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

深層学習の学習経過におけるクラスタ構造の推移の可視化

Visualization of cluster structure transition in deep learning

渡邊千紘 *1
Chihiro Watanabe

*1NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

Deep neural networks have achieved high performance in various tasks, and to interpret their prediction mechanism is an important open problem. Recently, a series of methods have been proposed for decomposing a trained neural network into a simple interpretable module structure. In this paper, to acquire knowledge about the training process of a deep neural network, we proposed a method for visualizing the cluster structure transition during the training phase. Our proposed framework consists of two parts: we first decompose the network at each training step into modules based on hierarchical clustering, and then reveal the relationships between clusters at different training steps based on the ratio of their common units. The experimental results showed that our proposed method could provide us with knowledge about division and integration of neural network modules, and also information about the role of each module in terms of input-output mappings.

1. はじめに

深層学習における学習結果を人間が解釈するための手法を構築することは，近年の機械学習分野における主な課題のひとつである．深層ニューラルネットは，画像認識や音声変換など，多様な課題において有効性が確認されているが，その学習結果として得られる関数は深い階層構造を通した非線形変換の繰り返しで表現され，予測の仕組みをそのまま人間が理解することは困難である．

近年，ニューラルネットの学習結果の解釈性向上を目的として，データから学習されたニューラルネットをモジュール構造に分割する手法が提案されている [1, 2, 3]．これらの手法は，ニューラルネット全体をひとつのブラックボックス関数とみなしてその挙動を解析する手法 [4, 5] や，ユニットや層など決められた単位での挙動を解析する手法 [6, 7] と異なり，ニューラルネットにおいて類似した役割を持つ部分構造を自動的に獲得できるという利点がある．

本論文では，学習済みニューラルネットをモジュール構造に分解する手法を，ニューラルネットの学習時における各エポックでの学習結果に対して適用することにより，深層学習の学習経過におけるクラスタ構造の推移を可視化する手法を提案する．この手法により，ニューラルネットの学習過程において，隠れ層におけるユニットの (1) 異なる役割を持つ部分への分割と，(2) 類似した役割を持つ部分の統合がどのようになされているかを知ることが可能になる．

実際に，画像認識を行うように学習された深層畳み込みネットワークに対し，提案法を適用することにより，学習の過程において隠れ層のユニットが分割・統合されていく様子を可視化する実験を行い，提案法の有効性と課題の検討を行った．

2. ニューラルネットのモジュール構造抽出

本節では，与えられた学習データを用いて学習済みのニューラルネットにおいて，各隠れ層のユニットが持つ役割を表す特徴ベクトルを定義し，それに基づいて階層的クラスタリングを適用することにより，ニューラルネットを類似する役割を持つモジュール構造に分解する手法を説明する．

2.1 隠れ層の各ユニットに対応する特徴ベクトルの定義

まず，学習済みのニューラルネットが与えられたとき，各隠れ層のユニット k が果たす役割を表す特徴ベクトル v_k を定義する．本論文では，特徴ベクトル v_k を，ユニット k の出力値と各入出力次元の値との相関に基づいて定義することとする．これにより，ユニット k が主に，入力データのどの次元の値を用いて，どの出力次元の値を予測するのに用いられているか（画像認識のタスクであれば，ユニット k が入力画像のどの画素の値を用いて，どの出力クラスの認識を行っているか）を定量的に求めることができる．

2.1.1 i 番目の入力次元の値が隠れ層のユニット k の出力値に与える影響の大きさ

i 番目の入力次元の値が隠れ層のユニット k の出力値に与える影響の大きさ v_{ik}^{in} を，以下の式で定義する．

$$v_{ik}^{\mathrm{in}} = \frac{E\left[\left(X_i^{(n)} - E[X_i^{(n)}]\right)\left(o_k^{(n)} - E[o_k^{(n)}]\right)\right]}{\sqrt{E\left[\left(X_i^{(n)} - E[X_i^{(n)}]\right)^2\right]E\left[\left(o_k^{(n)} - E[o_k^{(n)}]\right)^2\right]}}.$$

