Learn from A Rationalist: Distilling Intermediate Interpretable Rationales

Jiayi Dai[1,2], Randy Goebel[1,2]

[1] University of Alberta, [2] Alberta Machine Intelligence Institute

ICML International Conference On Machine Learning

paper available

Keywords interpretable ML, explainable AI, knowledge distillation

Background

Rationale extraction (RE)

RE offers an interpretable neural architecture through a select-predict architecture where two neural networks, i.e., generator *g* and predictor *f*, learn jointly to select a subset of features (i.e., a *rationale*) and make a prediction based on the rationale.

Chicken-and-egg dilemma

With only the remote supervision signal from the final prediction, the training process is difficult: the generator relies on the guidance of the predictor to select important features while the predictor relies on the output of the generator to learn task prediction. The issue is significantly exacerbated when the base neural networks are not sufficiently capable.

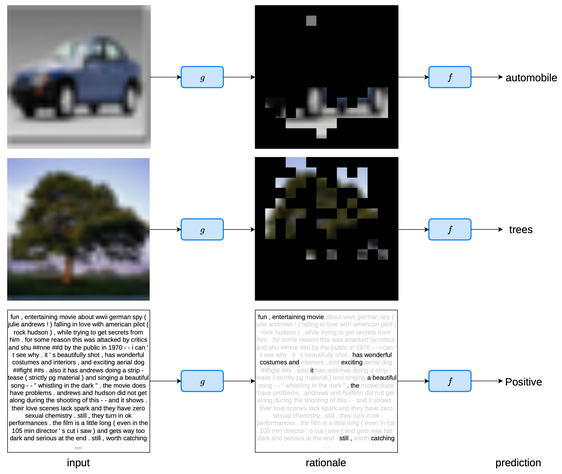

Figure 1. Examples of rationale extraction for ViT Base on CIFAR 10, CIFAR 100 and BERT Base on IMDB movie reviews (arranged from top to bottom).

Motivation

If discovering gravity law from scratch

One cannot derive the precise mathematical formula (the predictor) without first identifying that mass and distance are the critical governing variables (the rationale). Conversely, it is difficult to isolate these specific variables from the myriad of irrelevant factors (e.g., color, temperature, one's religious beliefs) without a working formula to verify their predictive power.

Interpretable and verifiable knowledge

However, once Isaac Newton published the gravity theory in 1687 which was written in an interpretable and verifiable way, even an average human could utilize the knowledge to make predictions on gravity force almost as accurately as Newton himself.

"If I have seen further, it is by standing on the shoulders of giants."

——Isaac Newton

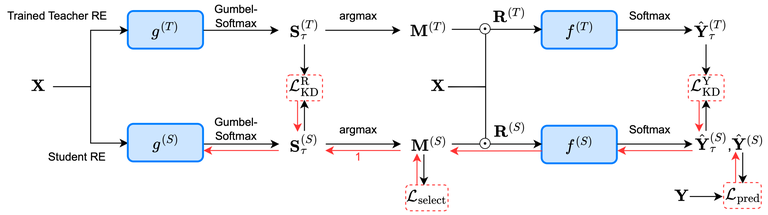

Figure 2. The schematic of rationale extraction with knowledge distillation.

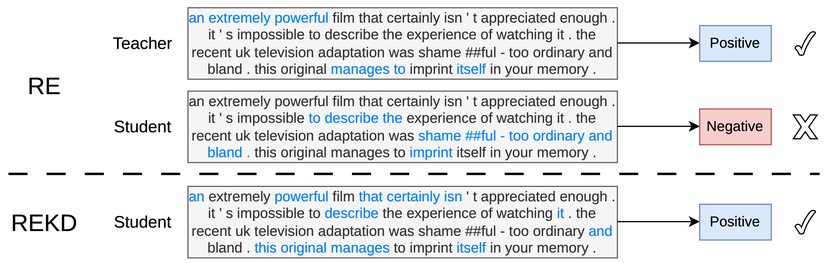

Figure 3. An actual example of REKD from IMDB movie reviews.

Rationale extraction with knowledge distillation (REKD)

Gumbel-Softmax

$$S_{l,i} = \frac{\exp((Z_{l,i} + G_{l,i})/\tau)}{\sum_{j=0}^{1} \exp((Z_{l,j} + G_{l,j})/\tau)}$$

$Z_{l,i}$ are the output logits from *g*; $G_{l,i}$ are i.i.d samples from Gumbel(0, 1); τ is a temperature controlling the sharpness of the distributions. It is used for differentiable rationale sampling.

RE

$$\mathcal{L}_{\text{RE}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{select}}\mathcal{L}_{\text{select}}$$

$$= \text{CE}(\hat{\mathbf{Y}}, \mathbf{Y}) + \lambda_{\text{select}} \left(\sum_{l=0}^{L-1} M_l - L \cdot p_{\text{target}} \right)^2$$

M, obtained by discretizing *S,* is a binary mask over *L* features; *p_target* is the desired rationale length. RE objective = predictive loss + rationale selection loss. It is to encourage the sufficiency and necessity of the selected rationales.

