3O1-13in

Deep Kernel を用いた高次元空間への階層的な写像とその最適化

Hierarchical Mapping for High Dimensional Spaces with Deep Kernel and the Optimization

椿真史 *1 Masashi Tsubaki　Kevin Duh*1 Kevin Duh　新保仁 *1 Masashi Shimbo　松本裕治 *1 Yuji Matsumoto

*1奈良先端科学技術大学院大学
Nara Institute of Science and Technology (NAIST)

In this paper, we discuss a role of hierarchical non-linear functions of Deep Learning for solving classification and clustering problems, and we pay attention to the fact that Deep Learning methods provide a vector space with the appropriate distance and similarity. Based on this assumption, we propose a method to optimize high dimensional space with Deep Kernel, and try to compare with Deep Learning architecture.

1. はじめに

機械学習において最も重要なのは，適切なベクトル空間の存在であり，これを前提として分類やクラスタリングが行われる．ここでの適切とは，データ間の類似度が分類やクラスタリングの問題を解く上で適切である，ということである．近年の Deep Neural Network(以下 DNN)，特に表現学習の成功は，このような適切な空間が学習されていることを意味する．

自然言語処理においては主に，個々の単語から文全体が持つ何らかの表現を構成したい時に，DNN が用いられる．具体的には，単語ベクトル表現から構成された文全体の表現に対する感情分析や関係認識，さらには画像とのマルチモーダルな対応付けなどにおいて大きな成功を収めている [Socher 12, Socher 14]．このような成功は，DNN の基本的な考えである非線形関数の階層的な適用が，単語という低次の要素から文という高次の対象を構成するために，重要な役割を果たすことを示唆している．そしてこれはつまり，高次の表現に対して，冒頭で述べたような適切な類似度を持つ空間が学習されることを意味する．

これらを踏まえて本稿では，以下の 2 つ問題を提起する．

- 画像や言語に限らず様々なデータにおいて，低次の要素から高次の対象を構成する際に，それに伴うより複雑な意味を表現するための適切な高次元空間が必要なのではないか？
- 上記の高次元空間は，DNN によって，つまり低次元空間における非線形関数の階層的な適用によって獲得される空間と，どのような関係にあるのか？

そこで我々は，Deep Kernel を用いた多層非線形類似度学習法を提案する．DNN が持つモデル上の制約として，低次から高次の表現を構成する際に，ある低次元に固定された表現を常に考慮しなければならない [Socher 12, Socher 14] のに対し，我々はカーネルの持つ高次元空間への写像というより柔軟な発想を用いる．そしてさらに，これを Deep Kernel へと拡張した上で，高次元空間の類似度を階層的に最適化する．最終的には，最も下層の低次の表現が，上層のより高次の表現を適切に構成できるように，元のベクトル空間のみが明示的に学習される (図 2)．

本稿の貢献は以下の 3 つである．

連絡先: 椿真史，奈良先端科学技術大学院大学，
masashi-t@is.naist.jp

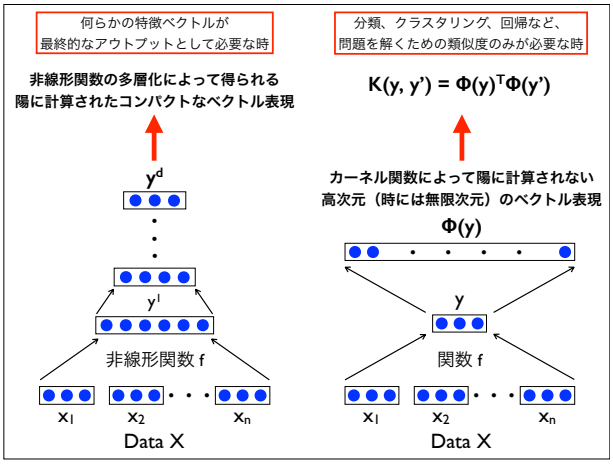

図 1: Deep Neural Network とカーネル法の比較．低次の要素から高次の対象の表現を構成する，という目的に焦点を当てた場合，前者は低次元空間における非線形関数の階層的な適用で，後者は高次元空間への写像によって，それを実現すると捉えることができる．

