The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

画像テキスト検索における不確かさの評価

Evaluation uncertainties in Image-Caption Retrieval

濱 健太　松原 崇　上原 邦昭
Hama Kenta　Matsubara Takashi　Uehara Kuniaki

神戸大学 大学院システム情報学研究科
Graduate School of System Infomatics, Kobe University

Deep learning algorithms are able to learn powerful representations for many tasks. These models' outputs are often taken blindly and assumed to be accurate, however, which are not always the case. This blind assumption causes many issues such as AI unsafety and social bias. Therefore, a meaningful measure of uncertainty is essential. It has been shown that Monte Carlo (MC) Dropout can model epistemic uncertainty, and enhance model performance in machine learning tasks. In this paper, we propose an evaluation method of uncertainty in image caption retrieval and verified its significance by qualitative evaluation. Also, we show that a learning model using MC Dropout improves accuracy in image caption retrieval.

1. はじめに

機械学習のアルゴリズムやシステムの発展と共に, その手法を適用したアプリケーションが様々な分野で成果を挙げている. これらのアルゴリズムの多くは, 人間には解釈できない入力と出力の間の複雑なマッピングを学習する. そのため精度が十分得られていれば, 実用上のアプリケーションでは出力結果を正しいものと仮定して利用するしかない. 異なる入力に対して, モデルがどの程度確信を持って結果を出力しているのかを評価するのは困難である. しかし安全性や社会的問題という観点から見ても, 実用上のアプリケーションでモデルの出力の不確かさを, 評価することは重要である. [Kendall 17] は自動画像分類アプリが人種差別的な画像分類を行なって, 問題となった例をあげている. 我々は精度が良いことに加え, 出力の不確かさを正しく評価できるシステムを必要としている. もし十分に不確かさを評価できれば, 上記のような問題を回避することができる.

機械学習のモデルの不確かさはその要因によって分類され, 様々な定義が存在する. 本研究では, モデルのパラメタや構造選択に関わる不確かさ (epistemic uncertainty) と, 出力に対するモデルの確信度合いを表す予測分布のエントロピー (predictive entropy) を扱う. epistemic uncertainty は十分な学習データがあれば減少し, モデルが持つ不確かさと捉えられる [Kendall 17]. Bayesian neural networks (BNN) [MacKay 92] は epistemic uncertainty を評価する方法の一つである. BNN はニューラルネットワークの重みがある事前分布に従うと仮定する. そこで, ネットワークの重みを何度もサンプルし, その出力のブレを見ることで epistemic uncertainty を評価できる [Gal 16].

近年, 様々な機械学習のモデルで不確かさを評価する研究がされている. 画像認識の分野では, セグメンテーションや深度推定への適用 [Kendall 15,Kendall 17] や, 自然言語処理の様々なタスクへの適用 [Xiao 18], また異常検知 [Choi 18], 行動認識 [Subedar 18], 時系列データ解析 [Zhu 17] など多岐に渡る. しかし, 画像テキスト検索においては, その双方向性から入力となるクエリと, 出力となるターゲットの両方の不確かさを考えることができる点で他のタスクと異なる性質を持っている. クエリ側の不確かさの評価結果から, より良い検索文を入力するようユーザに促す, システムの学習データの不足を検知し管理者に知らせる. ターゲット側の不確かさの評価結果から, 表示する検索結果を変えて精度の改善を測る等, 画像テキスト検索では不確かさに関わる情報はシステムの改善にとって重要である. 本研究は, 画像テキスト間検索における不確かさの評価方法を提案し, その不確かさが捉えた意味を定性的に評価し検証した. また提案手法が, 画像テキスト間検索で精度を向上させたことを示す.

連絡先: 濱 健太, 神戸大学 大学院システム情報学研究科, hamaken@ai.cs.kobe-u.ac.jp

2. 関連研究

2.1 Bayesian Neural Network (BNN)

ニューラルネットワークは重み W をパラメタとして持つ. 教師あり学習では, データセット $D = (x_i, y_i)_{i=1}^N$ に対して, 何らかの目的関数を最大化する W を推定するよう学習する. ベイズ推定では W は, 事前分布 $p(W)$ に従うと仮定し, 最適なパラメタ W ではなく, D に対する最適な事後分布 $p(W|D)$ を推定することを目標としている. またモデルの出力を $f^W(x)$ とすると, BNN の予測値 y は推定した事後分布 $p(W|D)$ から得られる W で, $f^W(x)$ を周辺化した値となる.