ただし，$E[\cdot]$ は全学習データに対する平均値を表し，$X_i^{(n)}$ を n 番目の学習データに対する i 番目の入力次元の値，$o_k^{(n)}$ を n 番目の学習データに対する（学習済みニューラルネットにおける）ユニット k の出力値とする．

2.1.2 隠れ層のユニット k の出力値が j 番目の出力次元の値に与える影響の大きさ

隠れ層のユニット k の出力値が j 番目の出力次元の値に与える影響の大きさ v_{kj}^{out} を，以下の式で定義する．

$$v_{kj}^{\mathrm{out}} = \frac{E\left[\left(o_k^{(n)} - E[o_k^{(n)}]\right)\left(y_j^{(n)} - E[y_j^{(n)}]\right)\right]}{\sqrt{E\left[\left(o_k^{(n)} - E[o_k^{(n)}]\right)^2\right]E\left[\left(y_j^{(n)} - E[y_j^{(n)}]\right)^2\right]}}.$$

ただし，$y_j^{(n)}$ を n 番目の学習データに対する j 番目の出力次元の値とする．

連絡先: chihiro.watanabe.xz@hco.ntt.co.jp

Algorithm 1 学習済みニューラルネットにおける隠れ層のユニットの特徴ベクトル $\{v_k\}$ に対する階層的クラスタリング

1: 隠れ層のユニット k の特徴ベクトルを v_k とする（$k = 1, \cdots, k_0$）．各時刻 t において，クラスタ m に割り当てられたユニットの集合を $\{C_m^{(t)}\}$ とおく．$t \leftarrow 1, C_m^{(1)} \leftarrow \{m\}$.
2: ユニット i, j の特徴ベクトル間のコサイン距離を以下で定義する：$f(i,j) \equiv 1 - (v_i \cdot v_j)/(\|v_i\|_2 \|v_j\|_2)$.
3: コサイン距離に基づくクラスタ C_1, C_2 間の距離を以下で定義する：$d(C_1, C_2) \equiv (\sum_{i \in C_1, j \in C_2} f(i,j))/(|C_1||C_2|)$.
4: **for** $t = 2$ to $k_0 - 1$ **do**
5: $(m_1^*, m_2^*) \leftarrow \arg\min_{(m_1, m_2)} d(C_{m_1}^{(t-1)}, C_{m_2}^{(t-1)})$.
6: $m_1^* < m_2^*$ と仮定し，以下のようにクラスタを更新する．

$$C_m^{(t)} \leftarrow \begin{cases} C_{m_1^*}^{(t-1)} \cup C_{m_2^*}^{(t-1)} & (m = m_1^*) \\ C_m^{(t-1)} & (m \leq m_2^* - 1, m \neq m_1^*) \\ C_{m+1}^{(t-1)} & (m_2^* \leq m) \end{cases}.$$

7: **end for**

2.1.3 隠れ層のユニット k の特徴ベクトル

2.1.1, 2.1.2 節の定義に基づき，隠れ層のユニット k の特徴ベクトルを以下で定義する：$v_k \equiv [v_{1k}^{\text{in}}, \cdots, v_{i_0 k}^{\text{in}}, v_{k1}^{\text{out}}, \cdots, v_{kj_0}^{\text{out}}]$. ただし，$i_0$ は入力次元の数，j_0 は出力次元の数を表す．定義より，特徴ベクトル v_k は，ユニット k の出力値と各入出力次元の値との相関値を並べたベクトルであり，ユニット k がどの入力次元の値を主に用い，どの出力次元の予測に寄与するかを表すという意味で，ユニット k の推論における役割を表すものとなっている．

