The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

確率分布を用いた画像テキストデータの埋め込みと検索

Embedding and retrieval of images and text data using probability distribution

濱 健太 *1　松原 崇 *1　上原 邦昭 *1
Kenta Hama　Takashi Matsubara　Kuniaki Uehara

*1 神戸大学 大学院システム情報学研究科計算科学専攻
Graduate School of System Infomatics, Kobe University

Multimodal data including images, sounds, texts is accumulated on the Internet. We can expect general-purpose data representation to perform tasks such as data discrimination, generation, and retrieval on various modalities datasets. The key idea for acquiring the representation is embedding a point from a data space of each modality in a point of common space. However, if data is embedded in a point, it becomes difficult to interpret the ambiguity of the data's meaning and the inclusive relation among the data. Of course, representation of data point does not necessarily need to be a point. In this study, we embed image and text into a normal distribution in a common space. This improves the performance of image retrieval.

1. はじめに

マルチモーダルデータの蓄積が進み, モダリティの異なるデータ間で, 相互の検索や変換を行い, 情報の理解, 要約をサポートするシステムの研究が盛んに行われている [Wang 16]. 複数のモダリティの異なるデータを扱う上では, 各モダリティを横断するデータの分類, 生成, 検索といったタスクに転用が可能な, 汎用的なデータ表現を獲得することが重要である. その表現獲得の核となるアイデアとして, 各モダリティのデータからそれら共通の表現空間上の点への埋め込みがある. 埋め込みとは, あるデータ x をある関数 f によって, データの意味を表現する空間上のベクトル $f(x)$ に写像することである. さらに, 埋め込まれた空間上で類似度を定義すれば, マルチモーダルデータ間の意味の近さの測定が可能になる. この空間上では意味の近いデータ同士の類似度が高いことが望ましいが, データ間の類似度を正しく定義されているか評価するのは困難である. そのため, 埋め込まれた空間の評価に, 共通空間上でのデータ間の検索の精度を利用することが提案されている [Kiros 14]. これは, 検索精度の良い共通空間は, 意味の近いデータを, 適切な類似度の点に埋め込んでいるという考え方に基づいている.

このようなマルチモーダル表現の獲得を目的として, 現在, 盛んに研究されている対象として, 画像とテキストデータ間の検索が挙げられる. 画像, テキストデータはインターネット上から容易に収集でき, 画像処理や自然言語処理の分野における研究結果も多く存在するため, 研究対象として選び易い.

画像とテキストデータ間の検索では, まず各データの特徴抽出を行い, その特徴量を何らかの基準で共通空間上のベクトルに変換する. 例えば, 画像の特徴抽出には ImageNet [Deng 09] で学習済みの CNN (Convolutional Neural Network) の中間層の出力を用いるものや, テキストの特徴抽出には word2vec [Mikolov 13] や LSTM (Long Short-Term Memory) の出力が用いられる. そして, これらの特徴量を rank-loss と呼ばれるマージンを用いた関数を使い, クエリーに対する望ましい出力ほど類似度が高くなるように埋め込むことが多い [Kiros 14]. しかし, 画像とテキストデータを共通空間上の点へ埋め込むと, 入力データの持つ意味の曖昧さや, データ間の意味の包含関係といった概念の解釈が困難になる. そのため, データが埋め込まれるべき箇所が複数あるような状況で問題が発生しやすい.

例えば, 単語の埋め込みを考える場合, “rock” という単語は “stone”,“river” などの自然を意味する単語と近い関係にあるのと同時に, “pop”, “jazz” などの音楽を意味する単語との意味も近いはずである. もしもデータを点へ埋め込む場合, “rock” という単語を挟んで, 自然を意味する単語全体と音楽を意味する単語全体が近くの点へ埋め込まれてしまうという問題が生じる. しかし, データの持つ意味の曖昧さを表現できる埋め込み手法があれば, “rock” という単語の曖昧さを上手く捉え, 周囲の単語に対する影響を軽減することが期待される.