しかし, 近年の深層ニューラルネットワークは高次元のパラメタ W を持ち, 構造も非線形で複雑なため, 事後分布を計算できない. そこで, 様々な近似計算の方法が提案されている. 特に, Monte Carlo dropout (MC dropout) [Gal 16] は元のモデルに対する大きな変更なしで利用できる. MC dropout は, ネットワークの非線形層の間に Dropout 層を追加し, モデルの評価時に Dropout を適用しながら複数回ネットワークから出力を得て, その平均をモデルの出力として利用する. [Gal 16] では, MC dropout があるモデルの変分ベイズ法の近似計算であることが数学的に示されている. Dropout のマスク行列をサンプルして得られるモデルの出力は, 事後分布 $p(W|D)$ からサンプルした W を用いて計算した $f^W(x)$ と解釈できる. また, モデルの不確かさはこの方法を繰り返して得られた複数の出力の分散を見ると, 近似的に評価できる.

2.2 Visual Semantic Embedding++

画像テキスト検索における一般的な手法として，Visual-Semantic Embeddings (VSE) [Kiros 14] がある．VSE は，CNN で画像の特徴量を抽出し，Recurrent Neural Network(RNN) でテキストの特徴量を抽出し，それらを類似度の計算が可能な共通の空間に線形変換で埋め込む．本研究では，ベースラインとして VSE++ [Faghri 17] と呼ばれる，VSE の損失関数を改良し精度を向上させたモデルを用いる．この VSE++では画像の特徴量は ImageNet で学習済みの VGG か ResNet を，テキストは LSTM を用いて抽出する．

VSE++の損失関数について説明する．入力画像とその特徴量をそれぞれ x, h とし，入力テキストとその特徴量をそれぞれ y, t とする．ここで，画像の特徴量とテキストの特徴量を共通空間に埋め込む行列を M_x, M_y とすると，共通空間上の画像，テキストの表現は $z_x = M_x h$, $z_y = M_y t$ という形で得られる．これらの表現の類似度は VSE++では以下の cos 類似度を用いて計算される．

$$\mathrm{sim}(z_x, z_y) = \frac{z_x \cdot z_y}{\|z_x\|_2 \|z_y\|_2}$$

VSE++は次の rank-loss と呼ばれる式を用いて損失関数を定義する．

$$\begin{aligned} r(x, y) = &\max_{\hat{y}} \max\{0, \alpha - \mathrm{sim}(z_x, z_y) + \mathrm{sim}(z_x, \hat{z_y})\} \\ &+ \max_{\hat{x}} \max\{0, \alpha - \mathrm{sim}(z_x, z_y) + \mathrm{sim}(\hat{z_x}, z_y)\} \end{aligned} \tag{1}$$

$\hat{z_x}, \hat{z_y}$ は入力データ x, y と関連しないデータ (負例) の埋め込み点を意味する．マージン α は関連するデータ (正例) と負例の類似度の差を調節するハイパーパラメタである．rank-loss を用いて，VSE++の損失関数は以下のように定義される．

$$L = \frac{1}{N} \sum_{n=1}^{N} r(x_n, y_n)$$

ただし，N は訓練データ数である．

3. 提案手法

3.1 不確かさの定義

モデルの出力の不確かさは，その要因によって大きく 2 種類に分類される [Kendall 17]．入力データセットそのものの複雑さや，観測ノイズなどが原因とされるそれ以上減らすことのできない不確かさ (aleatoric uncertainty) と，モデルへの入力データの不足や，それによる学習不足が原因とされる不確かさ (epistemic uncertainty) である．本研究では epistemic uncertainty とモデルの出力に対する確信度合いである predictive entropy の二つを取り扱う．まずは [Kendall 17] による，epistemic uncertainty の定義を説明する．BNN ではデータセット $D = (x_i, y_i)_{i=1}^N$ が与えられたときの，重み $\mathbf{W}$ の事後分布を推定する．扱う問題が分類問題の場合，入力 x から得られる予測値 y は次のようにして得られる．

$$y|\mathbf{W} \sim \mathrm{Categorical}(\mathrm{Softmax}(f^{\mathbf{W}}(x)) \tag{2}$$

事後分布 $p(\mathbf{W}|D)$ が与えられた際，新たな入力 x^* が，クラス y^* に分類される確率は次の式で与えられる．

$$p(y^*|x^*, D) = \int p(y^*|f^{\mathbf{W}}(x^*)) p(\mathbf{W}|D) d\mathbf{W} \tag{3}$$