2.2 階層的クラスタリングに基づくユニットの分類

2.1 節で定義した特徴ベクトル $\{v_k\}$ に基づき，隠れ層のユニットのクラスタリングを行う．本論文では，ユニット i, j の特徴ベクトル間の距離をコサイン距離

$$f(i,j) \equiv 1 - \frac{v_i \cdot v_j}{\|v_i\|_2 \|v_j\|_2}.$$

で測るものとし，クラスタ C_1, C_2 間の距離を以下で定義する．

$$d(C_1, C_2) \equiv \frac{1}{|C_1||C_2|} \sum_{i \in C_1, j \in C_2} f(i,j).$$

上記の定義に基づき，全てのユニットが異なるクラスタに属する状態を初期状態として，各時刻で距離が最小のクラスタ対を 1 つのクラスタに統合することを繰り返す．全てのユニットが 1 つのクラスタに属する状態になった時点でクラスタリングを終了する．階層的クラスタリングの手続きの全体を Algorithm 1 に記載した．

3. ニューラルネットの学習経過におけるクラスタ構造の推移の可視化

ニューラルネット学習時の各エポックごとに，学習されたニューラルネットに対し 2. 節に述べたモジュール構造抽出法を適用することで，ニューラルネットの学習が進むにつれ，そのモジュール構造や各モジュールの役割がどのように変化していくかを知ることができる．以下では，s エポックの学習が完了した時点でのニューラルネットのクラスタを $C_1^{(s)}, \cdots, C_{m_s}^{(s)}$ とし（$s = 1, \cdots, s_0$），ニューラルネットの学習経過におけるクラスタ構造の推移を可視化するための手法について説明する．

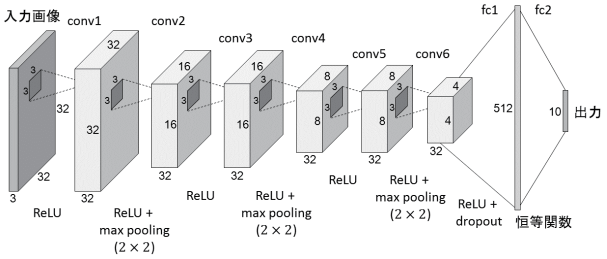

図 1: CIFAR-10 [8] の学習に用いたニューラルネットの構造

本論文では，ニューラルネットの学習経過におけるクラスタの分割や統合の様子を可視化するために，エポック (s_1, s_2) におけるクラスタ間の関係を行列 $R^{(s_1,s_2)}$ で表す．具体的には，s_1 エポックの学習完了時の m_1 番目のクラスタ $C_{m_1}^{(s_1)}$ と，s_2 エポックの学習完了時の m_2 番目のクラスタ $C_{m_2}^{(s_2)}$ の関係を，以下のように定義する．クラスタ $C_{m_1}^{(s_1)}$ に含まれるユニットの数 $|C_{m_1}^{(s_1)}|$ に対する，クラスタ $C_{m_1}^{(s_1)}$ とクラスタ $C_{m_2}^{(s_2)}$ に共通するユニットの数 $|C_{m_1}^{(s_1)} \cap C_{m_2}^{(s_2)}|$ の割合を行列 $R^{((s_1,s_2)}$ の (m_1, m_2) 成分として定義する．

$$R_{m_1,m_2}^{(s_1,s_2)} \equiv \frac{|C_{m_1}^{(s_1)} \cap C_{m_2}^{(s_2)}|}{|C_{m_1}^{(s_1)}|}.$$

上記の行列 $R^{(s_1,s_2)}$ から，エポック (s_1, s_2) 間でのクラスタ構造の推移について知識を得ることができる．

4. 実験

10 種類のクラスからなる画像データセット CIFAR-10 [8] を用いて，画像認識を行うニューラルネットの学習を行い，提案法を適用することにより，ニューラルネットの学習経過におけるクラスタ構造の推移の可視化を行った．

