Learning to Scaffold:
Optimizing Model Explanations for Teaching

Patrick Fernandes*

Marcos Treviso*

Danish Pruthi

André Martins

Graham Neubig

NeurIPS 2022

November 28th

How should we evaluate explanations?

- Explainability methods generally do not correlate with each other
- Most explanations do not help to predict the model’s outputs and/or failures

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

How should we evaluate explanations?

- Explainability methods generally do not correlate with each other
- Most explanations do not help to predict the model's outputs and/or failures

- **Simulability:** "can we recover the model's output based on the explanation?"
    - ✓ aligns with the goal of communicating the underlying model behavior
    - ✓ is easily measurable (both manually and automatically)
    - ✓ puts all explainability methods under a single perspective

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

How should we evaluate explanations?

- Explainability methods generally do not correlate with each other
- Most explanations do not help to predict the model’s outputs and/or failures

- **Simulability:** "can we recover the model’s output based on the explanation?"
  - ✓ aligns with the goal of communicating the underlying model behavior
  - ✓ is easily measurable (both manually and automatically)
  - ✓ puts all explainability methods under a single perspective

- Pruthi et al. (2021) proposed a framework for measuring simulability that disregards **trivial protocols** 🥰

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Simulability

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(\,T(x)\,,\;S_\theta(x)\,)\right]$$

Simulability

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x \sim \mathcal{D}_{\text{train}}} \left[\mathcal{L}_{\text{sim}}(\underbrace{T(x)}_{\textit{teacher}}, \underbrace{S_\theta(x)}_{\textit{student}}) \right]$$

Simulability

cross entropy

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(\underbrace{T(x)}_{\text{teacher}}, \underbrace{S_\theta(x)}_{\text{student}})\right]$$

Simulability

cross entropy

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(\underbrace{T(x)}_{\textit{teacher}}, \underbrace{S_\theta(x)}_{\textit{student}})\right]$$

(test time) $$\mathrm{SIM}(T, S_\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[1\{T(x) = S_\theta(x)\}\right]$$

Simulability

cross entropy

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x \sim \mathcal{D}_{\text{train}}} \left[\mathcal{L}_{\text{sim}}(\underbrace{T(x)}_{\textit{teacher}}, \underbrace{S_\theta(x)}_{\textit{student}}) \right]$$

(test time) $$\text{SIM}(T, S_\theta) = \mathbb{E}_{x \sim \mathcal{D}_{\text{test}}} \left[\underbrace{1\{ T(x) = S_\theta(x) \}}_{\textit{agreement}} \right]$$

Simulability

cross entropy

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(\underbrace{T(x)}_{\textit{teacher}}, \underbrace{S_\theta(x)}_{\textit{student}})\right]$$

(test time) $$\text{SIM}(T, S_\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\underbrace{\mathbb{1}\{T(x) = S_\theta(x)\}}_{\textit{agreement}}\right]$$

Introducing explanations: Teacher and Student explainers $E_T(x)$, $E_S(x)$

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Simulability

cross entropy

(training time) $\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(T(x), S_\theta(x))\right]$

teacher student

(test time) $\text{SIM}(T, S_\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\mathbb{1}\{T(x) = S_\theta(x)\}\right]$

agreement

Introducing explanations: Teacher and Student explainers $E_T(x)$, $E_S(x)$

$$\theta_E^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(T(x), S_\theta(x)) + \beta\mathcal{L}_{\text{expl}}(E_T(x), E_{S_\theta}(x))\right]$$

simulability loss

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Simulability

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}^{\text{cross entropy}}\big(\underbrace{T(x)}_{\text{teacher}},\ \underbrace{S_\theta(x)}_{\text{student}}\big)\right]$$

(test time) $$\text{SIM}(T, S_\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\Big[\underbrace{\mathbb{1}\{T(x) = S_\theta(x)\}}_{\text{agreement}}\Big]$$

Introducing explanations: Teacher and Student explainers $E_T(x)$, $E_S(x)$

$$\theta_E^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\Big[\underbrace{\mathcal{L}_{\text{sim}}\big(T(x),\ S_\theta(x)\big)}_{\text{simulability loss}} + \underbrace{\beta\mathcal{L}_{\text{expl}}\big(E_T(x),\ E_{S_\theta}(x)\big)}_{\text{explainer regularizer (e.g.. KL)}}\Big]$$

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Simulability

cross entropy

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(T(x), S_\theta(x))\right]$$

teacher student

(test time) $$\text{SIM}(T, S_\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\mathbb{1}\{T(x) = S_\theta(x)\}\right]$$

agreement

Introducing explanations: Teacher and Student explainers $E_T(x)$, $E_S(x)$

$$\underbrace{\theta^\star_E = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(T(x), S_\theta(x)) + \beta\mathcal{L}_{\text{expl}}(E_T(x), E_{S_\theta}(x))\right]}$$

simulability loss explainer regularizer (e.g.. KL)

(standard simulability) $\text{SIM}(T, S_{\theta^\star})$

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Simulability

cross entropy

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(\underbrace{T(x)}_{\text{teacher}}, \underbrace{S_\theta(x)}_{\text{student}})\right]$$

(test time) $$\mathrm{SIM}(T, S_\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\underbrace{\mathbb{1}\{T(x) = S_\theta(x)\}}_{\text{agreement}}\right]$$

Introducing explanations: Teacher and Student explainers $E_T(x)$, $E_S(x)$

$$\underbrace{\theta_E^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\underbrace{\mathcal{L}_{\text{sim}}(T(x), S_\theta(x))}_{\text{simulability loss}} + \underbrace{\beta\mathcal{L}_{\text{expl}}(E_T(x), E_{S_\theta}(x))}_{\text{explainer regularizer (e.g.. KL)}}\right]}$$

(standard simulability) $$\mathrm{SIM}(T, S_{\theta^\star}) < \mathrm{SIM}(T, S_{\theta_E^\star})$$ (scaffolded simulability)

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Simulability

cross entropy

(training time) $$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{sim}}(\underbrace{T(x)}_{\text{teacher}}, \underbrace{S_\theta(x)}_{\text{student}})\right]$$

(test time) $$\text{SIM}(T, S_\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\underbrace{1\{T(x) = S_\theta(x)\}}_{\text{agreement}}\right]$$

Introducing explanations: Teacher and Student explainers $E_T(x)$, $E_S(x)$

$$\theta^\star_E = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\underbrace{\mathcal{L}_{\text{sim}}(T(x), S_\theta(x))}_{\text{simulability loss}} + \underbrace{\beta\mathcal{L}_{\text{expl}}(E_T(x), E_{S_\theta}(x))}_{\text{explainer regularizer (e.g.. KL)}}\right]$$

(standard simulability) $\text{SIM}(T, S_{\theta^\star}) < \text{SIM}(T, S_{\theta^\star_E})$ (scaffolded simulability)

Can we **learn explainers** $\phi(E)$ that optimize **simulability**?

(scaffolded simulability) $\text{SIM}(T, S_{\theta^\star_E}) < \text{SIM}(T, S_{\theta_{\phi(E)}})$ (optim. scaffolded simulability)

Optimizing Explainers for Teaching

- **S**caffold-**Ma**ximizing **T**raining (**SMaT**) framework

$$\mathcal{L}_{\text{student}}(x; T, E_T, S_\theta, E_S) = \mathcal{L}_{\text{sim}}(\, T(x)\,,\; S_\theta(x)\,) + \beta\mathcal{L}_{\text{expl}}(\, E_T(x)\,,\; E_{S_\theta}(x))$$

Optimizing Explainers for Teaching

- **S**caffold-**Ma**ximizing **T**raining (**SMaT**) framework

$$\mathcal{L}_{\text{student}}(x; T, E_T, S_\theta, E_S) = \mathcal{L}_{\text{sim}}(\,T(x)\,,\ S_\theta(x)\,) + \beta\mathcal{L}_{\text{expl}}(\,E_T(x)\,,\ E_{S_\theta}(x))$$

