



Explanation Bottleneck Models (AAAI 2025 Oral)

Shin'ya Yamaguchi and Kosuke Nishida

NTT

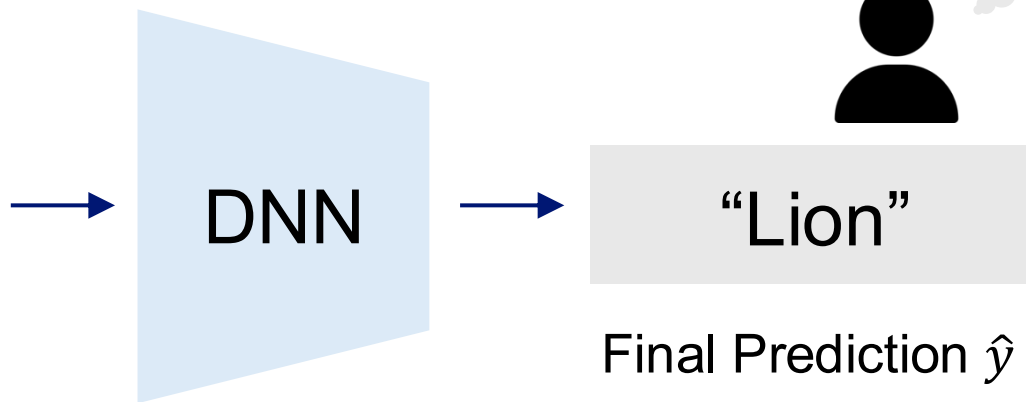
Background

❖ DNN は強力な推論能力を持つが予測の解釈性が低い

- ▷ なぜその予測結果が得られたのかを正確に説明できない
- ▷ 例. 画像分類モデル

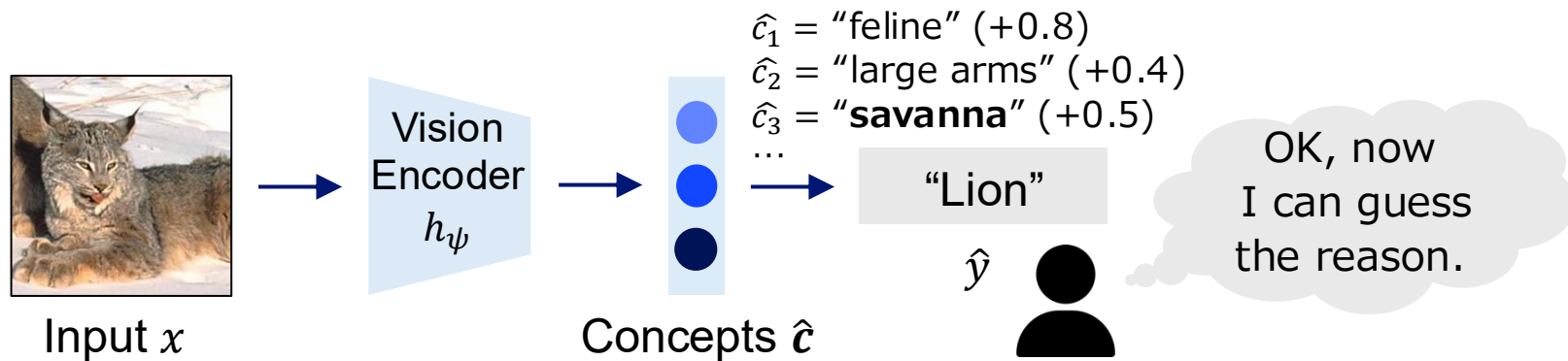


Input x
(Ground Truth: “Lynx”)