ほとんどの場合，事後分布の計算が解析的に不可能なため，変分ベイズ方が用いられる．変分ベイズ法では，パラメタ θ を用いて真の分布 $p(\mathbf{W}|D)$ を近似するための分布 $q_\theta(\mathbf{W})$ を置き，これら二つの分布を Kullback-Leibler (KL) divergence の最小化によって近づける．変分ベイズ法は第 2.1 項で述べた用に，MC Dropout で近似できる．本研究では Epistemic Uncertainty の評価に MC Dropout を用いる．モデルの評価時，最適化された事後分布の近似である $q_\theta(\mathbf{W})$ を用いて，式 (4) の $p(\mathbf{W}|D)$ と置き換えて，M 回のサンプルによるモンテカルロ積分により予測分布は次のように得られる．

$$p(y^*|x^*, D) \approx \frac{1}{M} \sum_{j=1}^{M} \mathrm{softmax}(f^{\mathbf{W}'_j}(x^*)) \tag{4}$$

ここで，epistemic uncertainty $U_m(x^*)$ は次のように定義する．

$$U_m(x^*) \approx \frac{1}{C} \sum_{y=1}^{C} \frac{1}{M} \sum_{j=1}^{M} \mathrm{softmax}(f^{\mathbf{W}'_j}(x^*))^2 - \mathbb{E}(y|x^*)^2 \tag{5}$$

C はクラス数，$\mathbf{W}'_j$ は分布 $q(\mathbf{W})$ からサンプルされたモデルのパラメタである．また predictive entropy は，モデルの予測分布 (4) のエントロピーとして定義する．

3.2 Image-Caption Retrieval

この節では，画像テキスト間検索というタスクを確率論の枠組みで捉え直す．初めに，議論を簡単にするため，一つのクエリに対してターゲットが正例と負例の二つのみの場合を考える．クエリを x，正例を y，負例を y' とする．このとき，点 x が状態 y または y' を取る確率が，ボルツマン分布に従うと仮定する．点 x が，状態 y を取る確率は次のように表される．

$$p(y|x) = \frac{e^{-E_y}}{e^{-E_y} + e^{-E_{y'}}} \tag{6}$$

ここで，点 x に対する状態 y が与えられている場合，対数尤度 $\log p(y|x)$ の最大化によって分布を学習できる．$\log p(y|x)$ は次のような式で表される．

$$\begin{aligned} \log p(y|x) &= \log \frac{1}{1 + e^{-E_{y'} + E_y}} \\ &= -\mathrm{softplus}(-E_{y'} + E_y) \end{aligned}$$

従って，対数尤度 $\log p(y|x)$ の最大化は，式 $\mathrm{softplus}(E_{y'} - E_y)$ を最小化することと等しい．この softplus 関数 (logistic loss) と，$\max(0, x)$ (hinge loss) は互いに近似としてよく用いられる．また 2 値分類問題において，目的関数に logistic loss もしくは hinge loss を用いた場合の識別境界が似た形になることは知られている [Bishop 06]．ここで，エネルギー関数 E_y を $E_y = -\mathrm{sim}(x, y)$ と仮定すると，$\log p(y|x)$ の最大化は式 (1) の最小化とほとんど等価である．クエリとターゲットのドメインを入れ替えることで，式 (1) の第二項目に関しても同様のことが言える．換言すれば，上記の確率場における対数尤度の最大化は，VSE の rank-loss の最小化と等しいと解釈できる．

次に，ターゲットの数が 2 つ以上の場合を考える．この場合，点 x が状態 y を取る確率は次のように表される．

$$p(y|\mathbf{x}) = \mathrm{softmax}(E_y) \tag{7}$$

正解となる状態 y が与えられている場合，この問題は多クラス分類問題として解釈される．2 値分類問題における softplus

表 1: Results on the MSCOCO dataset

Model	Dropout VGG	Dropout GRU	#models	Caption Retrieval R@1	R@5	R@10	Med r	Mean r	Image Retrieval R@1	R@5	R@10	Med r	Mean r
VSE++(baseline)	○	×	1	49.6	78.6	88.2	1.8	6.0	37.3	72.2	84.3	**2.0**	9.2
+ dropout	○	○	1	50.1	78.6	87.8	**1.6**	6.2	**38.1**	**73.1**	**85.2**	**2.0**	**8.5**
+ MC dropout	○	○	1	43.2	73.7	84.2	2.0	8.0	31.2	65.1	78.6	3.0	13.9
+ MC dropout	○	○	10	49.3	79.2	**88.6**	**1.6**	6.2	37.4	72.3	84.8	**2.0**	8.7
+ MC dropout	○	○	30	49.8	79.1	**88.6**	1.8	**5.9**	37.9	72.9	85.0	**2.0**	8.4
+ MC dropout	○	○	50	**50.4**	**79.5**	88.5	**1.6**	**5.9**	37.5	73.0	**85.2**	**2.0**	**8.5**