本研究では, テキストと画像の両方を点ではなく, 確率分布へ埋め込む方法を提案する. テキストと画像を確率分布へと埋め込むことができれば, テキストと画像の持つ意味の曖昧さや, マルチモーダルデータ間の意味の包含関係や共通部分といった概念を考えることも可能となる. 本研究ではこの確率分布への埋め込みを行う提案手法を, 検索タスクの state-of-the-art である VSE++ [Faghri 17] と組み合わせて比較し, 類似度の定義として望ましいことを示す.

2. 関連研究

2.1 VSE (Visual Semantic Embeddings)

ニューラルネットを用いて得られた特徴量から, 共通空間上の点へ埋め込むモデルとして, VSE [Kiros 14] について説明する. VSE は, 画像から説明文を生成するタスク (Image Captioning) を行うためのモデルである. 画像の特徴抽出には ImageNet で学習済みの AlexNet [Krizhevsky 12], テキストの特徴抽出には LSTM を用いて, 得られた特徴量は線形変換によって共通空間上のベクトルへ変換される.

まず, 画像データ x, 画像の特徴量 h, テキストデータ y, テキストの特徴量 t とする. 画像の特徴量は CNN の最終層から 1 層前の出力で, テキストの特徴量は LSTM の中間層の出力とし, $h = \mathrm{CNN}(x)$, $t = \mathrm{LSTM}$ と表すことにする. 次に, これらの特徴量を線形変換し, 共通の表現空間に埋め込む. この埋め込みに用いる行列を各データに対して M_x, M_y とすると, 埋め込まれた共通空間上のベクトルは $z_h = M_x h$, $z_t = M_y t$ と表される. もし, これらのデータ x, y が関連するデータであ

連絡先: 濱健太, 神戸大学大学院システム情報学研究科計算科学専攻, hamaken@ai.cs.kobe-u.ac.jp

れば, 共通空間上のベクトル z_h, z_t の類似度も高くなることが望ましい. VSE では類似度の定義に cos 類似度を用いている. cos 類似度は

$$\mathrm{sim}(x, y) = \frac{z_h \cdot z_t}{\|z_h\|_2 \|z_t\|_2}$$

ただし, $z_h = M_x \mathrm{CNN}(x), z_t = M_y \mathrm{LSTM}(y)$ と定義される. 関連するデータ対 z_h, z_t に関して, $\mathrm{sim}(x, y)$ が大きくなるように埋め込むために, VSE では rank-loss と呼ばれる以下の式を用いている.

$$r(x, y) = \sum_{\hat{y}} \max\{0, \alpha - \mathrm{sim}(x, y) + \mathrm{sim}(x, \hat{y})\} + \sum_{\hat{x}} \max\{0, \alpha - \mathrm{sim}(x, y) + \mathrm{sim}(\hat{x}, y)\}$$

$\hat{y}, \hat{x}$ はそれぞれ x, y に関連しないデータを表すものとする. α はマージンと呼ばれるハイパーパラメータであり, 関連するデータ対の類似度と, 関連しないデータ対の類似度の差をどの程度大きくするかを調節する役割を担っている. データ全体における損失関数は rank-loss を用いて,

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

と定義する. ただし, N はデータ数を表すものとする. 定義した損失関数を最小化するように学習を行えば, 共通空間への埋め込みが実現される.

2.2 VSE++

VSE++ (Improved Visual Semantic Embeddings) [Faghri 17] は VSE の改良として, 提案されたものである. VSE の rank-loss を以下のように変更している.

$$r(x, y) = \max_{\hat{y}} \max\{0, \alpha - \mathrm{sim}(x, y) + \mathrm{sim}(x, \hat{y})\} + \max_{\hat{x}} \max\{0, \alpha - \mathrm{sim}(x, y) + \mathrm{sim}(\hat{x}, y)\}$$

VSE では関連しないデータ点全てにおいての和をとっていた箇所が, 各データ点における損失の中でも最大の値となるデータ点のみを全体の損失に含めている. シンプルな変更であるが, VGG [Simonyan 14], ResNet [He 16] など画像の特徴抽出に使うモデルによらず精度が向上し, 別の埋め込み手法との併用も可能である. 本研究ではこの VSE++を基本モデルとし, 確率分布への埋め込みが可能となる改良を加える.