Optimizing Explainers for Teaching

- **S**caffold-**Ma**ximizing **T**raining (**SMaT**) framework

$$\mathcal{L}_{\text{student}}(x; T, E_T, S_\theta, E_S) = \mathcal{L}_{\text{sim}}(T(x), S_\theta(x)) + \beta\mathcal{L}_{\text{expl}}(E_T(x), E_{S_\theta}(x))$$

$$\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S}) = \underbrace{\mathcal{L}_{\text{sim}}(T(x), S_\theta(x))}_{\textit{simulability loss}} + \beta\mathcal{L}_{\text{expl}}(E_{\phi_T}(x), E_{\phi_S}(x))$$

parameterized explainers

Optimizing Explainers for Teaching

- **S**caffold-**Ma**ximizing **T**raining (**SMaT**) framework

$$\mathcal{L}_{\text{student}}(x; T, E_T, S_\theta, E_S) = \mathcal{L}_{\text{sim}}(T(x), S_\theta(x)) + \beta \mathcal{L}_{\text{expl}}(E_T(x), E_{S_\theta}(x))$$

$$\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S}) = \mathcal{L}_{\text{sim}}(T(x), S_\theta(x)) + \beta \mathcal{L}_{\text{expl}}(E_{\phi_T}(x), E_{\phi_S}(x))$$

simulability loss

parameterized explainers

- Bi-level optimization:

(inner opt.) $$\theta^\star(\phi_T), \phi_S^\star(\phi_T) = \arg\max_{\theta, \phi_S} \mathbb{E}_{x \sim \mathcal{D}_{\text{train}}} \left[\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S}) \right]$$

student parameters and student explainer parameters

Optimizing Explainers for Teaching

- **S**caffold-**Ma**ximizing **T**raining (**SMaT**) framework

$$\mathcal{L}_{\text{student}}(x; T, E_T, S_\theta, E_S) = \mathcal{L}_{\text{sim}}(T(x), S_\theta(x)) + \beta\mathcal{L}_{\text{expl}}(E_T(x), E_{S_\theta}(x))$$

$$\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S}) = \underbrace{\mathcal{L}_{\text{sim}}(T(x), S_\theta(x))}_{\textit{simulability loss}} + \beta\mathcal{L}_{\text{expl}}(E_{\phi_T}(x), E_{\phi_S}(x))$$

parameterized explainers

- Bi-level optimization:

(inner opt.) $$\theta^\star(\phi_T), \phi_S^\star(\phi_T) = \arg\max_{\theta, \phi_S} \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S})\right]$$

student parameters and student explainer parameters

(outer opt.) $$\phi_T^\star = \arg\max_{\phi_T} \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\mathcal{L}_{\text{sim}}(T(x), S_{\theta^\star(\phi_T)})\right]$$

teacher explainer parameters

Optimizing Explainers for Teaching

How can we optimize this?

- Bi-level optimization:

(inner opt.) $\theta^\star(\phi_T), \phi_S^\star(\phi_T) = \arg\max_{\theta, \phi_S} \mathbb{E}_{x \sim \mathcal{D}_{\text{train}}} \left[\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S}) \right]$

student parameters and student explainer parameters

(outer opt.) $\phi_T^\star = \arg\max_{\phi_T} \mathbb{E}_{x \sim \mathcal{D}_{\text{test}}} \left[\mathcal{L}_{\text{sim}}(T(x), S_{\theta^\star(\phi_T)}) \right]$

teacher explainer parameters

Optimizing Explainers for Teaching

How can we optimize this?

- Assume the explainers are differentiable

- Bi-level optimization:

(inner opt.) $$\theta^\star(\phi_T), \phi_S^\star(\phi_T) = \arg\max_{\theta,\phi_S} \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S})\right]$$

student parameters and student explainer parameters

(outer opt.) $$\phi_T^\star = \arg\max_{\phi_T} \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\mathcal{L}_{\text{sim}}(T(x), S_{\theta^\star(\phi_T)})\right]$$

teacher explainer parameters

Optimizing Explainers for Teaching

How can we optimize this?

- Assume the explainers are differentiable
- Explicit differentiation with a truncated gradient update

- Bi-level optimization:

(inner opt.) $\theta^\star(\phi_T), \phi_S^\star(\phi_T) = \arg\max_{\theta,\phi_S} \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S})\right]$

student parameters and student explainer parameters

(outer opt.) $\phi_T^\star = \arg\max_{\phi_T} \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\mathcal{L}_{\text{sim}}(T(x), S_{\theta^\star(\phi_T)})\right]$

teacher explainer parameters

Optimizing Explainers for Teaching

How can we optimize this?

- Assume the explainers are differentiable
- Explicit differentiation with a truncated gradient update
- Diff. through a gradient operation ⇔ JAX for Hessian-vector products

- Bi-level optimization:

(inner opt.) $\theta^\star(\phi_T), \phi_S^\star(\phi_T) = \arg\max_{\theta,\phi_S} \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[\mathcal{L}_{\text{student}}(x; T, E_{\phi_T}, S_\theta, E_{\phi_S})\right]$

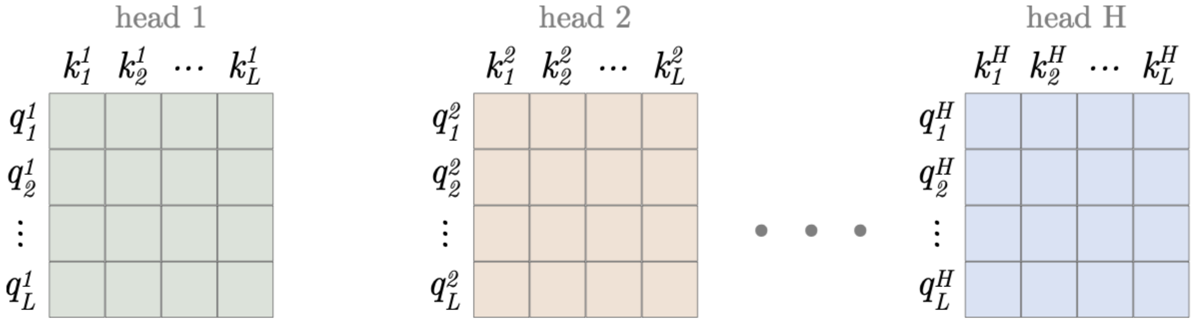
student parameters and student explainer parameters

(outer opt.) $\phi_T^\star = \arg\max_{\phi_T} \mathbb{E}_{x\sim\mathcal{D}_{\text{test}}}\left[\mathcal{L}_{\text{sim}}(T(x), S_{\theta^\star(\phi_T)})\right]$

teacher explainer parameters

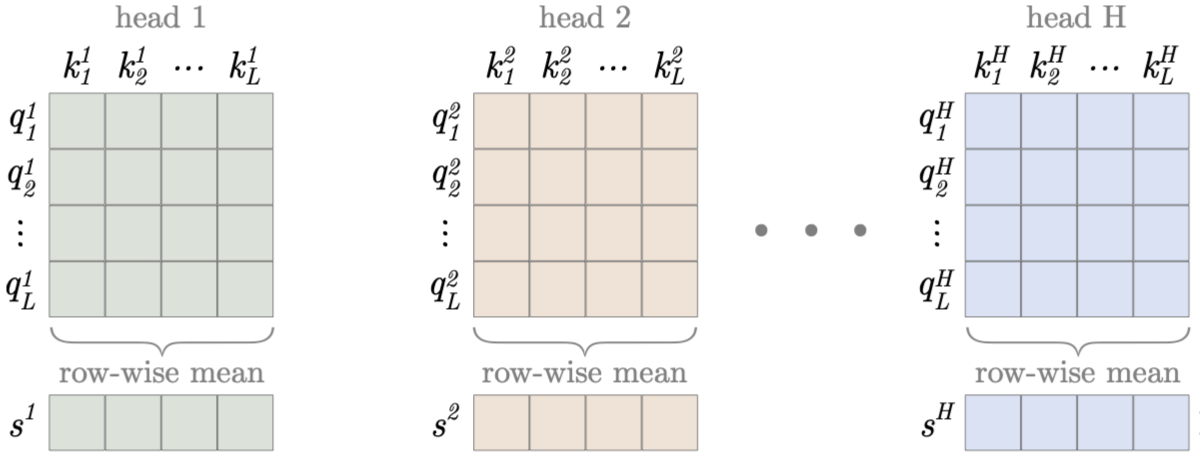
Differentiable, Parameterized Explainer

- Head-level parameterization:



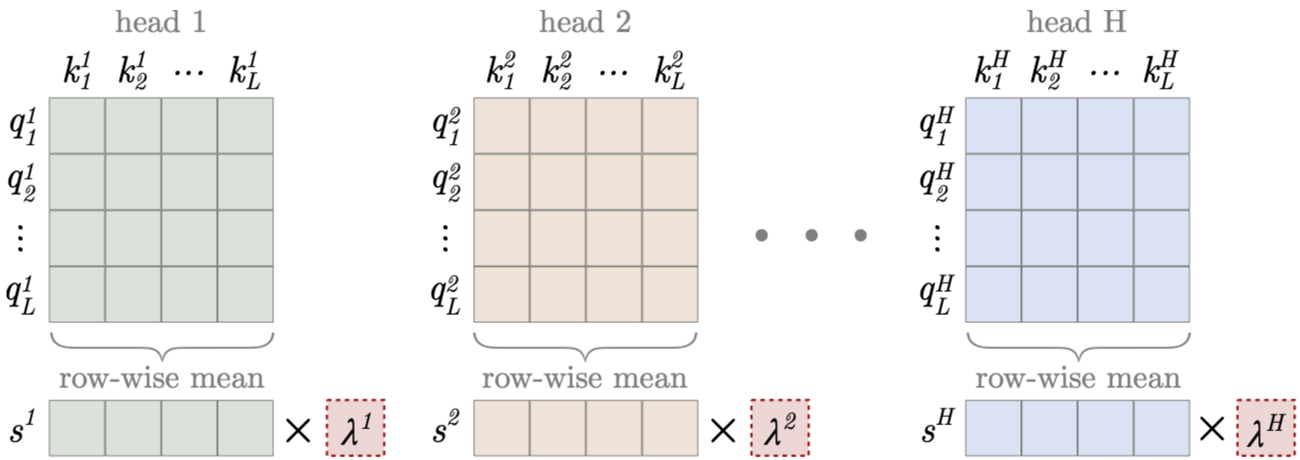
Differentiable, Parameterized Explainer

- Head-level parameterization:



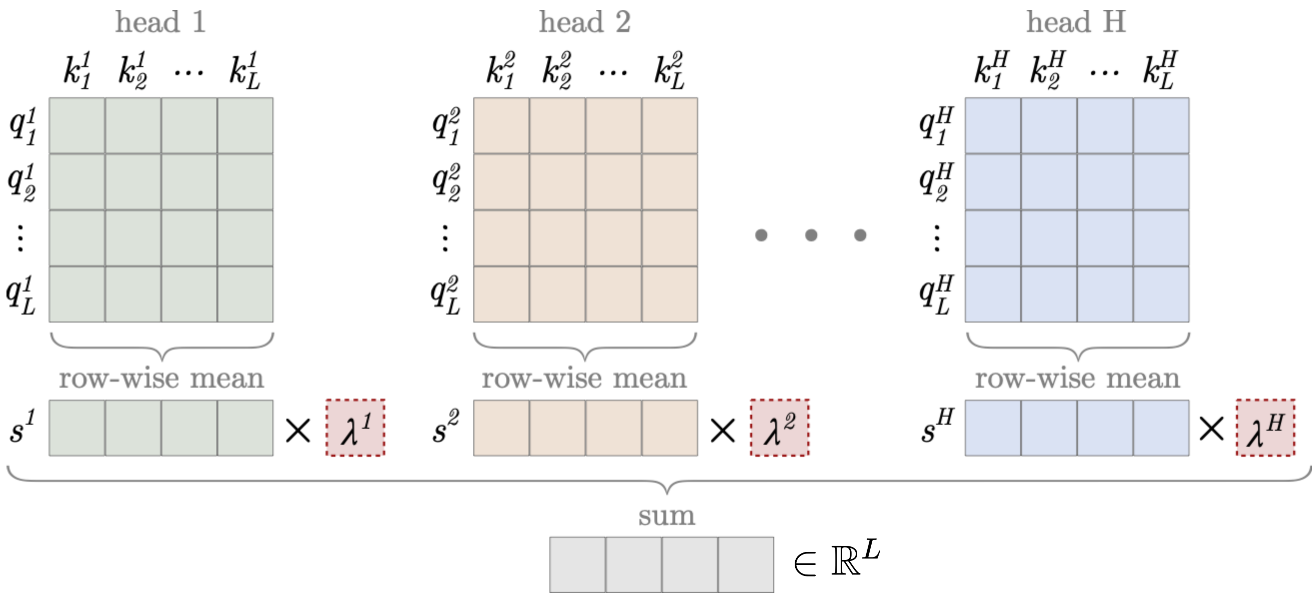
Differentiable, Parameterized Explainer

- Head-level parameterization:



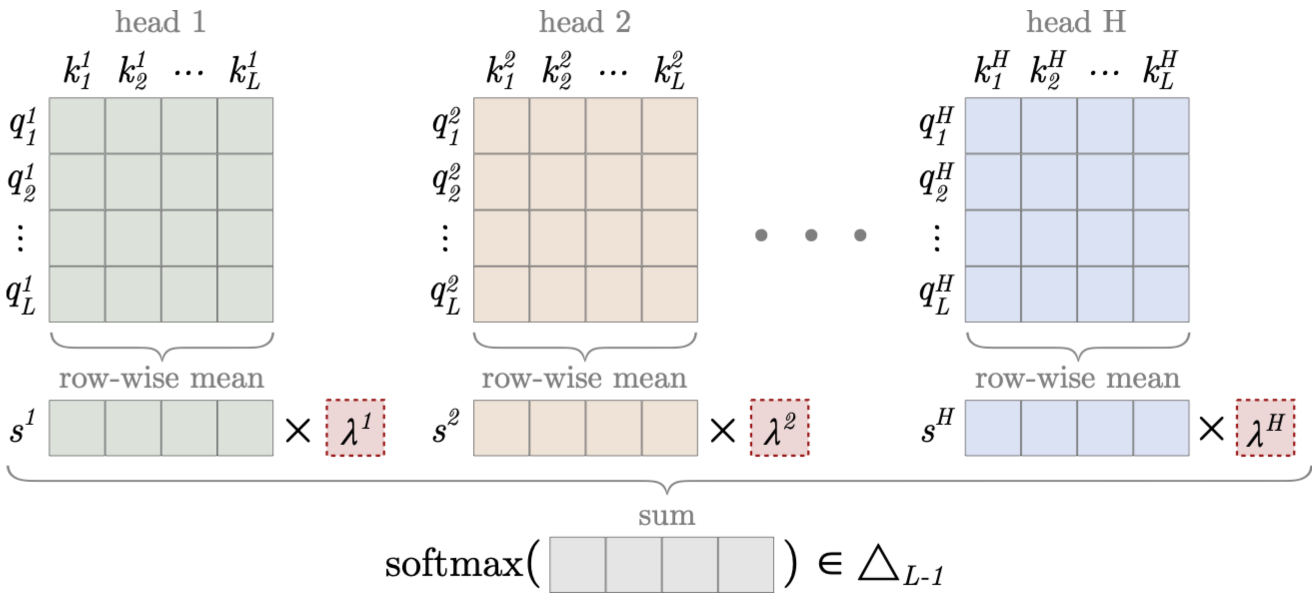
Differentiable, Parameterized Explainer

- Head-level parameterization:



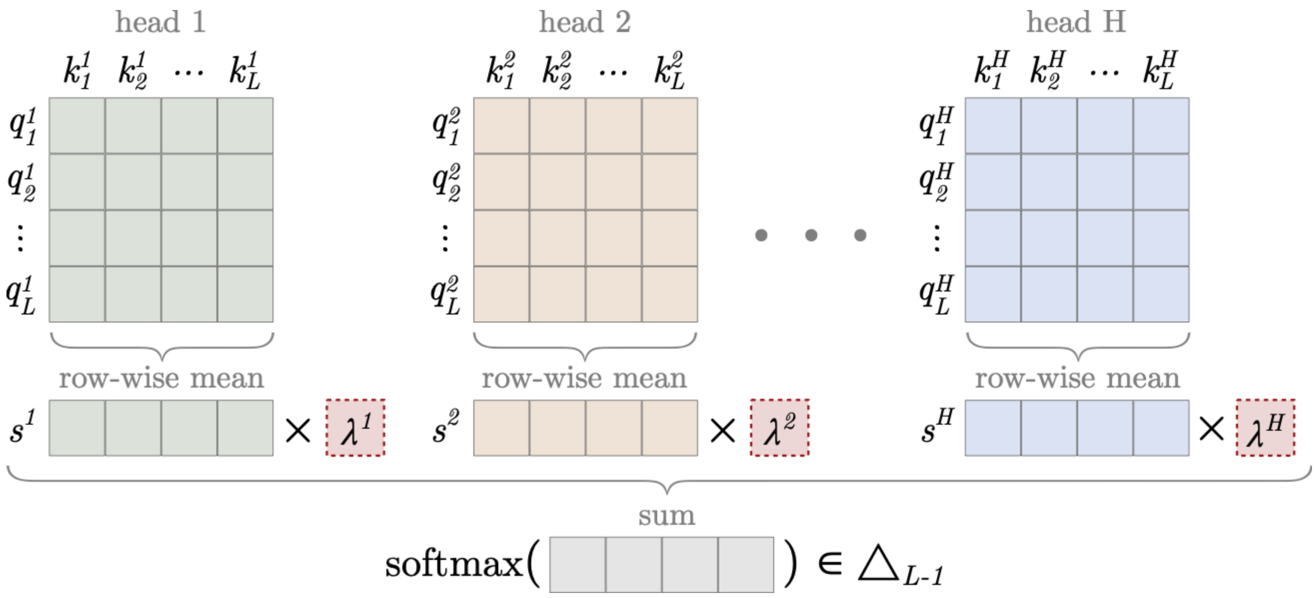
Differentiable, Parameterized Explainer

- Head-level parameterization:



Differentiable, Parameterized Explainer

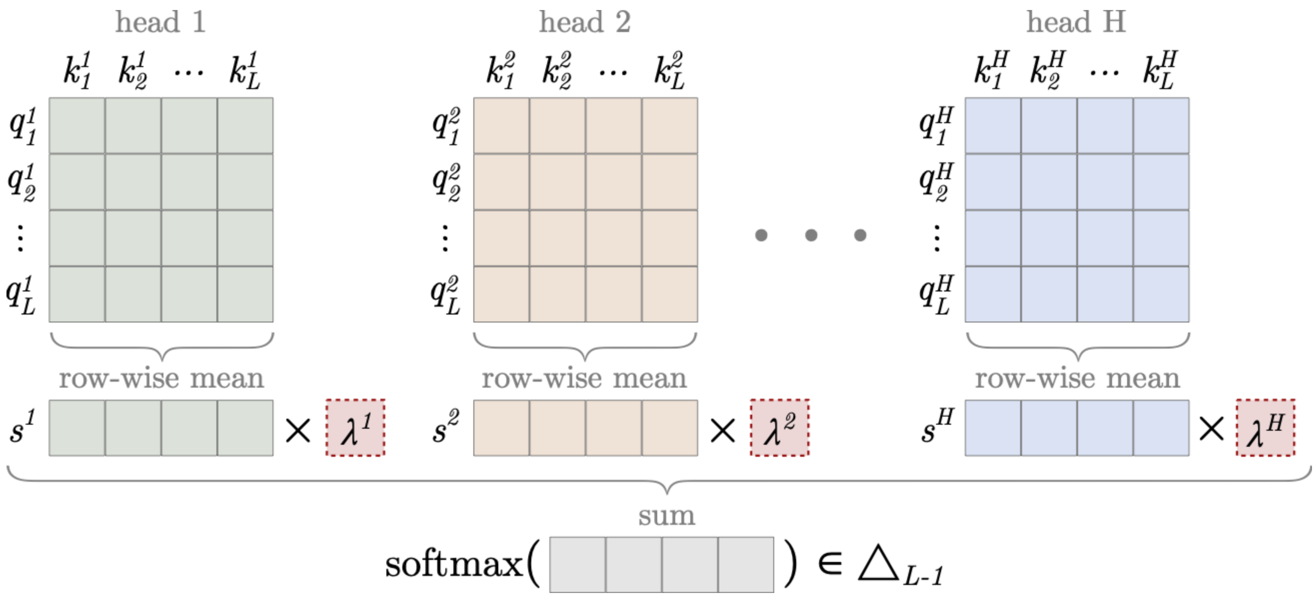
- Head-level parameterization:



$$\lambda_T = \mathrm{normalize}(\phi_T) \in \triangle_{H-1}$$

Differentiable, Parameterized Explainer

- Head-level parameterization:



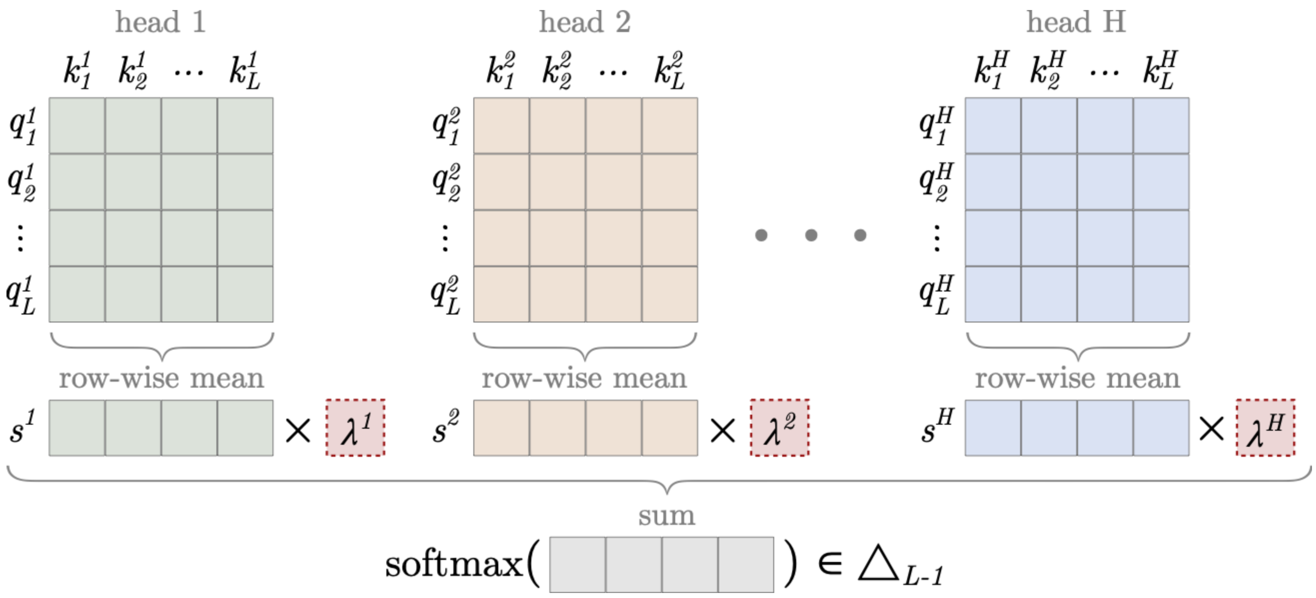
$$\lambda_T = \text{normalize}(\phi_T) \in \triangle_{H-1}$$

$$\text{sparsemax}(z) = \arg\min_{p \in \triangle_{H-1}} \|p - z\|_2$$

From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. Martins and Astudillo, 2016. (ICML)

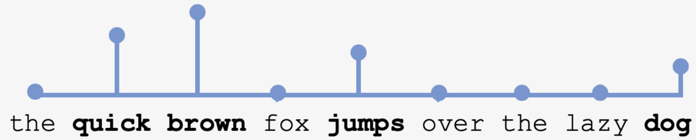
Differentiable, Parameterized Explainer

- Head-level parameterization:



$$\lambda_T = \mathrm{normalize}(\phi_T) \in \triangle_{H-1}$$

$$\mathrm{sparsemax}(z) = \arg\min_{p \in \triangle_{H-1}} \|p - z\|_2$$



From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. Martins and Astudillo, 2016. (ICML)

Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)

Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)