1. Deep Kernel を用いた多層非線形類似度学習法の提案は，我々の知る限り本稿が初めてである．
2. 提案法はシンプルかつ実装も容易であるが，文の意味的類似度評価データセットにおいて，様々な DNN のモデルを上回る，あるいは同等の性能を出すことに成功した．
3. DNN における非線形関数の適用とカーネル法における高次元空間への写像という 2 つの側面から，低次から高次の表現構成における非線形性について新たに考察した．

2. 提案法

訓練データは，$\{(S_i, S'_i), y_i\}_{i=1}^n$ の形式で与えられる（3.1.1 節）．S と S' は文，$y \in [-1, +1]$ はその類似度を表す．目標は，新たな文の類似度を予測することである．我々はまず，2 つのシンプルな計算法を用いて文のベクトル表現を得る（2.1

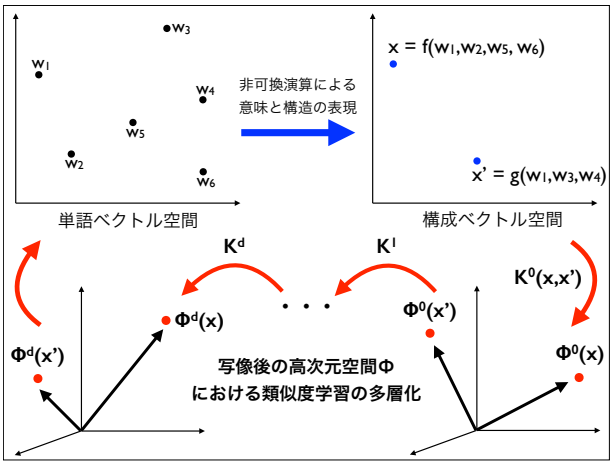

図 2: 提案法の直感的なイメージ．階層的にカーネル関数を適用することで階層的に高次元空間へ写像し，各々の空間の類似度をカーネルを通してすべて最適化する．最終的には，元の単語ベクトル空間のみが再学習される．

節）．次に，文の類似度計算をカーネルを用いて非線形に拡張し（2.2 節），さらに Deep Kernel へと拡張する（2.3 節）．最後に，多層非線形類似度学習における最適化法について，2 つのアプローチを述べる（2.4 節）．

2.1 文の意味と構造のベクトル表現

まず最もシンプルな文のベクトル表現として，文 S のベクトル $\mathbf{x}$ を，

$$\mathbf{x} = f_{ADD}(S) = \sum_{w \in S} \mathbf{d}(w) \tag{1}$$

と計算する．ここで，$\mathbf{d}(w)$ は単語 w の n 次元ベクトル表現とする．この計算法は，文内に現れるすべての単語の共起情報を考慮することができる反面，N-gram や係り受け関係などの系列や構造の情報は一切考慮できない欠点がある．そこで次に，文ベクトル $\mathbf{x}$ を，D_S を文 S 内の係り受け関係にある単語ペアの集合とした上で，

$$\mathbf{x} = f_{SUBT}(D_S) = \sum_{(w_i, w_j) \in D_S} (\mathbf{d}(w_i) - \mathbf{d}(w_j)) \tag{2}$$

と計算する．ここで，w_i と w_j は係り受け関係にある単語ペアである．これは，最も基本的な非可換演算である減算を用いることで，文内の係り受け関係にある単語間 w_i と w_j の順序情報をエンコードする．このように意味と構造の情報を低次元空間で表現した上で，後述するカーネルを用いた非線形類似度学習法を適用し，高次元空間においてより詳細に意味と構造を最適化する．本稿での文ベクトル計算は，これら 2 つのシンプルな手法を用いるに留める [*1]．

2.2 カーネルを用いた非線形類似度計算

まず我々は，意味的類似度計算のためのカーネル関数 K に，自然言語処理において幅広く用いられる，線形カーネルのコサイン類似度 K_{cos} を用いる．

$$K_{\cos}(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{x}'}{\sqrt{\mathbf{x}^{\mathrm{T}}\mathbf{x}}\sqrt{\mathbf{x}'^{\mathrm{T}}\mathbf{x}'}} \tag{3}$$