3. 提案手法

本研究では, 各入力画像, テキストデータを多変量正規分布に変換する. その際, 計算を簡単にするために, 正規分布の各次元に独立性を仮定し, 分散共分散行列は対角成分のみ持つことにしている. 提案手法のモデル全体は図 1 のようになっている.

まず入力データを VSE++と同様に, 画像は ImageNet で事前学習済みの CNN (VGG, ResNet) に入力し, テキストは LSTM に入力する. なお本研究における実装には GRU [Cho 14] を用いている. 各ニューラルネットワークを通して得られた特徴量は, 線形変換によって正規分布のパラメータである平均と分散に変換される. 埋め込みは $M_{x_\mu}, M_{x_\sigma}, M_{y_\mu}, M_{y_\sigma}$ の 4 つの行列を用いて行われる. これらの行列を用いて, 図 1 の入力画像データとテキストデータの平均と分散は, $\mu_i = M_{x_\mu} i, \sigma_i = \mathrm{diag}(M_{x_\sigma} i), \mu_t = M_{y_\mu} t, \sigma_t = \mathrm{diag}(M_{y_\sigma} t)$ と表される. よって, 画像 x の特徴量 i とテキスト y の特徴量 t の埋め込みは, $z_i = \mathcal{N}(\mu_i, \sigma_i)$, $z_t = \mathcal{N}(\mu_t, \sigma_t)$ という正規分布の形で表現される. 以上の操作と VSE++は, CNN, LSTM を用いて特徴抽出を行うところまでは同じで, その後, 特徴量を 1 つの点ではなく, 2 つの平均と, 分散を表現する点へ変換するように埋め込み行列を 2 つずつ用意する点で異なっている.

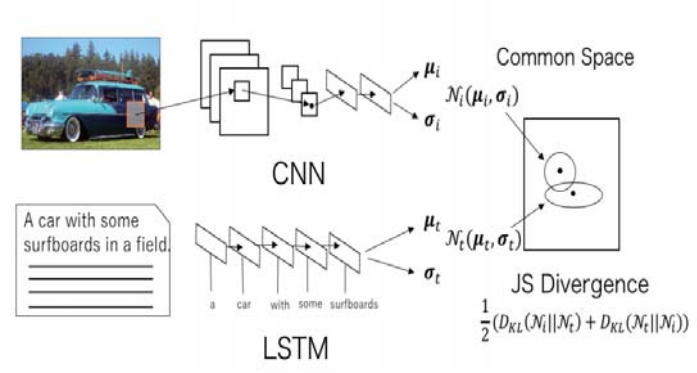

図 1: The whole model of the proposed method.

この埋め込みは, 画像, テキストの特徴量を n 次元実数空間上の点 $x \in R^n$ ではなく, 多変量正規分布の平均 $\mu \in R^n$, 分散共分散行列 $\Sigma \in R^{(n,n)}$(ただし $\Sigma = \mathrm{diag}(\sigma), \sigma \in R^n$) へと変換すれば, 関数空間上の点 $f \in \{\mathcal{N}(\mu, \Sigma) | \mu \in R^n, \Sigma = \mathrm{diag}(\sigma), \sigma \in R^n\}$ への埋め込みを行なっていると解釈できる. よって, VSE++の共通空間である R^n と空間そのものが異なるため, cos 類似度をそのまま利用することはできない. そこで, 確率分布間の類似度を定義するため, 確率分布間の差異を計算する際に用いられることの多い, JS ダイバージェンスを利用する.