Experiments: simulability

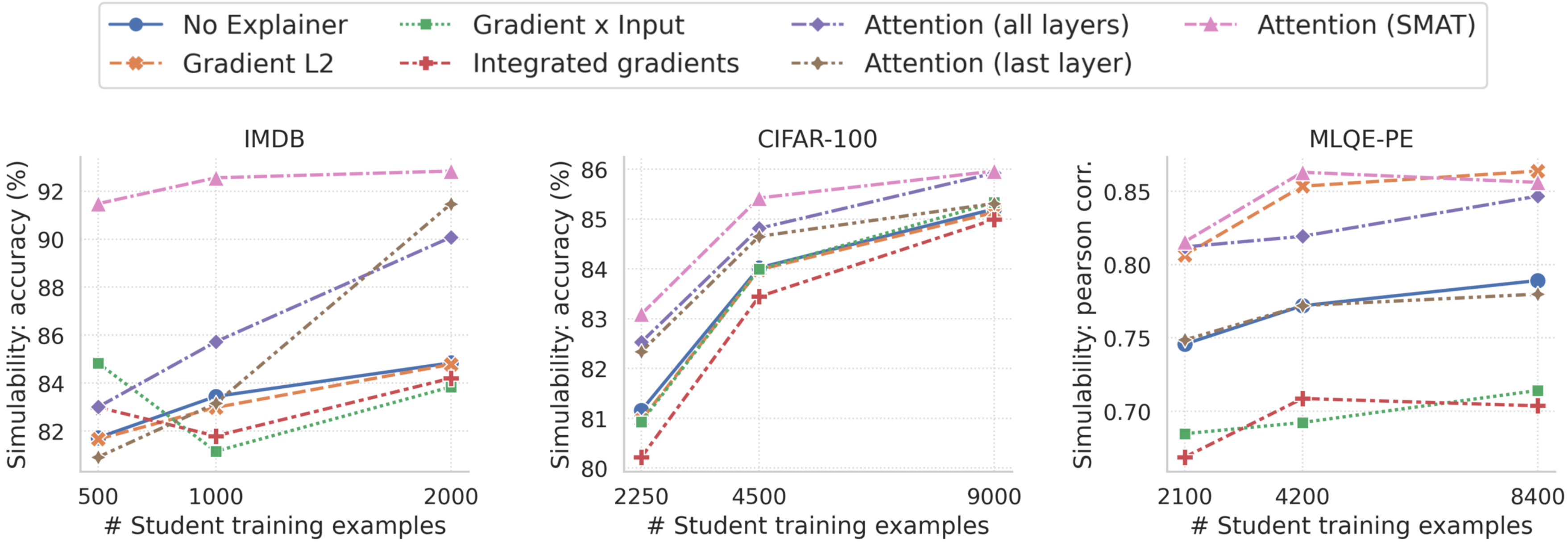
- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)

Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)

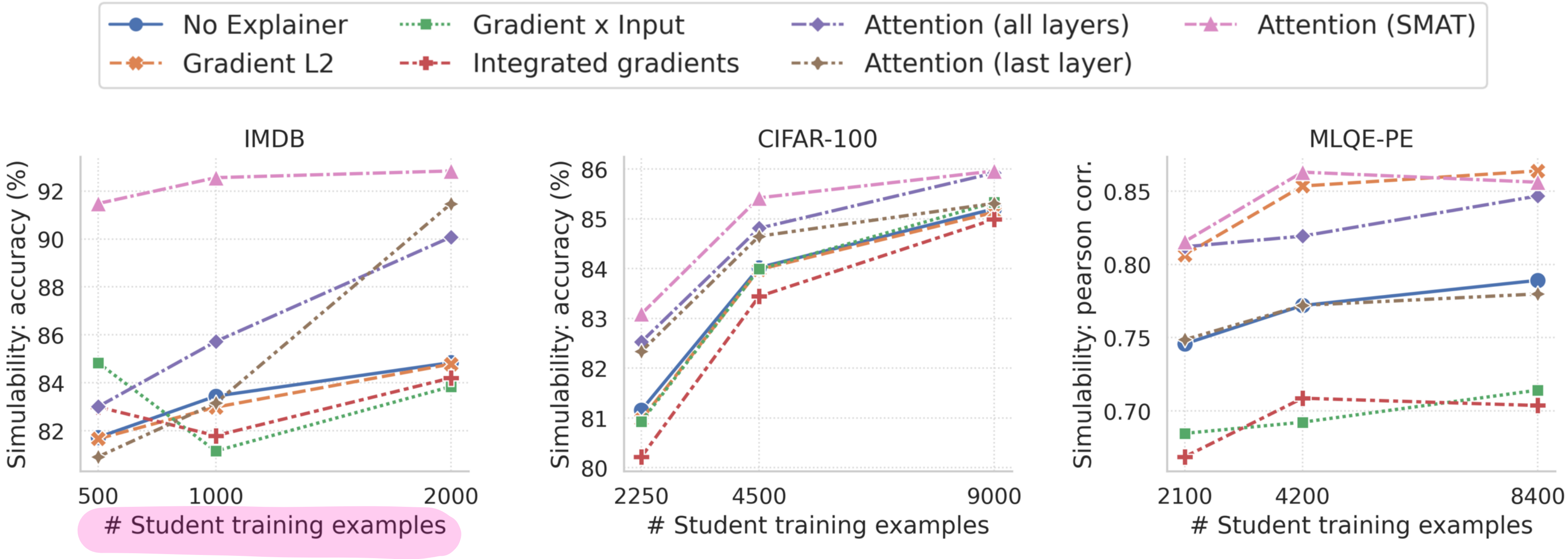
Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)



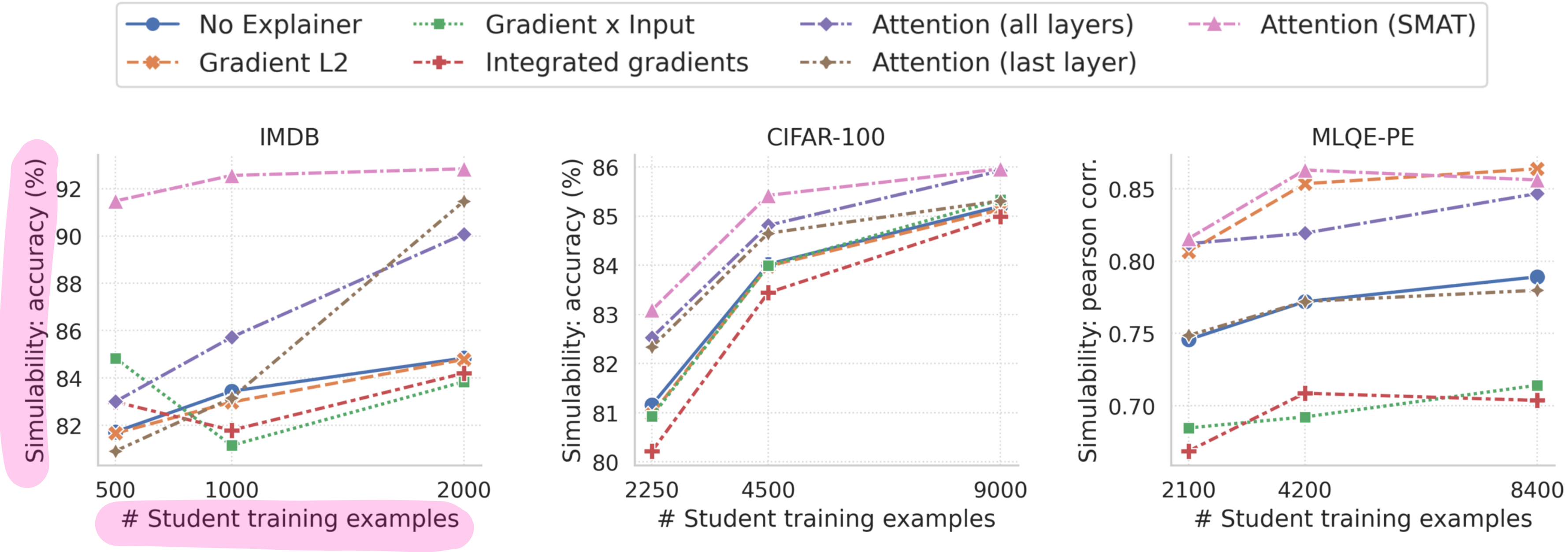
Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)



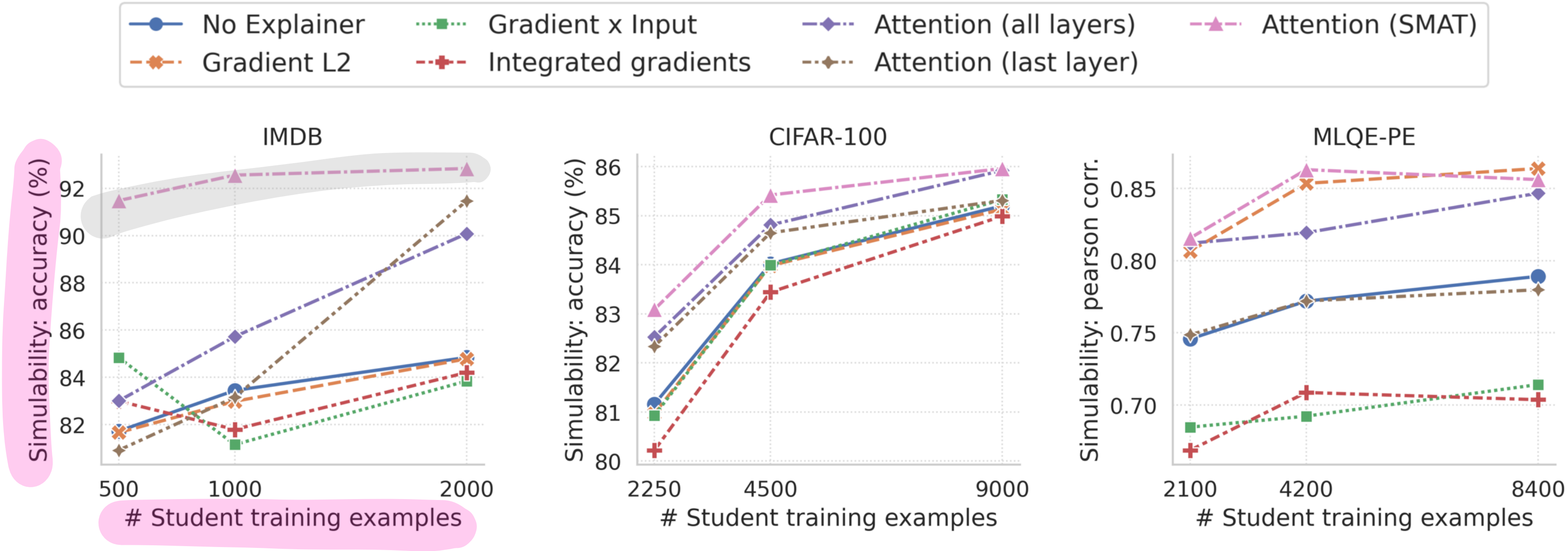
Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)