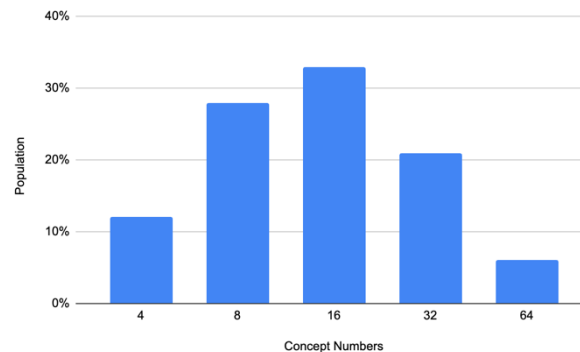
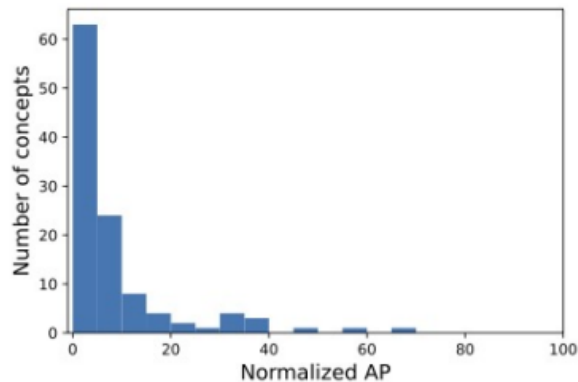
❖ Concept Bottleneck Models (CBMs, [1])

- ▶ **コンセプト**と呼ばれる解釈可能な中間出力を持つ二段階予測モデル
 - コンセプト := データの意味的な構成要素 (アノテーションで付与)
 - 事前定義したコンセプトによって最終予測の原因を解釈できる



❖ 事前定義コンセプト以外の説明を生成できない

- ▷ 事前定義コンセプト集合はタスクごとに設計
 - 説明可能なコンセプトの数は限定される
 - コンセプト数の増加は二段階予測モデルの学習を難しくする [2]
 - 人間ユーザにとっても多数のコンセプトの提示は解釈性を低下させる [2]