KD

"Given input *X*, these features $R^{(T)}$ are important and, because of these features, the final prediction is $Y^{(T)}$".

$$\mathcal{L}_{\text{KD}} = \lambda_R \mathcal{L}^{\text{R}}_{\text{KD}} + \tau^2 \mathcal{L}^{\text{Y}}_{\text{KD}}$$

$$= \lambda_R \sum_{l=0}^{L-1} D_{\text{KL}}(\mathbf{S}^{(T)}_{\tau,l} \parallel \mathbf{S}^{(S)}_{\tau,l}) + \tau^2 D_{\text{KL}}\left(\mathbf{Y}^{(T)}_{\tau} \parallel \mathbf{Y}^{(S)}_{\tau}\right)$$

L^R_{KD} and L^Y_{KD} are the KL Divergence loss terms for rationale and prediction distillations.

Distilling rationale is performed by matching the Gumbel-Softmax distributions between the student and the teacher; Distilling prediction is to match their classification distributions.

Loss scale τ^2 is applied to the prediction distillation as in standard KD. However, the gradient magnitude for rationale distillation has been balanced with RE loss due to the temperature scale in Gumbel-Softmax and, hence, no τ^2 scale.

Unified annealing scheduler

$$\tau_k = \tau_0 e^{-\gamma k} \quad \text{where} \quad \gamma = \frac{1}{K} \log\left(\frac{\tau_0}{\tau_K}\right)$$

A unified temperature scheduler for Gumbel-Softmax and KD. τ_0 and τ_k are the start and end temperatures; *K* is the total steps indexed by *k*. It creates a curriculum learning strategy for KD.

REKD

The final objective is to add up the RE and KD loss terms.

$$\mathcal{L}_{\text{REKD}} = \alpha\mathcal{L}_{\text{RE}} + (1-\alpha)\mathcal{L}_{\text{KD}}$$

$$= \alpha(\mathcal{L}_{\text{pred}} + \lambda_{\text{select}}\mathcal{L}_{\text{select}}) + (1-\alpha)(\lambda_R \mathcal{L}^{\text{R}}_{\text{KD}} + \tau^2 \mathcal{L}^{\text{Y}}_{\text{KD}})$$

It is to provide the student with teacher supervision while it explores rationale extraction.

Results

The Bitter Lesson of XAI evaluation

Should machine rationales be aligned with human rationale annotations? It is often referred to as *plausibility*.

No! The knowledge in the data does not have to be aligned with human knowledge. RE is to learn from the data about what features are important for what predictions.

Instead, an objective evaluation metric is to measure the task predictive performance under a rationale constraint.

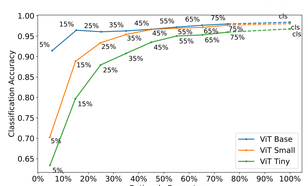

Figure 4. Rationale ratio vs. accuracy for ViT Base, Small and Tiny on CIFAR 10 by varying target from 5% to 75% and classification.

The accuracy of RE drops more from classification to a low target rationale ratio when the neural model is less capable.

Task	ViT Model	Accuracy	R %
CLS	Base	.984(.002)	100%
	Small	.981(.001)	100%
	Tiny	.968(.002)	100%
RE	Base (T)	.964(.009); .020 ↓	15.2%(.4%)
	Small	.889(.019); .092 ↓	14.9%(.2%)
	Tiny	.797(.054); .171 ↓	14.8%(.5%)
REKD	Small (S)	.968(.006); .079 ↑	14.8%(.2%)
	Tiny (S)	.936(.003); .139 ↑	14.9%(.3%)

Table 1. RE and REKD predictive accuracy with 15% rationale length with ViT Base as Teacher and Small/Tiny as student on CIFAR10. CLS refers to the classification accuracy, i.e., rationale is full input.

Task	ViT Model	Accuracy	R %
CLS	Base	.947(.004)	100%
	Small	.944(.003)	100%
	Tiny	.903(.003)	100%
RE	Base (T)	.830(.028); .117 ↓	15.1%(.7%)
	Small	.779(.028); .165 ↓	14.8%(.9%)
	Tiny	.645(.005); .258 ↓	14.8%(.5%)
REKD	Small (S)	.845(.015); .066 ↑	15.2%(.2%)
	Tiny (S)	.777(.007); .132 ↑	15.1%(.3%)

Table 2. RE and REKD predictive accuracy with 15% rationale length with ViT Base as Teacher and Small/Tiny as student on CIFAR100 (with 20 coarse classes).

Task	BERT Model	Accuracy	R %
CLS	Base	.914(.004)	100%
	Small	.889(.003)	100%
	Mini	.877(.003)	100%
RE	Base (T)	.912(.004); .002 ↓	9.7%(.6%)
	Small	.881(.005); .008 ↓	10.3%(.6%)
	Mini	.863(.005); .014 ↓	10.2%(.8%)
REKD	Small (S)	.906(.002); .025 ↑	9.8%(.4%)
	Mini (S)	.892(.002); .029 ↑	10.0%(.3%)

Table 3. RE and REKD predictive accuracy with 10% rationale length with BERT Base as Teacher and Small/Mini as student on IMDB movie reviews (binary classification).

REKD significantly improves the predictive performance of the student RE models with reduced variance across both vision and language tasks (CIFAR10/100 and IMDB movie reviews) using multiple variants of vision transformer (ViT) and BERT models as base RE models. In IMDB, student BERT models under a 10% rationale constraint even outperforms classification results with full inputs.