以降，本稿で述べるすべてのカーネルは正規化されているものとし，以下のように表現する．

$$K_{\cos}(\phi(\mathbf{x}), \phi(\mathbf{x}')) = \frac{K(\mathbf{x}, \mathbf{x}')}{\sqrt{K(\mathbf{x}, \mathbf{x})}\sqrt{K(\mathbf{x}', \mathbf{x}')}} \tag{4}$$

ここで ϕ は，次に述べる非線形カーネルによって写像される高次元空間である．つまり我々は，正規化されたカーネルを用いることで，高次元空間 ϕ においても，適切な意味的類似度であるコサイン類似度を考えることができる．

次に我々は，以下の 2 つの非線形カーネルを用いる．

$$K^{\ell=0}_{poly}(\mathbf{x}, \mathbf{x}') = (c_0 + K_{\cos}(\mathbf{x}, \mathbf{x}'))^p \quad s.t \quad c_0 \geq 0, \quad p \in \mathbb{N} \tag{5}$$

$$K^{\ell=0}_{rbf}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1 - K_{\cos}(\mathbf{x}, \mathbf{x}')}{\sigma_0^2}\right) \quad s.t \quad \sigma_0 \geq 0 \tag{6}$$

K_{poly} は多項式カーネル，K_{rbf} は RBF カーネル [*2] である．ここでの ℓ は，次節で述べる Deep Kernel における Layer 数を表し，Layer 数 $\ell = 0$ の時を通常の多項式カーネルと RBF カーネルとする．

2.3 Deep Kernel による多層化

前節を踏まえて，最終的に本稿で提案する Deep Kernel は，以下のように再帰的に定義される．

$$K^{\ell}_{poly}(\mathbf{x}, \mathbf{x}') = (c_\ell + K^{\ell-1}_{poly}(\mathbf{x}, \mathbf{x}'))^p \quad s.t \quad c_\ell \geq 0, \quad p \in \mathbb{N} \tag{7}$$

$$K^{d}_{rbf}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1 - K^{\ell-1}_{rbf}(\mathbf{x}, \mathbf{x}')}{\sigma_\ell^2}\right) \quad s.t \quad \sigma_\ell \geq 0 \tag{8}$$

前述の通り，ℓ は Deep Kernel における Layer 数である．このようにカーネルを多層化することで，カーネル内パラメータ (多項式では c，RBF では σ) が増え，ℓ 次元ベクトルとなる．Deep Kernel により，表現力の高い空間へ階層的に写像することができ，またそれは ℓ 次元のパラメータベクトルによって適切な類似度を持つ空間として制御される．我々は，高次元空間に写像されたベクトル $\phi(\mathbf{x})$，つまり最終的な文ベクトルを陽に得ることなく，カーネルを通してその意味と構造の類似度のみを計算し学習することで，新たな単語ベクトル表現を陽に得ることができる (図 2)．

*1 より複雑な文ベクトルの計算法については [Socher 12, Socher 14] を参照されたい．これらを本稿の提案法に組み込むことは今後の課題である．

*2 一般的に用いられる RBF カーネルはユークリッド距離を用いた $K_{rbf}(\mathbf{x}, \mathbf{x}') = \exp\left(-||\mathbf{x} - \mathbf{x}'||^2/2\sigma^2\right)$ である．これはまず $||\mathbf{x} - \mathbf{x}'||^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x} + \mathbf{x}'^{\mathrm{T}}\mathbf{x}' - 2\mathbf{x}^{\mathrm{T}}\mathbf{x}'$ と内積を用いて書き下すことができる．次に，これらの内積を任意のカーネルに置き換えることが可能であるため，すべてをコサイン類似度に置き換え $||\mathbf{x} - \mathbf{x}'||^2 = K_{\cos}(\mathbf{x}, \mathbf{x}) + K_{\cos}(\mathbf{x}, \mathbf{x}) - 2K_{\cos}(\mathbf{x}, \mathbf{x}')$ を得る．そして最終的に，$||\mathbf{x} - \mathbf{x}'||^2 = 2 - 2K_{\cos}(\mathbf{x}, \mathbf{x}')$ が得られるため，本稿での RBF カーネルは式 (6) となることに注意されたい．