コンセプトの数が増えるほど訓練は困難に

人間はほどほどの数のコンセプトを好む

Method

❖ RQ: 事前コンセプト集合なしで解釈可能な DNN を実現できるか？

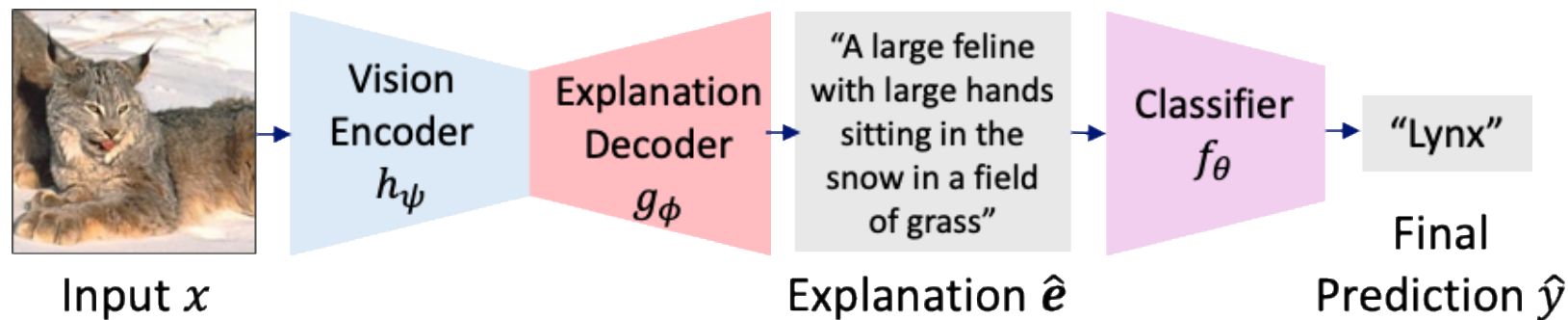
- ▶ より柔軟で人間にとって解釈性の高い説明生成とは？

❖ Key Idea: 自然言語で予測の説明を生成する

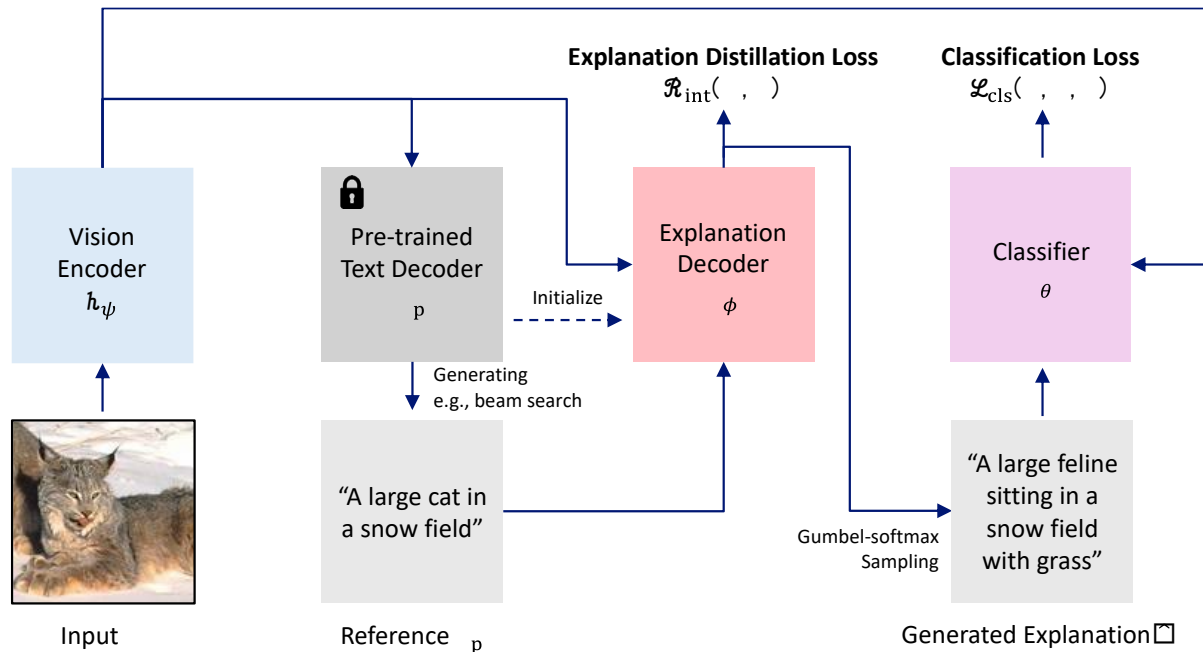
- ▶ 事前学習済みのマルチモーダルエンコーダ・デコーダモデル [4] を活用
→ 自然言語とデータ (e.g., 画像) の対応関係を記述可能
- ▶ コンセプトの代わりに自然言語の説明から最終出力を予測できるか？

❖ Explanation Bottleneck Models (XBMs)

- ▶ 入力データから自然言語説明を生成 → テキスト分類器で最終ラベルを出力
- ▶ 事前学習済みのエンコーダ・デコーダをファインチューニングして構成



- ❖ エンコーダ h_ψ , テキストデコーダ g_ϕ , テキスト分類器 f_θ を
ターゲットタスク (分類) と**説明蒸留損失** (Explanation Distillation) で訓練



目的関数

$$\min_{\theta, \phi, \psi} \mathcal{L}_{\text{cls}}(\theta, \phi, \psi) + \lambda \mathcal{R}_{\text{int}}(\phi, \psi)$$

$$\mathcal{L}_{\text{cls}}(\theta, \phi, \psi) = \mathbb{E}_{(x, y) \in \mathcal{D}} \ell_{\text{CE}}(f_\theta(g_\phi(h_\psi(x))), y)$$

$\mathcal{R}_{\text{int}}(\phi, \psi)$: 説明蒸留損失

❖ 分類損失 $\mathcal{L}_{cls}$ のみのナイーブな最適化の課題

- ▷ テキストデコーダ g_ϕ が自然言語の構造を無視して最適化される
→ 出力テキストが崩壊し、説明として無意味な文字列を生成 (e.g., 22222...)