ニューラルネットの学習においては，図 1 に示す畳み込みネットワークを用い，学習率は Adam [9] に基づいて決めることとし，バッチサイズを 100, エポック数を 30 とした．全ての畳み込み層において，サイズ 1 の padding を適用した．

学習されたニューラルネットに対し，全学習データの中からランダムサンプリングした 2500 個の入力データを用いて特徴ベクトルの計算を行った．ここで，あるユニットの出力値の入力データに対する分散が 0 であるとき，各入出力次元の値との相関値が定義できないため，そのようなユニットは削除してから階層的クラスタリングを適用した．各エポックの学習完了時におけるニューラルネットに対し，2.2 節に述べた手法に基づき階層的クラスタリングを行い，各クラスタ内の 2 点におけるコサイン距離が 0.83 以下となるようなクラスタ構造のうち，最大のクラスタ数を持つものを結果として用いた．3. 節に述べた手法に基づき，学習経過におけるクラスタ構造の推移を可視化した結果を図 2 に示した．また，1, 10, 20, 30 エポックにおける各クラスタのセントロイドをそれぞれ図 3, 4, 5, 6 に示した．

図 2 より，$(1, 10)$, $(10, 20)$ エポック間では，行クラスタに含まれるユニットのうち大部分が 1 つの列クラスタに割り当てられるケースが多く，行クラスタが複数の列クラスタに均等に分割されるケースは少ないことが分かる．一方，$(20, 30)$ エポック間では，行クラスタが分割され異なる列クラスタに統合されるケースが比較的多く見られる．また，図 3, 4, 5, 6 か

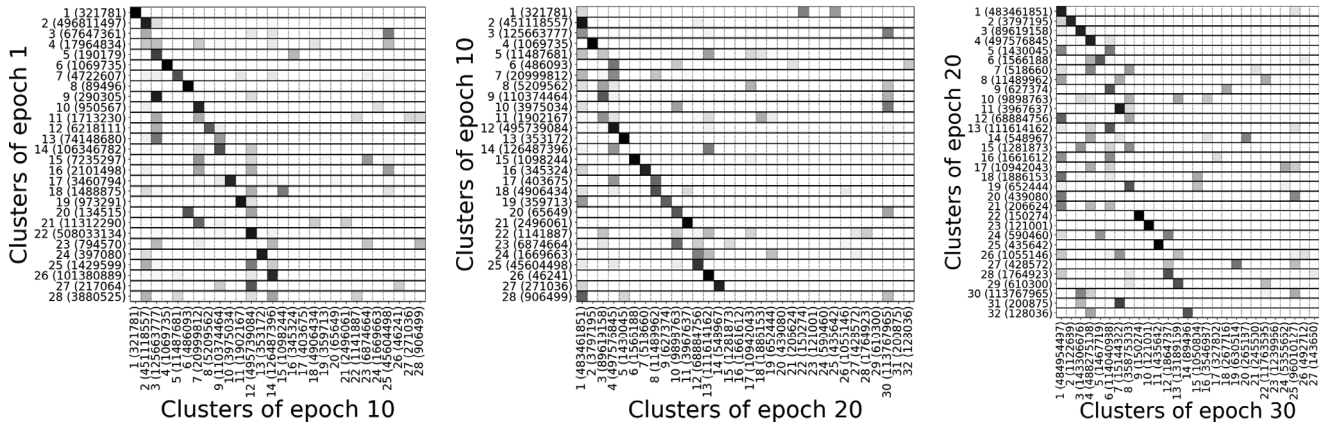

図 2: ニューラルネットの学習経過におけるクラスタ構造の推移．左から，$(1, 10)$, $(10, 20)$, $(20, 30)$ エポックの学習完了時におけるクラスタ間の関係を表す．行列の各行は，その行に対応するクラスタに含まれるユニットが，各列に対応するクラスタに分割される（行クラスタ内のユニット数に対する）割合を表す．かっこ内の数字は各クラスタに割り当てられたユニットの数を表す．