2.4 多層非線形類似度学習

最終的に用いるカーネルとロス関数は以下の通りである．

$$K^{\ell}(S,S') = K^{\ell}(f_{ADD}(S), f_{ADD}(S')) \times K^{\ell}(f_{SUBT}(D_S), f_{SUBT}(D_{S'})) \quad (9)$$

$$L(\Theta) = \sum_{i=1}^{n} \frac{1}{2}\{y_i - K^{\ell}(S_i, S'_i)\}^2 + \frac{\lambda}{2}||\Theta||^2 \quad (10)$$

$L(\Theta)$ は，データセットで与えられた文の類似度 y とカーネル $K^{\ell}(S_i, S'_i)$ との正則化付き二乗誤差である．Θ は学習するパラメータ集合であり，単語ベクトルとカーネル内パラメータ (多項式カーネルの場合は $\mathbf{c} \in \mathbb{R}^{\ell}$，RBF カーネルの場合は $\boldsymbol{\sigma} \in \mathbb{R}^{\ell}$) である．

学習法には 2 つの戦略が存在する．1 つは，$\mathbf{c}$ や $\boldsymbol{\sigma}$ をロス関数におけるベクトルパラメータと見なして一度に最適化する手法であり，本稿ではこれを Layer-all と呼ぶ．もう 1 つは，c_i や σ_i を下層 ($i = 0$) から 1 つずつ学習していく手法であり，本稿ではこれを Layer-wise と呼ぶ [*3]．Layer-wise は Layer-all と異なり，適切な類似度を持つ高次元空間を，1 層ずつ階層的に学習していくことになる．

3. 実験結果と考察

3.1 実験

3.1.1 データセット

提案法は，SemEval 2014 の Sentences Involving Compositional Knowledge (SICK) [Marelli 14] のデータセット [*4] を用いて評価した．このデータセットは，2 つの文の意味的な類似度を人手でスコアリングしたものであり，訓練データとテストデータは各々約 5000 文対から成る．評価には，提案法によって計算された二つの文ベクトルの類似度と人手の類似度スコアとの，ピアソンの相関係数 r とスピアマンの相関係数 ρ，そして平均二乗誤差 (MSE) を用いる [*5]．我々の目標は，単語ベクトル表現を用いた意味構成モデルによって，新たに与えられた文の意味的類似度を正確に予測することである．

3.1.2 比較する既存研究

既存研究では主に，2 つの文に含まれる単語や N-gram のマッチングやオーバーラップ，品詞や木構造のアライメント，さらには WordNet などの外部知識などを用いて様々な素性を考え，それらを用いてサポートベクター回帰で学習するものが一般的である．SemEval2014 の SICK に関しても，同様の素性エンジニアリングの手法に基づいたアプローチが多数を占めている（Illinois-LH_run1 [Lai 14] UNAL-NLP_run1 [Jimenez 14] Meaning_Factory_run1 [Bjerva 14] ECNU_run1 [Zhao 14]）．特に，単語ベクトル空間と構成モデルのみを用いた意味的なアプローチのみでは，相関係数が 0.7 程度に留まるという報告がある [Marelli 14]．一方で，Deep Neural Network を用いた手法も幾つか提案されている [Socher 14, Tai 15]．これらは主に，Recursive Neural Network あるいは Recurrent Neural Network を用いて，単語ベクトルから文ベクトルを直接計算するモデルである．文の依存構造や木構造を考慮した上で多数の重み行列や非線形関数を階層的に適用し，低次元の文表現を学習する手法となっている．

∗3 これは，Deep Learning における Pre-training と同様の戦略である．

∗4 http://alt.qcri.org/semeval2014/task1/

∗5 SemEval 2014 のオフィシャルのランキングにおいては，ピアソンの相関係数 r が用いられている．

カーネル	r	ρ
コサイン	0.7521	0.7445
多項式 (p=2)	**0.8414**	**0.7886**
多項式 (p=3)	0.8410	0.7870
多項式 (p=4)	0.8387	0.7845
RBF	0.8373	0.7860
多項式 (Deep (Layer-all))	0.8230	0.7788
多項式 (Deep (Layer-wise))	0.772	0.7319
RBF(Deep (Layer-all))	0.7780	0.7692
RBF(Deep (Layer-wise))	0.7508	0.7312