❖ アイデア: 事前学習デコーダの分布 $p_\phi(e)$ によってモデル分布 $q(e)$ を罰則

$$\mathcal{R}_{\text{int}}(\phi, \psi) = D_{\text{KL}}(q \parallel p_\phi) = \sum_{e \in \mathcal{E}} q(e|x) \log \left(\frac{q(e|x)}{p_\phi(e|x)} \right) = \mathbb{E}_{e \sim q(e|x)} \log \left(\frac{q(e|x)}{p_\phi(e|x)} \right)$$

- ▷ 系列レベル蒸留 (teacher forcing) [5] で定式化
→ サンプリングプロセス全体への backprop を避けつつ自然言語構造を維持

$$\mathcal{R}_{\text{int}}(\phi, \psi) \approx - \sum_{e \in \mathcal{E}} \mathbb{I}_{e=e_p} \log p_\phi(e|x) = -\log p_\phi(e = e_p|x)$$

❖ 事前学習モデルから参照説明テキストを生成 → 説明損失蒸留を計算

▷ Gumbel-softmax + STE で勾配を伝播

Algorithm 1 Training of XBMs

Require: Training dataset $\mathcal{D}$, vision encoder h_ψ , text decoder g_ϕ , classifier f_θ , pre-trained parameters $(\psi_p, \phi_p, \theta_p)$, training batchsize B , step size η , trade-off parameter λ

Ensure: Trained models $(h_\psi, g_\phi, f_\theta)$

```
1:  $\theta \leftarrow \theta_p, \phi \leftarrow \phi_p, \psi \leftarrow \psi_p$  # Initialize parameters
2: while not converged do
3:    $\{(x^i, y^i)\}_{i=1}^B \sim \mathcal{D}$ 
4:    $\{e_p^i\}_i^B \leftarrow \{\text{generate}(g_p, h_p(x^i))\}_i^B$  # Generating reference explanation
5:    $\{\hat{e}^i\}_i^B \leftarrow \{\text{g\_sampling}(g_\phi, h_\psi(x^i))\}_i^B$  # Gumbel-softmax sampling
6:   # Computing batch-mean losses
7:    $\mathcal{L}_{\text{cls}}^B \leftarrow \frac{1}{B} \sum_{i=1}^B \ell_{\text{CE}}(f_\theta(h_\psi(x^i), \hat{e}^i), y^i)$ 
8:    $\mathcal{R}_{\text{int}}^B \leftarrow \frac{1}{B} \sum_{i=1}^B \ell_{\text{CE}}(g_\phi \circ h_\psi(x^i), e_p^i)$ 
9:   # Updating parameters through backprop.
10:   $\theta \leftarrow \theta - \eta \nabla_\theta (\mathcal{L}_{\text{cls}}^B + \lambda \mathcal{R}_{\text{int}}^B)$ 
11:   $\phi \leftarrow \phi - \eta \nabla_\phi (\mathcal{L}_{\text{cls}}^B + \lambda \mathcal{R}_{\text{int}}^B)$ 
12:   $\psi \leftarrow \psi - \eta \nabla_\psi (\mathcal{L}_{\text{cls}}^B + \lambda \mathcal{R}_{\text{int}}^B)$ 
13: end while
```

❖ XBM は三種類の説明形式を提供

- ▷ 説明テキスト $\hat{e}$ $\rightarrow$ 訓練したデコーダ g_ϕ の出力 (標準)
- ▷ コンセプトフレーズ c $\rightarrow$ 説明テキスト $\hat{e}$ を NLP パーサで分解
- ▷ ヒートマップ $\rightarrow$ テキスト分類器 f_θ 内の画像-テキスト対応 (画像入力時のみ)