ら，学習経過における各クラスタの代表的な役割について知ることができる．たとえば，30 エポックの学習完了時におけるクラスタ 21（図 6）は，入力画像の中心に馬のような形があり，背景が緑色であることの情報を主に用いており，出力への寄与としては主に馬の認識を行うのに用いられていることがわかる．また，図 2 より，このクラスタは 20 エポックの学習完了時におけるクラスタ 7 に含まれるユニットの一部からなり，このクラスタ 7 も類似した役割を持つことが図 5 より読み取れる．他にも，例えば 30 エポックの学習完了時におけるクラスタ 25（図 6）は，主に自動車やトラックの画像と，それ以外の画像との判別に用いられていることが分かる．

5. 考察

本論文の提案手法に基づき，ニューラルネットの学習経過におけるクラスタ構造の推移を可視化することが可能になったが，この手法により得られる結果は，各クラスタ内の 2 点におけるコサイン距離の上限を示すハイパーパラメータを変化させることにより異なるものになる．このハイパーパラメータの値を大きくすれば，粗いクラスタ構造が得られ，各エポックにおける数の代表的なクラスタとその役割を知ることができるが，役割が大きく異なるユニットも同一クラスタに割り当てられてしまい，各クラスタの役割がセントロイドから読み取りづらくなる（符号のキャンセリングにより，どの入出力とも相関が見られなくなる）可能性がある．このことは，本論文の実験結果において，各クラスタにおける入出力との相関値の平均が比較的小さな値となったことの原因になっている（実際，1, 10, 20, 30 エポックにおいて，ユニット単位での入力次元との相関値の範囲はそれぞれ $[-0.71, 0.99]$, $[-0.71, 0.99]$, $[-0.71, 0.99]$, $[-0.71, 0.99]$ であり，出力次元との相関値の範囲はそれぞれ $[-0.59, 0.60]$, $[-0.60, 0.60]$, $[-0.58, 0.61]$, $[-0.58, 0.59]$ であった）．一方，この値を小さくすれば，コサイン距離の意味で近い特徴ベクトルのみを含むきめ細やかなクラスタ構造が得られるが，クラスタ数が膨大になり全体構造を把握することが難しくなるという問題がある．このようなクラスタ構造の解釈性におけるトレードオフを数量的に評価し，最適な結果を選択する手法を構築することは，提案法の重要な課題である．

6. 結論

本論文では，階層的クラスタリングに基づき，学習経過の各段階におけるニューラルネットのモジュール構造を抽出し，クラスタ構造の推移を可視化する手法を提案した．また，実際に画像認識を行うネットワークに提案法を適用し，その有効性と課題について検討を行った．

参考文献

[1] C. Watanabe, K. Hiramatsu, and K. Kashino. Modular representation of layered neural networks. *Neural Networks*, 97:62–73, 2018.

[2] C. Watanabe, K. Hiramatsu, and K. Kashino. Recursive extraction of modular structure from layered neural networks using variational Bayes method. In *Proceedings of Discovery Science 2017, Lecture Notes in Computer Science*, volume 10558, pages 207–222, 2017.

[3] C. Watanabe, K. Hiramatsu, and K. Kashino. Modular representation of autoencoder networks. In *Proceedings of 2017 IEEE Symposium on Deep Learning, 2017 IEEE Symposium Series on Computational Intelligence*, 2017.

[4] J. Chen, L. Song, M. Wainwright, and M. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 883–892, 2018.

[5] X. Zhang, A. Solar-Lezama, and R. Singh. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In *Advances in Neural Information Processing Systems 31*, pages 4879–4890, 2018.

[6] T. Zahavy, N. Ben-Zrihem, and S. Mannor. Graying the black box: Understanding DQNs. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1899–1908, 2016.

[7] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pages 6076–6085, 2017.

