mathbf{1.0} = \underline{0.0} \quad r_k = [\texttt{wurde}_{0.44}]$

$Suff^{\textbf{IG}}_{k=10} = 1.0 - \mathbf{0.01} = 0.99 \quad r_k = [\texttt{Röntgen}]$

Sufficiency

SHAP's best-scoring label: *Zusammenfassung* $= j$

"Röntgenologisch wurde der V.a. eine Stauungspneumonie gestellt."

$$f(x)_j - f(r_j)_j = \mathbf{0.0}_{SHAP} | 0.99_{IG}$$

$Suff^{\textbf{SHAP}}_{k=10} = 1.0 - \mathbf{1.0} = \underline{0.0} \quad r_k = [\texttt{wurde}_{0.44}]$

$Suff^{\textbf{IG}}_{k=10} = 1.0 - \mathbf{0.01} = 0.99 \quad r_k = [\texttt{Röntgen}]$

$\vdots$

$Suff^{\textbf{IG}}_{k=100} = 1.0 - \mathbf{0.01} = 0.99 \quad r_k = [\texttt{Röntgen}_{0.13}, \texttt{a}_{0.04}, \texttt{.}_{0.01}, \texttt{Stau}_{0.03}, \texttt{p}_{0.01}, \texttt{ne}_{0.01}, \texttt{onie}_{0.04}]$

Sufficiency - Conclusion

- Comprehensiveness & Sufficiency reflect model preference of specific labels in their overall scoring.
- IG's tendency to disregard *most important* tokens also apparent here.

Conclusion & Lookout

Conclusion & Remarks

- ferret's preference of deleting tokens r over masking them out is questionable.

Conclusion & Remarks

- ferret's preference of deleting tokens r over masking them out is questionable.
- Sufficiency or Comprehensiveness should not be deployed without the other.
  - Contrasting results in Comprehensiveness & Sufficiency are a good sign of model bias.

Conclusion & Remarks

- ferret's preference of deleting tokens r over masking them out is questionable.
- Sufficiency or Comprehensiveness should not be deployed without the other.
  - Contrasting results in Comprehensiveness & Sufficiency are a good sign of model bias.
- ferret's Faithfulness measures the alignment of the explanation with the actual inner-workings of the model (to some degree) well.

Lookout

- Since IG tends to ascribe negative values to seemingly important tokens, find out why. Moreover, analyze if choice of baseline[1] has an impact.
- Experiment with inclusion of negative attribution tokens from IG into r.

[1] https://distill.pub/2020/attribution-baselines/