JS ダイバージェンスは, KL ダイバージェンスを用いて定義される. 二つの確率分布を p, q とすると, KL ダイバージェンスは $D_{KL}(p||q) = \int p \log \frac{p}{q} dx$ で定義される. p,q は正規分布なので, $D_{KL}(p||q)$ は代数計算によって求めることができ, さらに各次元の独立性を仮定しているので, 計算も容易である. JS ダイバージェンスは KL ダイバージェンスを用いて

$$D_{JS}(p||q) = \frac{1}{2}(D_{KL}(p||q) + D_{KL}(q||p))$$

と定義される. ここで類似度は符号を反転し, $\mathrm{sim}(x, y) = -D_{JS}(p||q)$ と定義する. もちろん, 類似度の定義に KL ダイバージェンスを用いることも考えられる. しかし, 事前実験により, 類似度に KL ダイバージェンスより JS ダイバージェンスを用いた方が精度が良かったため, 本研究では JS ダイバージェンスによる埋め込みを行っている.

モデルの学習は二段階で行う. まず第一段階として, 画像データ, テキストデータの平均に対して, 通常の VSE++と同様に, cos 類似度を用いて埋め込みを行う. その後, 第二段階として, 各データの分散を学習する. この際, 一段階目で学習したテキストデータの平均を固定して学習を行う. これは, 事前実験によりデータの平均を固定すると, 固定せず学習した場合より精度の向上が見られたためである. また, 第二段階の分散の学習では先ほど定義した JS ダイバージェンスを用いた類似度を使う.

表 1: Effect of feature and fine-tuning.

	Caption Retrieval					Image Retrieval				
Model	R@1	R@5	R@10	Med r	Mean r	R@1	R@5	R@10	Med r	Mean r
VSE++(VGG)	52.0	82.0	89.9	**1.0**	5.3	39.6	74.5	85.7	**2.0**	9.6
Ours(VGG)	**52.6**	**82.3**	**90.5**	**1.0**	**5.2**	**40.5**	**75.0**	**86.2**	**2.0**	**8.5**
VSE++(ResNet)	59.2	86.1	93.2	**1.0**	4.0	43.8	78.0	88.2	**2.0**	7.6
Ours(ResNet)	**59.4**	**86.7**	**93.8**	**1.0**	**3.8**	**44.9**	**78.9**	**88.9**	**2.0**	**6.7**
VSE++(VGG, finetune)	57.6	86.0	93.0	**1.0**	**4.1**	45.2	79.1	**88.9**	**2.0**	8.3
Ours(VGG, finetune)	**58.4**	**86.5**	**93.6**	**1.0**	**4.1**	**45.4**	**79.3**	**88.9**	**2.0**	**7.5**
VSE++(ResNet, finetune)	**64.3**	89.5	95.9	**1.0**	**3.1**	50.0	**83.3**	**91.4**	1.6	6.0
Ours(ResNet, finetune)	**64.3**	**90**	**96.3**	**1.0**	**3.1**	**50.5**	**83.3**	**91.4**	**1.4**	**5.4**

4. 実験及び結果

学習データセットとして, Image Caption Retrieval で一般に用いられる, Microsoft COCO データセット [Lin 14](以下 MS COCO) を使用し, 画像テキスト間の双方向の検索を評価する. MS COCO は, 物体検知, セグメンテーション, キャプショニングのための大規模な画像とテキストのデータセットである. Image Caption Retrieval では, キャプショニング用のデータを用いる. キャプショニング用のデータは, 1 枚の画像に対して, その説明を行うテキスト (キャプション) が 5 つ与えられている. データの分割は, VSE++の論文と同様に, 訓練データに 113,287 枚, 検証データに 5000 枚, 評価データに 5000 枚の画像を使用している. 画像データは, CNN への入力時に画像全体からランダムに 224×224 の大きさで切り出している.