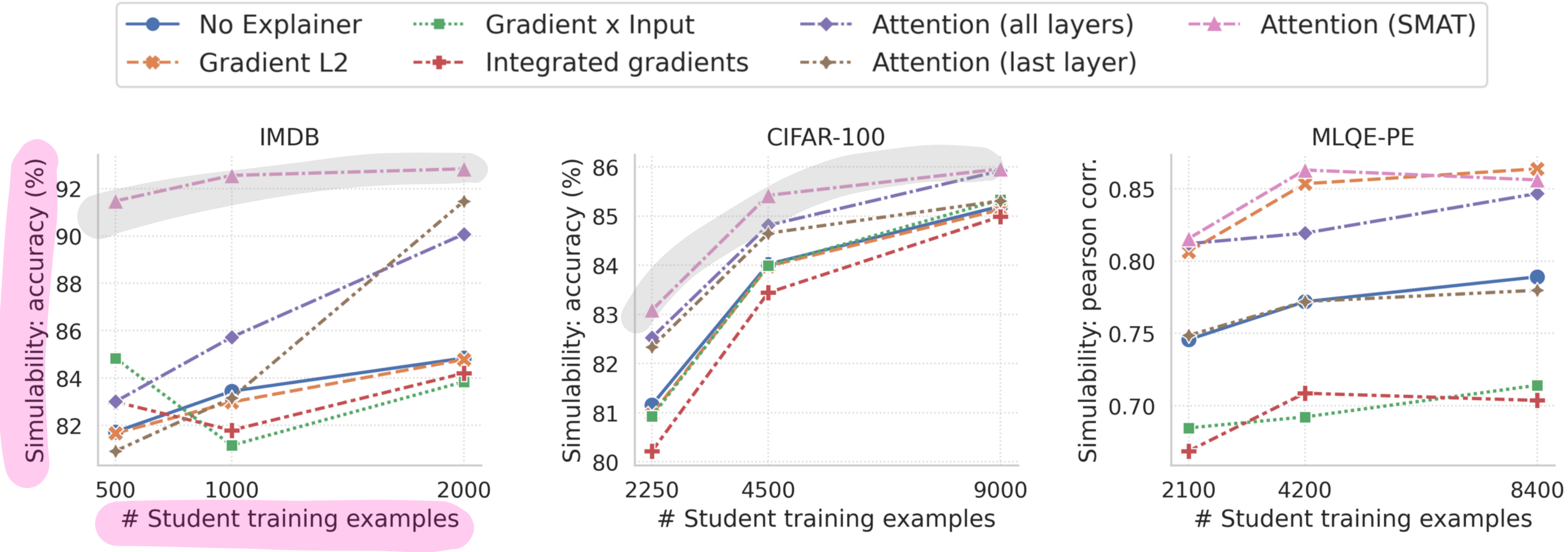
Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)



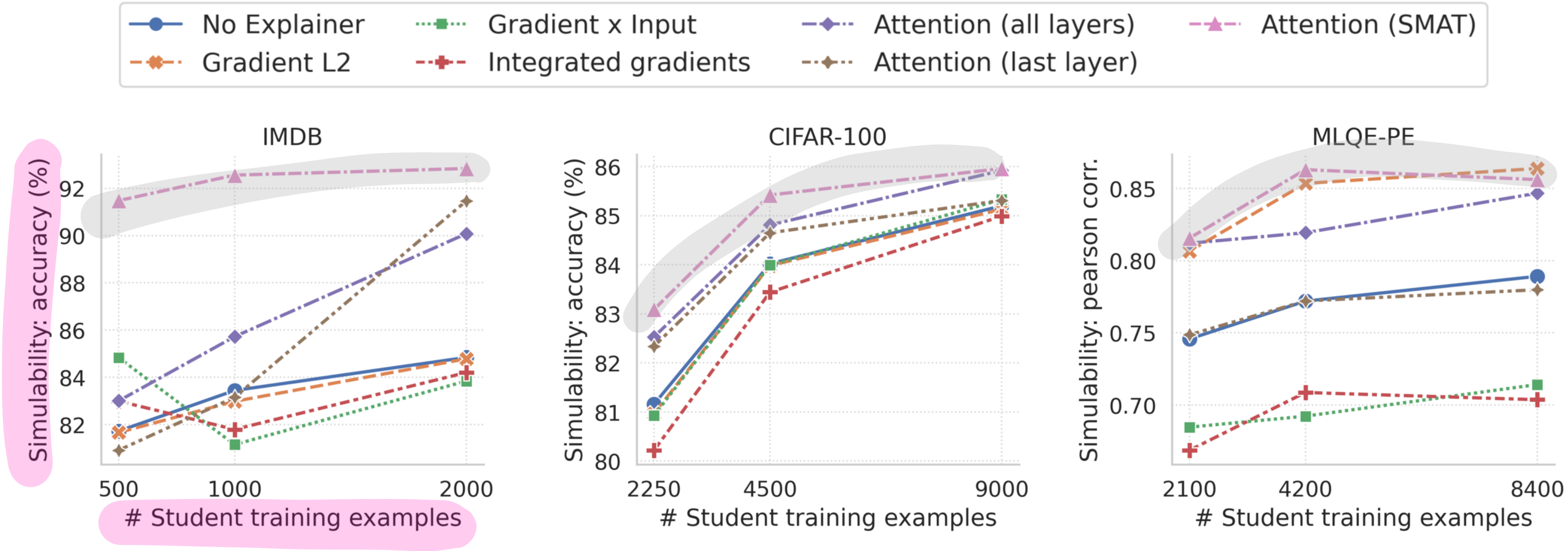
Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)



Experiments: simulability

- Text classification (IMDB)
- Image classification (CIFAR-100)
- Machine Translation Quality Estimation (MLQE-PE)



Experiments: plausibility

- *Plausiblity (human-likeness)* of the explainers

Text Classification

	AUC
Grad. L2	0.65
Grad. × Input	0.51
Integrated Grad.	0.53
Attn. (*all layers*)	0.68
Attn. (*last layer*)	0.61
Attn. (**SMaT**)	**0.73**
Attn. (*best layer*)*	0.75
Attn. (*best head*)*	0.75

Image Classification

	Rank	*TrueSkill*
Grad. × Input	3-4	-2.7±.67
Integ. Grad.	3-4	-2.1±.67
Attn. (*all lx.*)	2	0.7±.67
Attn. (**SMaT**)	**1**	**4.3±.70**

Quality Estimation

	OVERALL	
	src.	tgt.
Gradient L2	**0.67**	0.59
Gradient × Input	0.61	0.54
Integrated Gradients	0.62	0.53
Attention (*all layers*)	0.62	0.59
Attention (*last layer*)	0.54	0.50
Attention (**SMaT**)	0.66	**0.60**
Attention (*best layer*)*	0.65	0.65
Attention (*best head*)*	0.67	0.66

Experiments: plausibility

- *Plausiblity (human-likeness)* of the explainers

Text Classification

	AUC
Grad. L2	0.65
Grad. × Input	0.51
Integrated Grad.	0.53
Attn. (*all layers*)	0.68
Attn. (*last layer*)	0.61
Attn. (**SMaT**)	**0.73**
Attn. (*best layer*)*	0.75
Attn. (*best head*)*	0.75