[8] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

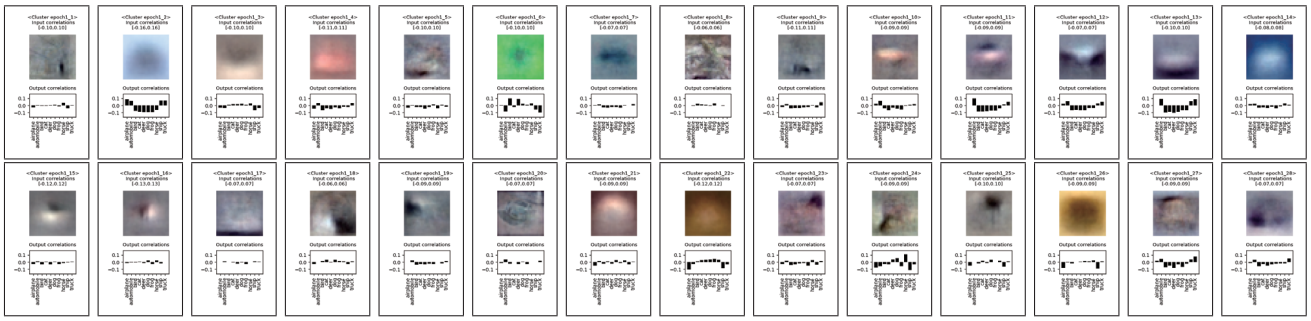

図 3: 1 エポックの学習完了時におけるニューラルネットの各クラスタのセントロイド．各クラスタの図について，上：クラスタに含まれるユニットの出力値と，各入力次元との相関値の平均（入力画像におけるどの画素の値を用いているか）．見やすさのため，各画像は，大かっこ内に示した実際の相関値の範囲を，その絶対値の最大値が 1 に対応するように引き伸ばした上で可視化している．下：クラスタに含まれるユニットの出力値と，各出力次元との相関値の平均（どの出力クラスの認識に寄与するか）．