表 1: 様々なカーネルを用いた時の，ピアソンとスピアマンの相関係数．

3.1.3 実装の詳細

提案法で再学習する単語ベクトル表現については，初期値として 300 次元の Global vector [Pennington 14] [*6] を用いた．また構文解析には Enju [*7] を用いた．

カーネル多層化では 3 層について実験を行った．特に多項式カーネルについては，層数が上がるに従い次数 p を上げ，上位層でより高次の素性の組み合わせを考慮できるモデルとした．

最終的なコスト関数は式 (10) の $L(\Theta)$ であり，これを最小化する．最適化には AdaGrad [Duchi 11] を用いた．単語ベクトルの学習率は $\alpha = 10^{-1}$，カーネル内パラメータの学習率は $\beta = 10^{-3}$，正則化項については $\lambda = 10^{-6}$ とした．データセットに対してはイテレーション数を上限 1000 に統一し実験を行い，比較検証した．また 3 層の Layer-wise の学習では，イテレーション数を 3 等分することで，階層的に高次元空間の学習を行った．

3.2 結果と考察

3.2.1 線形 vs. 非線形

線形カーネルであるコサイン類似度よりも，多項式カーネルと RBF カーネルを用いた非線形類似度学習によって，ピアソンの相関係数が最大で 0.1 ポイント程度上昇する結果となった．これは，単語ベクトル空間とは異なる高次元空間，つまり単語より表現力の高い空間において文の類似度を学習することが，非常に有効であることを示している．

3.2.2 ADD vs. SUBT

単語ベクトル間の演算において，可換の ADD と非可換の SUBT とを比較した．ADD では，文の系列や構造の情報はすべて失われてしまうが，文に出現する単語の共起情報をすべて考慮できるため，全体的に相関係数が高い結果となっている．また ADD と非可換演算である SUBT とを組み合わせたモデルでは，非線形カーネル，特に多項式カーネルを用いた場合に相関係数の上昇が見られた一方で，線形カーネルでは逆に相関係数が下がる結果となった．これは，SUBT による文の意味と構造の表現が，単語と同一次元の空間においては適切なエンコードではないが，異なる高次元空間においてその表現がより適切に学習されていることを示している．

3.2.3 Kernel vs. Deep Kernel

カーネルの多層化では，相関係数の上昇は見られなかった．特に RBF カーネルについては，相関係数が下がる結果となった．これは RBF カーネルによる無限次元空間への写像によって，表現力の非常に高い空間において類似度学習が適用されているため，過学習を引き起こすなどの問題が起きていると考え

∗6 http://nlp.stanford.edu/projects/glove/

∗7 http://www.nactem.ac.uk/enju/index.ja.html

手法	r	ρ	MSE
Illinois-LH_run1 [Lai 14]	0.7993	0.7538	0.3692
UNAL-NLP_run1 [Jimenez 14]	0.8043	0.7458	0.3593
Meaning_Factory_run1 [Bjerva 14]	0.8268	0.7722	0.3224
ECNU_run1 [Zhao 14]	0.8280	0.7689	0.3250
DT-RNN [Socher 14]	0.7863	0.7305	0.3983
SDT-RNN [Socher 14]	0.7886	0.7280	0.3859
Constituency Tree LSTM [Tai 15]	0.8491 (2)	0.7873 (3)	0.2852 (2)
Dependency Tree LSTM [Tai 15]	**0.8627 (1)**	**0.8032 (1)**	**0.2635 (1)**
提案法 (多項式カーネル (p=2))	**0.8414** (3)	**0.7886** (2)	**0.3040 (3)**
提案法 (多項式カーネル Deep (Layer-all))	0.8230	0.7788	0.3259
提案法 (多項式カーネル Deep (Layer-wise))	0.772	0.7319	0.3563