$\hat{e}$ = "A white large truck
parked in a parking lot
with a building in the
background"

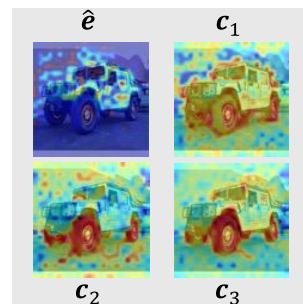
(i) Text Explanation

c_1 = "A white large truck" (0.48)

c_2 = "a parking lot" (0.23)

c_3 = "a building in the background" (0.20)

(ii) Concept Phrase w/ Score



(iii) Cross-Attention Heatmap

Experiments

❖ XBM の実装

- ▶ 事前学習マルチモーダルエンコーダ・デコーダ: BLIP (or LLaVA-v1.5)
- ▶ テキスト分類器: BERT

❖ ベースライン

- ▶ Fine-tuned BLIP-ViT: ブラックボックス (解釈性なし)
- ▶ Label-free CBM [6]: CLIP ベースの CBM
- ▶ Frozen-BLIP: テキスト分類器 f_θ のみを更新する差分 (一般キャプション → ラベル予測)

❖ データセット: Aircraft / Bird / Car / ImageNet (画像データセット)

❖ テスト分類精度 (%)

❖ CLIP-Score

- ▶ 入力画像と説明テキストのコサイン類似度を CLIP 空間上で計測
- ▶ 入力画像と説明の意味的な関連性を評価

❖ GPT-2 Perplexity

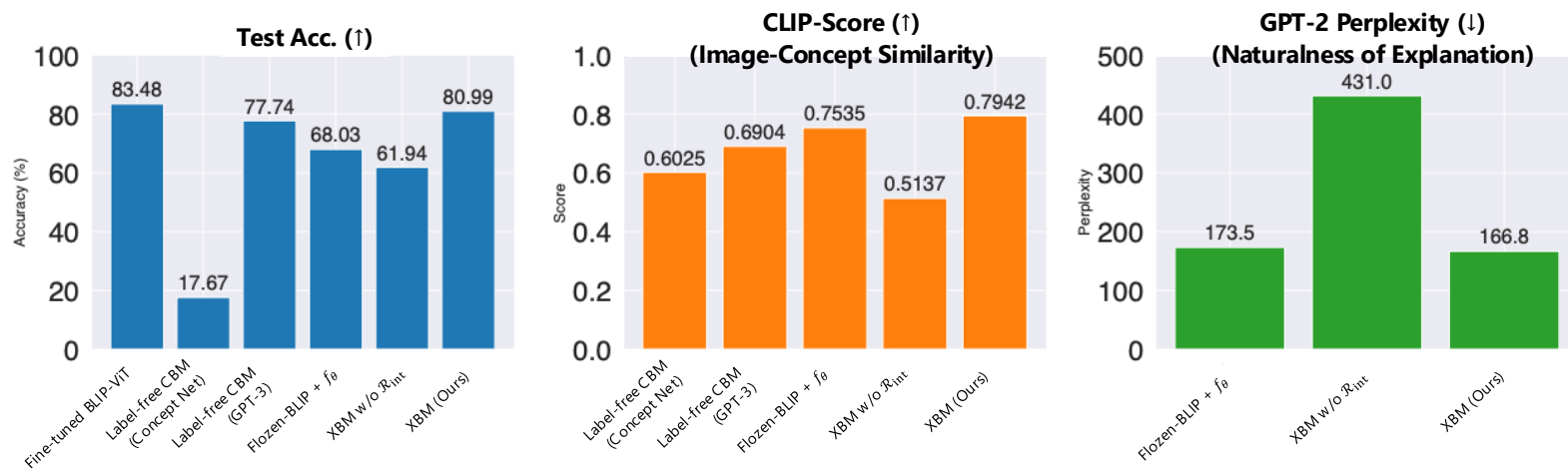
- ▶ 自然な言語分布を持つ GPT-2 を参照としたときの説明テキストの流暢性
- ▶ 生成テキストを GPT-2 に入力し、トークンごとの尤度から計算