関数と hinge loss の関係は, softmax 関数と multi-class SVM の損失関数との関係と同じである. この研究では, 多クラス分類問題における対数尤度最大化と, VSE の rank-loss 最小化がほとんど等価であるという考えに基づいて, 画像テキスト間検索におけるクエリ x^* に対する epistemic uncertainty を次のように定義する.

$$U_m(x^*) \approx \frac{1}{C}\sum_{y=1}^{C}\frac{1}{M}\sum_{j=1}^{M}\mathrm{softmax}(\mathrm{E_{j,y}})^2 - \mathbb{E}(\mathrm{y}|\mathrm{x}^*)^2 \qquad (8)$$

ただし, $\mathbb{E}(y|x^*) \approx \frac{1}{M}\sum_{j=1}^{M}\mathrm{softmax}(\mathrm{E_{j,y}})$, $E_{j,y} = \mathrm{sim}(z_{x_j}, z_{y_j})$. また, z_{x_j}, z_{y_j} はそれぞれ, ドロップアウトを用いた x, y の j 回目の埋め込み表現である.

4. 実験・結果

MC dropout における検索精度を確認するため, 次の設定で実験を行なった. 実験には Microsoft COCO (MS COCO) [Lin 14] データセットを用いた. またデータの分割は [Faghri 17] と同様, 訓練用画像 113,287 枚, 検証用画像 1,000 枚, 評価用画像 5,000 枚にした. また評価は画像 1,000 枚ずつのスコア 5 回分の平均を用いた. 訓練時には入力画像に大きさ 224×224 のランダムクロップを適用し, 評価は大きさ 224×224 でセンタークロップした画像で行なった.

本研究では, VSE++のテキスト側のエンコーダである GRU への入力と出力に, ドロップアウト層を適用する. ただし, 画像側のエンコーダである VGG はすでにドロップアウトが適用済みであるため, 新しく層は追加しない. 共通空間の次元は 2048 次元で, 単語の埋め込み次元は 600 とし, ドロップアウト率は 0.1 で学習を行なった. その他のパラメタや最適化アルゴリズムは全て [Faghri 17] のものを使用した.

評価指標は, 画像テキスト検索において一般的な R@1, R@5, R@10, Med r, Mean r を用いる. R@k は, 評価データ中全てのクエリに対して, 正解となるターゲットが検索結果の k 番目以内に含まれる割合を表す. Med r は, 全てのクエリの正解となるターゲットの順位の中央値であり,Mean r は平均である.

結果は表 1 のようになった. Baseline はドロップアウト層を GRU に追加していない通常の VSE++であり, Baseline + ドロップアウトは, GRU にドロップアウトを追加し評価時は全ての重みを使用する (weight averaging) モデルである. そして提案手法である Baseline+MC dropout は, 評価時に MC dropout を用いた. 表 1 から分かるように, baseline + ドロップアウトは画像検索の精度を向上させ, 提案手法はさらにそこから, テキスト検索の精度を向上させている. また, 評価時のサンプル数が大きいほど精度が向上することが確認できた.

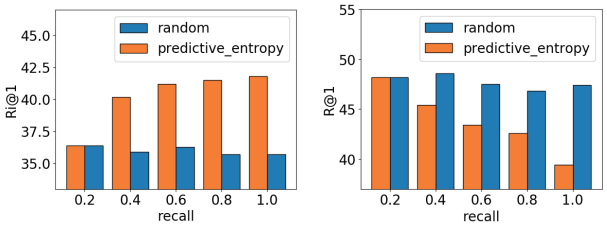

図 1: Left is R@1 for image retrieval. Right is Ri@1 for text retrieval. Horizontal axis means the rate of decreasing the query.

5. 検証

この節では, 定義した不確かさが検索タスクにおいて持つ意味を検証する. predictive entropy は予測結果に対するモデルの確信の度合いを表す. 図 1 は評価データのクエリから predictive entropy の大きいデータを除去しながら, R@1 の値をプロットしたものである. テキスト検索 (左) の場合, 除去する割合が大きくなるほど精度が向上しているため, predictive entropy がクエリーからターゲットを予測することの難しさを捉えていると考えられる. しかし, 画像検索 (右) の場合は精度が低下している. 画像からテキストへの検索は訓練データセットの構造上, 一つの画像に対する正解のテキストが 5 つ存在し, rank-loss は全てに画像が近くように学習を行う. predictive entropy が小さいデータはどれか一つのテキストに対して画像が過度に近づくなどして, 汎化学習が上手くできていない可能性がある. 次に, 評価用のデータセット内で predictive entropy が大きい入力データと小さい入力データを表示した. 図 2 は上段は predictive entropy が最大のものを左から順に, 下段は小さい順に表示している. predictive entropy が大きい画像は熊や馬などの動物が写っている画像が多いが, 小さい画像は人が写っているものが多い. これは画像認識において, 人の画像は動物よりも学習時に多く与えられるため, 識別性能が高いためだと考えられる. また表 2 は, predictive entropy の大きいテキストと小さいテキストの例である. 大きいテキストは, 具体性の低い文章であるが, 小さいテキストは状況を具体的に説明しているテキストになっている.

