Learning to Scaffold: Optimizing Model Explanations for Teaching

Patrick Fernandes, w/
Marcos Treviso, Danish Pruthi,
André F. T. Martins, Graham Neubig
Carnegie Mellon University, Instituto Superior Técnico

CMU PORTUGAL SUMMIT 2022
NEW FRONTIERS IN TECH

*(to appear at **NeurIPS 2022**)*

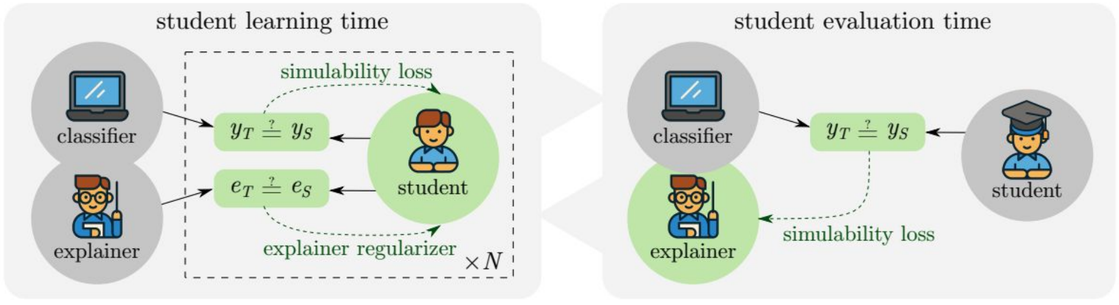
- Deep learning model's predictions are hard to **interpret/explain**
  - There is no widely accepted definition of what is a **good** explanation
- **Simulability**, as defining by Pruthi et al. (2022) is a promising criteria
  - Do explanations help "student" models/humans **learn to simulate** the "teacher"?

$$\theta_E^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}_{\text{train}}} \Big[\underbrace{\mathcal{L}_{\text{sim}}\left(S_\theta(x), T(x)\right)}_{\text{simulability loss}} + \beta \underbrace{\mathcal{L}_{\text{expl}}\left(E_S(S_\theta, x), E_T(T, x)\right)}_{\text{explainer regularizer}} \Big]$$

  - Requires an optimization procedure, so originally only used for evaluation

Can we optimize good explainers to maximize this simulability metric?

Optimizing Explaners for Teaching



- Suppose student/teacher explainers are **optimizable** (have parameters)

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

- **Crucial insight:** the optimal student depends on the teacher explainer
  - **Bi-level optimization**

$$\theta^*(\phi_T), \phi_S^*(\phi_T) = \arg\min_{\theta,\phi_S} \mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}_{\text{train}}} \left[\mathcal{L}_{\text{student}}(S_\theta, E_{\phi_S}, T, E_{\phi_T}, x)\right]$$

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

- Scaffold-Maximizing Training (**SMaT**)

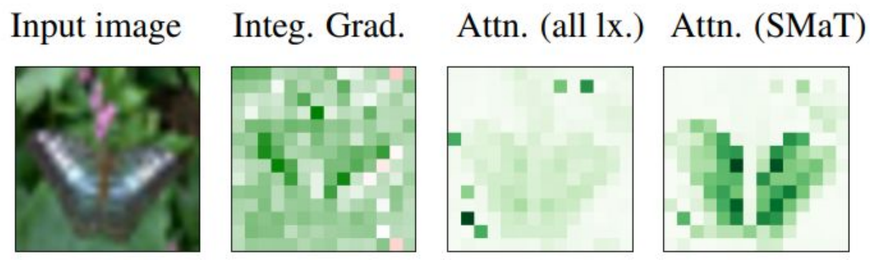
Experiments

- SMaT-trained explanations consistently lead to students that simulate better
  - Both in **text** and **image** tasks!

	2100	4200	8400
No Explainer	.7457 ± [.7366:.7528]	.7719 ± [.7660:.7802]	.7891 ± [.7860:.7964]
Gradient × Input	.6846 ± [.6781:.6894]	.6922 ± [.6885:.6965]	.7141 ± [.7136:.7147]
Integrated gradients	.6686 ± [.6677:.6694]	.7086 ± [.6994:.7101]	.7036 ± [.6976:.7037]
Attention (*all layers*)	.8120 ± [.7955:.8125]	.8193 ± [.8186:.8280]	.8467 ± [.8464:.8521]
Attention (*last layer*)	.7486 ± [.7484:.7534]	.7720 ± [.7672:.7726]	.7798 ± [.7717:.7814]
Attention (**SMaT**)	**.8156** ± [.8096:.8183]	**.8630** ± [.8412:.8724]	**.8561** ± [.8512:.8689]

(results on the ML-QE Quality Estimation dataset)

- Explanations align with how humans would explain similar decisions



(label: butterfly)

Paper & Code

- Available at:

Acknowledgements

This work was partially funded by Fundação para a Ciência e a Tecnologia under the scope of the CMU Portugal and P2020 project MAIA (LISBOA-01-0247- FEDER045909).

Carnegie Mellon Portugal
fct Fundação para a Ciência e a Tecnologia
Language Technologies Institute
TÉCNICO LISBOA
instituto de telecomunicações
Unbabel