$$\text{Perplexity} = \exp\left(-\frac{1}{N} \sum_n \sum_k \mathbb{I}_{n=k} \log(p_{\text{GPT-2}}(t_{n,k}))\right)$$

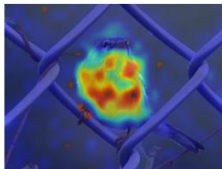
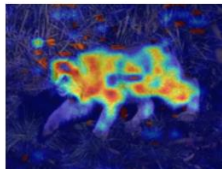
❖ XBM は CBM ベースラインをテスト精度と CLIP-Score で上回る

- ▷ Frozen-BLIP < XBM → 最適化によって精度・解釈性の両方が改善
- ▷ XBM w/o $\mathcal{R}_{int}$ < XBM → 説明蒸留損失は説明テキストの自然さを大きく改善

Performance Evaluation on Target Dataset: Bird (CUB)



❖ XBM による説明テキストは予測ラベルの内容を具体的に記述

	Bird (Yellow Bellied Flycatcher) 	ImageNet (Lynx) 
Pre-trained BLIP (Caption)	A bird perched on a wire fence with leaves on the ground and a blurry background.	Cat walking through the grass in the woods at night with it's eyes open.
Label-free CBMs (Top-3 Concept)	olive-colored sides (0.77) green head (0.55) a small, green body (0.52)	feline (0.98) long, sharp claws (0.53) mau (0.17)
XBMs w/o $\mathcal{R}_{int}$ (Text Explanation)	222222222222 222222222222	when when when when when when when when when when
XBMs (Text Explanation)	A small green and yellow bird perched on a wire fence with leaves on the side.	Furry feline walking in the woods at night with its eyes open and one paw on the ground.
XBMs (Top-3 Concept Phrase)	a small green and yellow bird (0.39) leaves on the side (0.32) a wire fence (0.21)	Furry feline (0.39) one paw on the ground (0.19) the woods (0.17)
XBMs (Cross-Attn. Heatmap)		

$\mathcal{R}_{int}$ は自然な
テキスト生成に貢献

– コンセプトフレーズによって
CBM 風の出力も可能

❖ 一般的なテキストから徐々に分類に必要な詳細特徴に集中

Input (California Gull)	Epoch 0	Epoch 20	Epoch 40
	"Someone standing on a rock in front of the water"	"A seagull standing on a rock by the water's edge"	"A seagull with a yellow beak standing on beach"
	"Animals that are standing on the sand near the water"	"A seagull standing on a beach next to a bunch of sea lions"	"A seagull with a beak on a beach next to a group of sea lions"
Top-10 Word Occurrence	seagull, standing, water, beach, looking, back, body, sand, rock, sky	seagull, standing, water, beach, back, body, sky, sand, rock, grass	seagull, standing, water, beak , beach, back, body, yellow , sky, mouth

❖ Explanation Bottleneck Models (XBMs)

- ▶ 概要: 説明テキストを中間出力に持つ新しい解釈可能モデルを提案
→ 事前定義コンセプトを超えた柔軟な語彙による説明生成を可能に
- ▶ 工夫: 説明蒸留損失によって自然言語の文法構造や自然さを維持
- ▶ 結果: 最適化によって精度・解釈性の両方を改善
- ▶ 示唆: コンセプトのような単純なラベルではなく、テキストによる柔軟な解釈性を持つなDNNを構築することが可能
- ▶ 今後: 解釈可能なマルチモーダル LLM の実現