パラメータの最適化アルゴリズムには, Adam [Kingma 14] を用いている. ハイパーパラメータは, 一段階目のエポック数を 30, 学習率は最初の 15 エポックを 2.0×10^{-4}, 残りの 15 エポックを 2.0×10^{-5} に設定している. 二段階目の学習はエポック数を 30, 学習率は最初の 15 エポックを 2.0×10^{-5}, 残りの 15 エポックを 2.0×10^{-6} に設定している. 学習率が一段階目に対して 0.1 倍されているのは, 画像の平均の固定を行なっていないため, 一段階目で学習された平均の局所解を抜け出して, 別の局所解へ向かうことを防ぐためである.

また, 本研究では CNN の fine-tuning と提案手法を組み合わせた実験も行なっている. CNN の fine-tuning は, 一段階目が終了した後に CNN のパラメータの固定をはずし, エポック数を 15, 学習率 2.0×10^{-6} に設定している. その後, CNN のパラメータを固定して, 先述した二段階目の学習と同様の設定で分散を学習させている. これらの学習の rank-loss のマージンは, fine-tuning を行わない場合は 0.2 を, 行う場合は 0.15 を使用している. 検索結果の評価指標には一般に Image Caption Retrieval で用いられる R@k, Med r, Mean r を用いている. 評価データ 5000 枚は 1000 枚ずつに対して検索を行い, 5 つの評価値の平均を評価結果としている.

実験では, 4 つの条件設定において既存手法である VSE++と提案手法である正規分布への埋め込みを比較する. 4 つの条件とは, CNN に VGG か ResNet を用いた場合を, それぞれ fine-tuning ありの場合となしの場合の 4 パターンに対応している. これらの比較は, VGG や ResNet の出力といった異なる特徴量を用いた場合にも本手法が適用できること, fine-tuning と組み合わせることが可能であることを示すために行っている. 結果を表 1 に示す.

表 1 の上から, VGG, ResNet, VGG + fine-tuning, ResNet + fine-tuning の場合における既存手法 (上) と提案手法 (下) の比較である. R@1,R@5,R@10,Med r, Mean r 全ての指標において, 提案手法の精度が向上している. また画像検索 (Image Retrieval) とテキスト検索 (Caption Retrieval) を比較すると, 評価値の改善は画像検索の mean r が顕著である. 特に, VGG, ResNet, VGG + fine-tuning, ResNet + fine-tuning の各条件で Caption Retrieval の Mean r は平均 0.075 の改善であるのに対して, Image Retrieval の Mean r は平均 0.85 ほど小さくなっている. 以上のことから, 提案手法が画像検索側の Mean r に対して, 大きく影響を及ぼしていることが分かる.

5. 考察

表 1 より, VGG, ResNet の中間層の出力どちらを用いた場合においても, 精度の向上が見られていることから, 提案手法が画像特徴量によらず適用可能であると考えられる. この事実から, 画像の分類タスクにおける CNN モデルの改良が進んだ場合, そのモデルに提案手法を適用すれば, 検索精度のさらなる向上が期待できる.

次に, Mean r の精度が向上した理由について考察を行う. 図 2 は, 画像検索において Mean r を大きく改善させた例の一つである. 図 2 下の表における Rank は, クエリーとなるテキストに対して正解となる上の画像の出力順位を示している. 上から 4 番目のテキストの Rank を見ると, VSE++は 665 と他と比較してかなり大きい値を取っている. テキストの内容に注目すると, このテキストのみ “bus” という単語が含まれていない. これは, VSE++が LSTM の学習において, テキスト内の “bus” という単語の有無のみで, 画像とテキストを類似度の高い点へ埋め込んでしまっていると考えられる. しかし, 分布への埋め込みを行う提案手法の Rank は 282 であり, 大きく改善している. これは図 2 上のバスの画像の分散を大きくとるか, “bus” という単語の含まれていない, 上から 4 番目のテキストの分散を大きくとるように, 提案手法が学習したからだと考えられる. このことから, 提案手法は各データの持つ意味の曖昧さを捉えていると考えられる.