Image Classification

	Rank	*TrueSkill*
Grad. × Input	3-4	-2.7±.67
Integ. Grad.	3-4	-2.1±.67
Attn. (*all lx.*)	2	0.7±.67
Attn. (**SMaT**)	**1**	**4.3±.70**

Quality Estimation

	OVERALL	
	src.	tgt.
Gradient L2	**0.67**	0.59
Gradient × Input	0.61	0.54
Integrated Gradients	0.62	0.53
Attention (*all layers*)	0.62	0.59
Attention (*last layer*)	0.54	0.50
Attention (**SMaT**)	0.66	**0.60**
Attention (*best layer*)*	0.65	0.65
Attention (*best head*)*	0.67	0.66

Experiments: plausibility

- *Plausiblity (human-likeness)* of the explainers

Text Classification

	AUC
Grad. L2	0.65
Grad. × Input	0.51
Integrated Grad.	0.53
Attn. (*all layers*)	0.68
Attn. (*last layer*)	0.61
Attn. (SMaT)	**0.73**
Attn. (*best layer*)*	0.75
Attn. (*best head*)*	0.75

Image Classification

	Rank	*TrueSkill*
Grad. × Input	3-4	-2.7±.67
Integ. Grad.	3-4	-2.1±.67
Attn. (*all lx.*)	2	0.7±.67
Attn. (SMaT)	**1**	**4.3±.70**

Quality Estimation

	OVERALL src.	OVERALL tgt.
Gradient L2	**0.67**	0.59
Gradient × Input	0.61	0.54
Integrated Gradients	0.62	0.53
Attention (*all layers*)	0.62	0.59
Attention (*last layer*)	0.54	0.50
Attention (SMaT)	0.66	**0.60**
Attention (*best layer*)*	0.65	0.65
Attention (*best head*)*	0.67	0.66

attention (all layers): i ' ve seen river ##dance in person and nothing compares to the video , but the show is awesome . the dancers are amazing . the music is impact ##ing . and the overall performance is outstanding . i ' ve never seen anything like it ! i suggest that you see this show if you can ! ! !

attention (SMaT): i ' ve seen river ##dance in person and nothing compares to the video , but the show is awesome . the dancers are amazing . the music is impact ##ing . and the overall performance is outstanding . i ' ve never seen anything like it ! i suggest that you see this show if you can ! ! !

Experiments: plausibility

- *Plausiblity (human-likeness)* of the explainers

Text Classification

	AUC
Grad. L2	0.65
Grad. × Input	0.51
Integrated Grad.	0.53
Attn. (*all layers*)	0.68
Attn. (*last layer*)	0.61
Attn. (SMaT)	**0.73**
Attn. (*best layer*)*	0.75
Attn. (*best head*)*	0.75

Image Classification

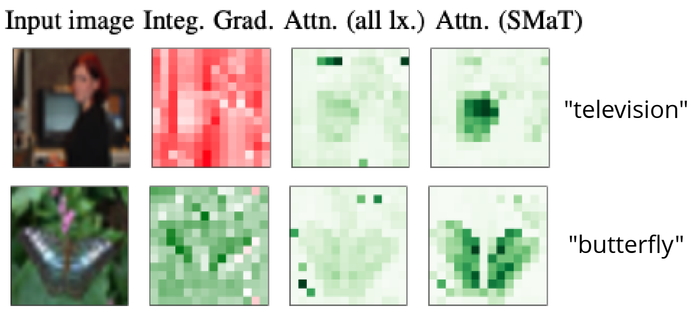
	Rank	*TrueSkill*
Grad. × Input	3-4	-2.7±.67
Integ. Grad.	3-4	-2.1±.67
Attn. (*all lx.*)	2	0.7±.67
Attn. (SMaT)	**1**	**4.3±.70**

Quality Estimation

	OVERALL	
	src.	tgt.
Gradient L2	**0.67**	0.59
Gradient × Input	0.61	0.54
Integrated Gradients	0.62	0.53
Attention (*all layers*)	0.62	0.59
Attention (*last layer*)	0.54	0.50
Attention (SMaT)	0.66	**0.60**
Attention (*best layer*)*	0.65	0.65
Attention (*best head*)*	0.67	0.66

attention (all layers): i ' ve seen river ##dance in person and nothing compares to the video , but the show is awesome . the dancers are amazing . the music is impact ##ing . and the overall performance is outstanding . i ' ve never seen anything like it ! i suggest that you see this show if you can ! ! !

attention (SMaT): i ' ve seen river ##dance in person and nothing compares to the video , but the show is awesome . the dancers are amazing . the music is impact ##ing . and the overall performance is outstanding . i ' ve never seen anything like it ! i suggest that you see this show if you can ! ! !

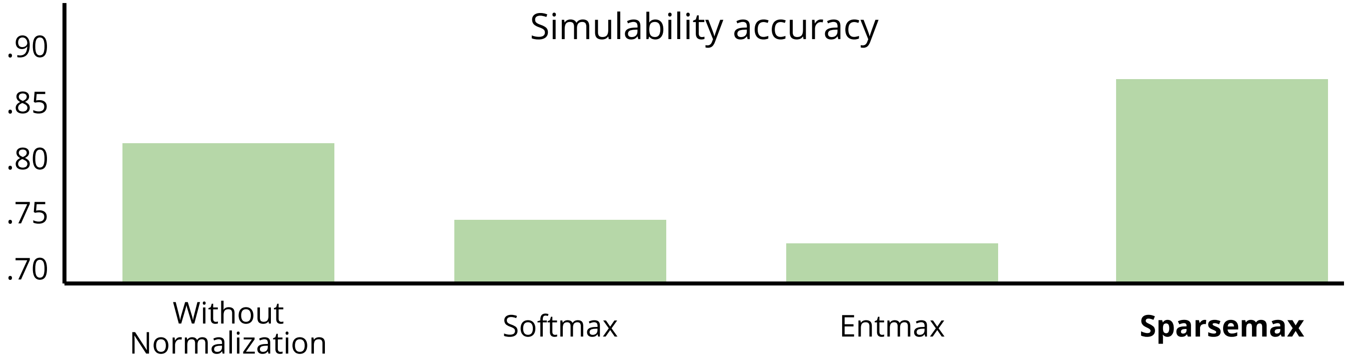


Experiments: head projection

- Normalization functions $\lambda_T = \mathrm{normalize}(\phi_T) \in \triangle_{H-1}$

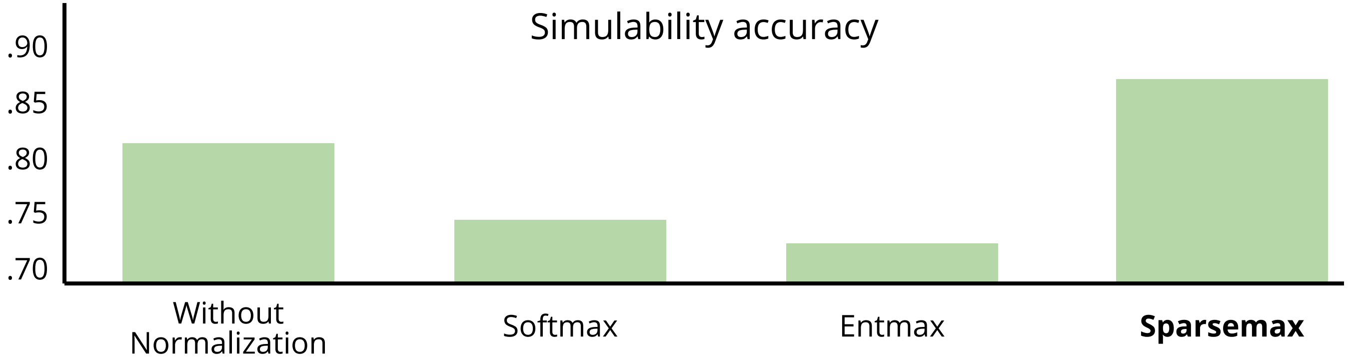
Experiments: head projection

- Normalization functions $\lambda_T = \mathrm{normalize}(\phi_T) \in \triangle_{H-1}$