表 2: 素性エンジニアリングをベースとした SemEval 2014 の上位のチームと RNN を上回る性能を達成することに成功した一方で，Constituency Tree LSTM と Dependency Tree LSTM については，提案法が同程度あるいは若干下回る結果となった．

られる．この過学習をどのように防ぐかは，今後の大きな課題である．

3.2.4 Layer-all vs. Layer-wise

これら 2 つの多層非線形類似度学習の戦略については，議論の余地が大いにある．Layer-wise については，1 層の学習をどの程度行ってより上位層へ移行するのかというハイパーパラメータ設定の問題がある，また，RBF カーネルを用いた場合の無限次元空間では，1 層ですでに任意の関数を近似できる性質を持つため，過学習する危険が非常に高い．これらすべて考慮した上で，Layer-all と Layer-wise の学習戦略を再考する必要がある．

3.2.5 提案法 vs. 既存研究

提案法は，ピアソンとスピアマンの相関係数，平均二乗誤差 (MSE) のすべてにおいて，素性エンジニアリングをベースとした SemEval 2014 の上位のチームと，[Socher 14] らの提案した RNN を上回る結果となった．一方で，Constituency Tree LSTM と Dependency Tree LSTM [Tai 15] については，提案法が同程度かあるいは若干下回る結果となった．しかし，LSTM の手法と比較すると，提案法はよりシンプルかつ実装も容易であり，今後様々な拡張を考えることができる．特に我々の結果は，単語ベクトル空間における意味と構造の非可換演算と，カーネルによる非線形類似度学習のみによって達成されており，その点において遥かに優位性があると考えられる．

4. 結論

本稿で我々は，Deep Kernel を用いた多層非線形類似度学習を提案した．実験結果は，低次の要素から高次の対象を構成する際に，それに伴うより複雑な意味を表現するための，適切な類似度を持つ高次元空間の学習の重要性を示唆している．関連する研究として [Vinyals 12] は，非線形関数を階層的に適用することで，カーネル SVM と同程度の性能を持つ線形 SVM の提案に成功している．DNN における低次元空間と，カーネルにおける高次元空間でそれぞれ学習されるデータについて，その非線形性という側面を考察する価値は大いにある．

参考文献

[Bjerva 14] Bjerva, J., Bos, J., Goot, van der R., and Nissim, M.: The Meaning Factory: Formal Semantics for Recognizing Textual Entailment and Determining Semantic Similarity, in *SemEval* (2014)

[Duchi 11] Duchi, J., Hazan, E., and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *JMLR* (2011)

[Jimenez 14] Jimenez, S., Duenas, G., Baquero, J., Gelbukh, A., Bátiz, A. J. D., and Mendizábal, A.: UNAL-NLP: Combining soft cardinality features for semantic textual similarity, relatedness and entailment, in *SemEval* (2014)

[Lai 14] Lai, A. and Hockenmaier, J.: Illinois-lh: A denotational and distributional approach to semantics, in *SemEval* (2014)

[Marelli 14] Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R.: SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment, in *SemEval* (2014)

[Pennington 14] Pennington, J., Socher, R., and Manning, C. D.: Glove: Global vectors for word representation, in *EMNLP* (2014)

[Socher 12] Socher, R., Huval, B., Manning, C. D., and Ng, A. Y.: Semantic Compositionality through Recursive Matrix-Vector Spaces, in *EMNLP-CoNLL* (2012)

[Socher 14] Socher, R., Le, Q. V., Manning, C. D., and Ng, A. Y.: Grounded Compositional Semantics for Finding and Describing Images with Sentences, *TACL* (2014)

[Tai 15] Tai, K. S., Socher, R., and Manning, C. D.: Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, *arXiv preprint arXiv:1503.00075* (2015)

[Vinyals 12] Vinyals, O., Jia, Y., Deng, L., and Darrell, T.: Learning with recursive perceptual representations, in *NIPS* (2012)

[Zhao 14] Zhao, J., Zhu, T. T., and Lan, M.: ECNU: One Stone Two Birds: Ensemble of Heterogenous Measures for Semantic Relatedness and Textual Entailment, in *SemEval* (2014)