Caption	Rank VSE++	Rank Ours
Passenger bus stationary in traffic on city street.	3	**2**
A modern city commuter bus in traffic during the day	2	**1**
A bus parked on the side of the road while in traffic	**5**	6
a modern train is parked against the sidewalk curb	665	**282**
A bus stops at a curb in a busy city street.	6	**3**

図 2: An example of greatly improving Mean r.

表 1 の fine-tuning ありとの比較を見ると, 評価値の向上が小さいものも見られる. これは, fine-tuning することによって, モデルの表現力が大きくなっていることが関係すると考えられる. ImageNet での事前学習に用いる画像の枚数は 120 万枚ほどなのに対して, MS COCO の訓練データ数は 11 万枚ほどしかないので, fine-tuning によって, 平均の学習で過学習を起こしている可能性がある. そのため, その後の分散の学習による精度の向上が困難になったと考えられる. しかし, 画像検索の mean r に関しては大きく改善しているため, 提案手法による改善が, 誤差の影響によるものとは考えにくい.

6. 結論

本研究では VSE++の埋め込みの出力を正規分布の平均と分散に変更し, rank-loss の類似度を JS ダイバージェンスを用いて定義するという変更を加えて, 画像データとテキストデータの共通空間上の点への埋め込みから, 確率分布への埋め込みに拡張する方法を提案した. 提案手法は学習を行う際, 通常の VSE++と同様に学習を行うのではなく, 平均のみを事前に VSE++で学習した上で, 平均を固定して追加で分散の学習を行うために, 分散を上手く学習することが可能となり, 検索精度が向上された. この結果から, 提案手法が共通空間上で望ましい類似度を定義していることの傍証が得られた.

本研究の今後の課題について述べる. まず埋め込みに用いる確率分布を変更するといった工夫が可能である. 本研究で正規分布に仮定した各次元の独立性を外すか, 多峰性の混合正規分布に埋め込むことも考えられる. また, 分布への埋め込みを行っていることから, 新たな正則化の導入も考えられる. 例として, KL ダイバージェンスを用いて分布間の包含関係を表現すれば, ランダムクロップされた画像の分布が切り出される前の画像に包含されるといった仮定を, 正則化条件として加えることができる. このような点への埋め込みでは導入できない, 分布への埋め込みのための正則化によって, モデルの汎化能力が向上することが期待される.

本研究は科研費 (16K12487) の支援 , 総務省 SCOPE(受付番号 172107101) の委託を受けて行われた.

参考文献

[Cho 14] Cho, K., Merrienboer, van B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp. 1724–1734 (2014)

[Deng 09] Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F.: ImageNet: A large-scale hierarchical image database, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 248–255 (2009)

[Faghri 17] Faghri, F., Fleet, D. J., Kiros, R., and Fidler, S.: VSE++: Improved Visual-Semantic Embeddings, *arXiv*, pp. 1–11 (2017)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 770–778 (2016)

[Kingma 14] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *arXiv*, pp. 1–15 (2014)

[Kiros 14] Kiros, R., Salakhutdinov, R., and Zemel, R. S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, *arXiv*, pp. 1–13 (2014)

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems(NIPS)*, pp. 1106–1114 (2012)

[Lin 14] Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, in *Proceedings of the 13th European Conference of Computer Vision(ECCV)*, pp. 740–755 (2014)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *arXiv*, pp. 1–12 (2013)

[Simonyan 14] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *arXiv*, pp. 1–14 (2014)

[Wang 16] Wang, W., Yang, X., Ooi, B. C., Zhang, D., and Zhuang, Y.: Effective deep learning-based multimodal retrieval, *Very Large Data Bases Journal(VLDB J)*, Vol. 25, No. 1, pp. 79–101 (2016)