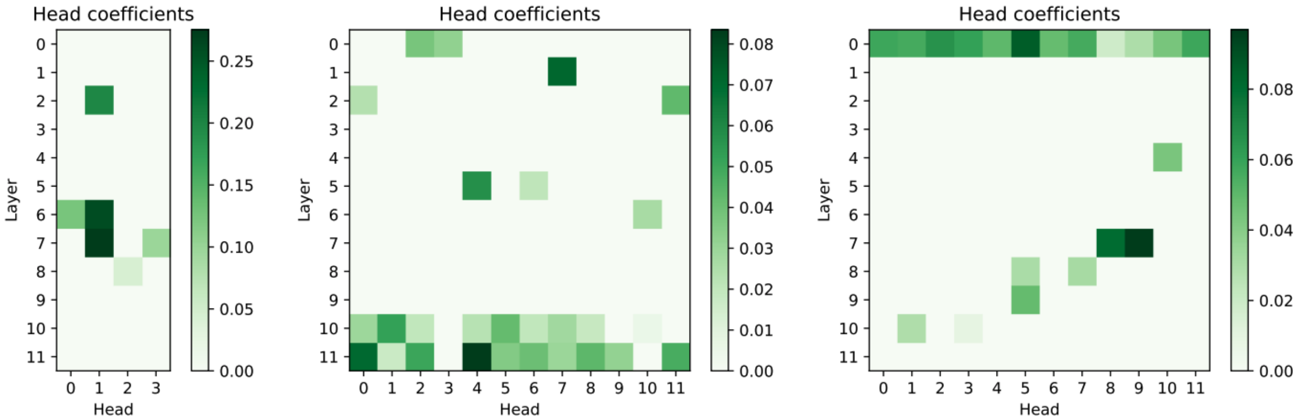


Experiments: head projection

- Normalization functions $\lambda_T = \text{normalize}(\phi_T) \in \triangle_{H-1}$

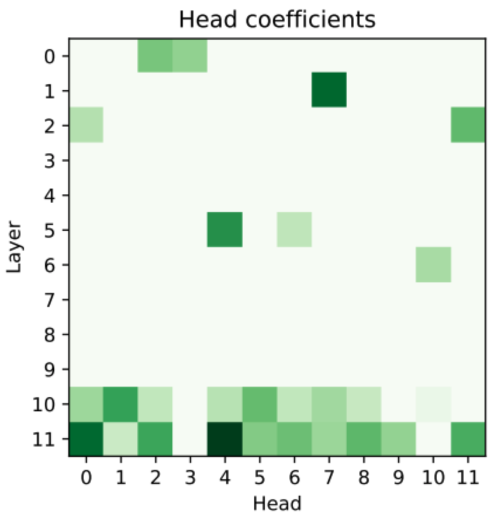


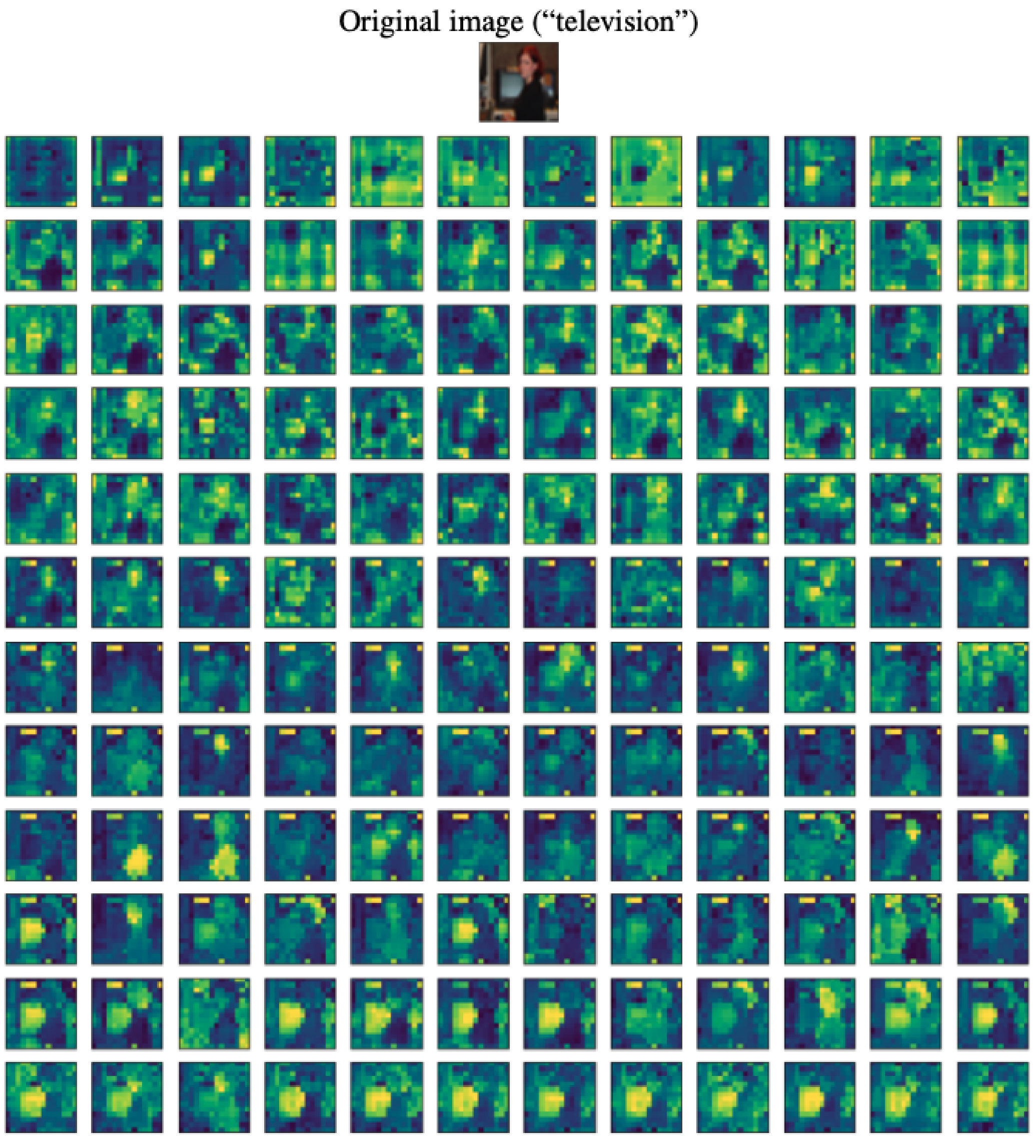
- Only a small subset of attention heads are deemed relevant by SMaT



Experiments: head projection

CIFAR-100





Experiments: head projection

Conclusions

- SMaT is a framework that optimizes explanations for teaching students
  - SMaT leads to **high simulability**
  - SMaT learns **plausible explanations**

- We hope this work motivates the interpretability community to consider **scaffolding** as valuable criterion **for evaluating and designing new methods**

(paper) arxiv.org/abs/2204.10810

(code) github.com/CoderPat/learning-scaffold

Introduction

- **Simulability** is particularly appealing for evaluating explanations
  - ✓ aligns with the goal of communicating the underlying model behavior
  - ✓ is easily measurable (both manually and automatically)
  - ✓ puts all explainability methods under a single perspective

Introduction

- **Simulability** is particularly appealing for evaluating explanations
  - ✓ aligns with the goal of communicating the underlying model behavior
  - ✓ is easily measurable (both manually and automatically)
  - ✓ puts all explainability methods under a single perspective

- Pruthi et al. (2021) proposed a framework for measuring simulability that

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Introduction

- **Simulability** is particularly appealing for evaluating explanations
    - ✓ aligns with the goal of communicating the underlying model behavior
    - ✓ is easily measurable (both manually and automatically)
    - ✓ puts all explainability methods under a single perspective

- Pruthi et al. (2021) proposed a framework for measuring simulability that
    - ⭐ disregards **trivial protocols**

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Introduction

- **Simulability** is particularly appealing for evaluating explanations
  - ✓ aligns with the goal of communicating the underlying model behavior
  - ✓ is easily measurable (both manually and automatically)
  - ✓ puts all explainability methods under a single perspective

- Pruthi et al. (2021) proposed a framework for measuring simulability that
  - ⭐ disregards **trivial protocols**

```
punctuation symbols  ⟹  positive
stop words           ⟹  negative
```

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)

Introduction

- **Simulability** is particularly appealing for evaluating explanations
  - ✓ aligns with the goal of communicating the underlying model behavior
  - ✓ is easily measurable (both manually and automatically)
  - ✓ puts all explainability methods under a single perspective

- Pruthi et al. (2021) proposed a framework for measuring simulability that
  - ⭐ disregards **trivial protocols**
  - 🧶 requires **an optimization procedure**

```
punctuation symbols  ⟹  positive
stop words           ⟹  negative
```

Evaluating Explanations: How much do explanations from the teacher aid students? Pruthi et. al. 2021. (TACL)