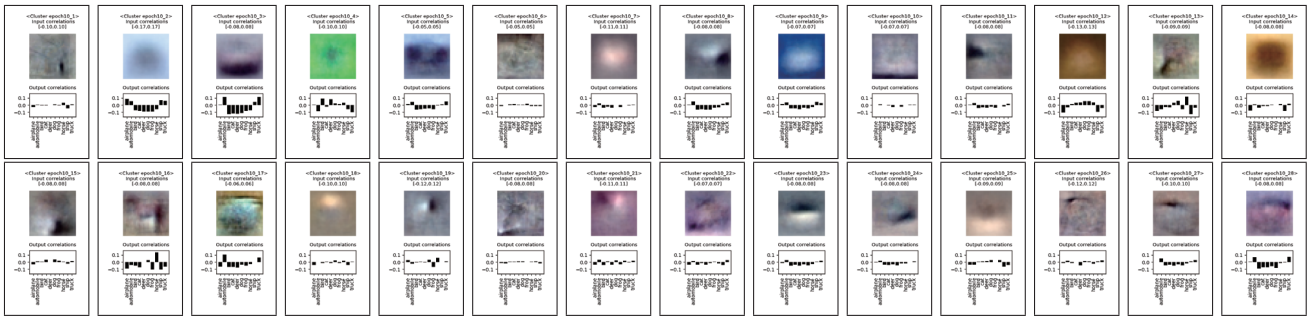

図 4: 10 エポックの学習完了時におけるニューラルネットの各クラスタのセントロイド．

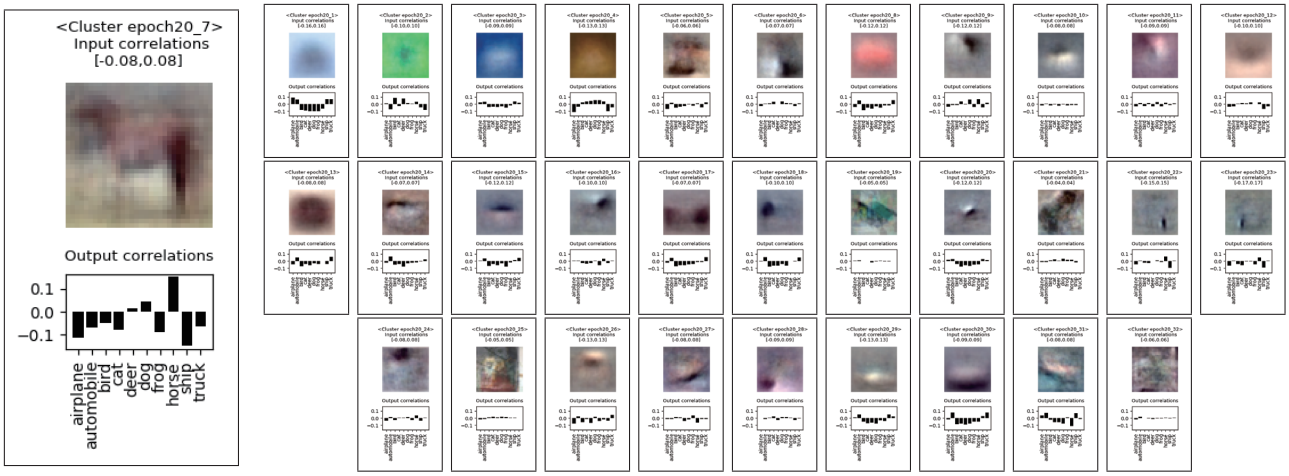

図 5: 20 エポックの学習完了時におけるニューラルネットの各クラスタのセントロイド．

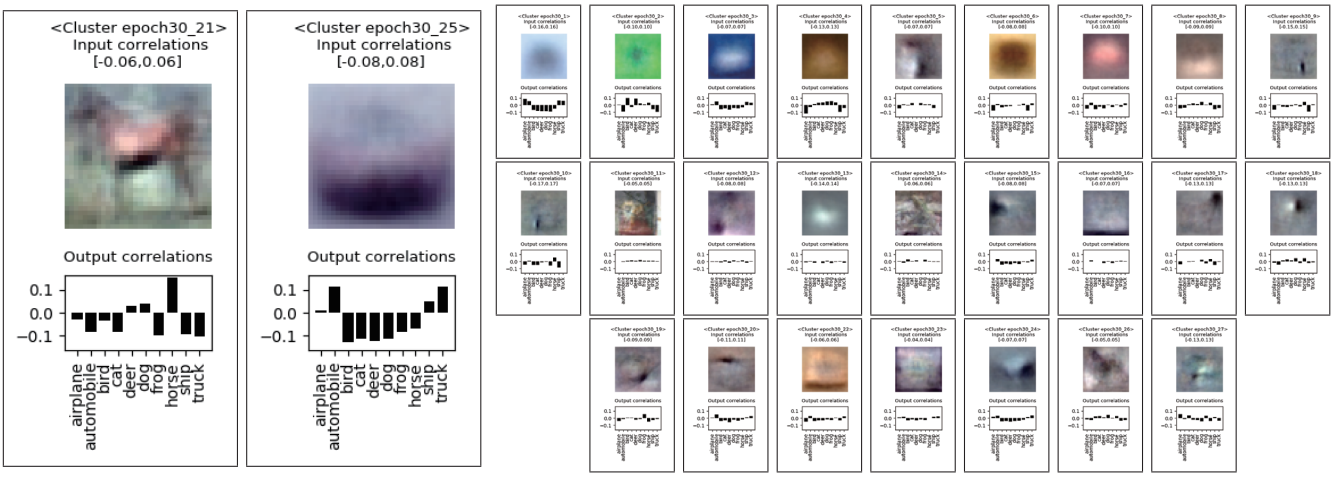

図 6: 30 エポックの学習完了時におけるニューラルネットの各クラスタのセントロイド．