次に, 定義した epistemic uncertainty が学習不足を検知するかどうかを検証する. 図 3 は訓練データに MS COCO を用いて, 評価データに flickr30k [Young 14] のデータ 1000 枚を用いた場合と, MS COCO データ 1000 枚を用いた場合の入力データの epistemic uncertainty のヒストグラムである. 評価データに学習データと異なる flickr30k を用いた場合の方が, epistemic uncertainty が大きいことから, この不確かさが学習の不足を捉えていることが確認できる.

図 2: Examples of input images in MSCOCO dataset with high and row data uncertainties. The top row shows top 5 examples with high epistemic uncertainty. The bottom row shows bottom 5 examples.

表 2: Examples of input texts in MSCOCO dataset with high or low data uncertainties.

highly uncertain texts
· A person is doing something that is quite fun.
· A person is separated from everyone else in the picture and doing something there.
lowly uncertain texts
· A bench is sitting amongst the trees and the bushes.
· A boat floating on a lake next to a shore.

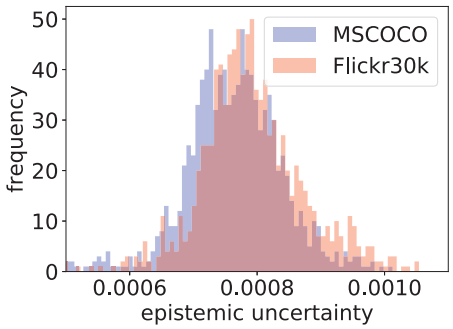

図 3: Distribution of epistemic uncertainty. Blue denotes MSCOCO, and red denotes Flickr30k.

6. 結論

本研究では, 検索タスクにおける不確かさの指標として epistemic uncertainty と predictive entropy を評価する方法を提案し, それらが意味のある指標であることを定性的実験により確かめた. また, 不確かさの評価のため用いた MC dropout という手法が, 検索タスクにおいて精度を向上させることを確認した. 今後の課題として, 他の不確かさの定義の検討や, 本研究で得られた不確かさの指標を用いて検索の実性能を向上させる方法の考案, 不確かさを用いた学習率などのパラメタの自動調節などを考えている. 本研究は総務省 SCOPE(受付番号 172107101) の委託を受けて行われた.

参考文献

[Bishop 06] Bishop, C. M.: *Pattern Recognition and Machine Learning*, Springer (2006)

[Choi 18] Choi, H. and Jang, E.: Generative ensembles for robust anomaly detection, *arXiv* (2018)

[Faghri 17] Faghri, F., Fleet, D. J., Kiros, R., *et al.*: VSE++: Improved Visual-Semantic Embeddings, *arXiv* (2017)

[Gal 16] Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in *ICML* (2016)

[Kendall 15] Kendall, A., Badrinarayanan, V., and Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, *arXiv* (2015)

[Kendall 17] Kendall, A. and Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision?, in *NIPS* (2017)

[Kiros 14] Kiros, R., Salakhutdinov, R., and Zemel, R. S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, *arXiv* (2014)

[Lin 14] Lin, T., Maire, M., Belongie, S. J., Hays, J., *et al.* : Microsoft COCO: Common Objects in Context, *ECCV* (2014)

[MacKay 92] MacKay, D. J.: A practical Bayesian framework for backpropagation networks, *Neural computation* (1992)

[Subedar 18] Subedar, M., Krishnan, R., Meyer, P. L., *et al.*: Uncertainty aware multimodal activity recognition with Bayesian inference, *arXiv* (2018)

[Tang 13] Tang, Y.: Deep learning using linear support vector machines, *arXiv* (2013)

[Xiao 18] Xiao, Y. and Wang, W. Y.: Quantifying Uncertainties in Natural Language Processing Tasks, *arXiv* (2018)

[Young 14] Young, P., Lai, A., Hodosh, M., *et al.*: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Transactions of the Association for Computational Linguistics* (2014)

[Zhu 17] Zhu, L. and Laptev, N.: Deep and confident prediction for time series at uber, in *ICDMW, 2017 IEEE International Conference on* (2017)