知能プログラミング演習 I

第 7 回: 畳み込みニューラルネットワーク II

梅津 佑太
2 号館 404A: umezu.yuta@nitech.ac.jp

課題のダウンロード

前回作ったディレクトリに移動して今日の課題のダウンロードと解凍

step1: `cd ./DLL`

step2: `wget http://www-als.ics.nitech.ac.jp/~umezu/Lec7.zip`

step3: `unzip Lec7.zip`

✓ **まだ** DLL **のフォルダを作ってない人は**, step1 **の前に**
`mkdir -p DLL`
でフォルダを作成する

講義ノート更新しました.

今日の講義内容

1. **畳み込みニューラルネットワークの逆伝播**

前回の復習

畳み込みニューラルネットワークの構成要素:

- **畳み込み層**: **入力** $\mathbb{R}^{d \times d \times K} \ni Z \mapsto Z' \in \mathbb{R}^{d' \times d' \times M}$:

$$u_{ijm} = \langle W_m, Z_{ij} \rangle + b_m \mapsto z'_{ijm} = f(u_{ijm})$$

- **プーリング層**: **入力** $\mathbb{R}^{d \times d \times K} \ni Z \mapsto Z' \in \mathbb{R}^{d' \times d' \times K}$:

$$u_{ijk} = \langle W_{ijk}, Z_k \rangle \mapsto z'_{ijk} = u_{ijk}$$

  ✓ $W_{ijk} \in \mathbb{R}^{d \times d}$ **の作り方は後述**

- **全結合層**: **入力** $\mathbb{R}^{d \times d \times K} \ni Z \mapsto Z' \in \mathbb{R}^{d'}$:

$$u_i = \langle W_i, Z \rangle + b_i \mapsto z'_i = f(u_i)$$

ただし, $\langle \cdot, \cdot \rangle$ **はテンソルの内積**

$$\langle A, B \rangle = \sum_{ijk} a_{ijk} b_{ijk}, \quad A, B \in \mathbb{R}^{d \times d \times K}$$

$$\langle A, B \rangle = \sum_{ij} a_{ij} b_{ij} = \mathrm{tr}(A^\top B), \quad A, B \in \mathbb{R}^{d \times d}$$

逆伝播

- c.f., **これまでの逆伝播**: **適当なサイズの行列** W_{r+1} ($r+1$ **層目のパラメータ**) **と** δ_{r+1} **を用いて**,

$$\delta_r = W_{r+1}^\top \boldsymbol{\delta}_{r+1} \odot \nabla f(\boldsymbol{u}_r)$$
$$\Rightarrow \delta_{r,j} = \tilde{\boldsymbol{w}}_{r+1,j}^\top \boldsymbol{\delta}_{r+1} \nabla f(u_{r,j}) = \langle \tilde{\boldsymbol{w}}_{r+1,j}, \boldsymbol{\delta}_{r+1} \rangle \nabla f(u_{r,j})$$

とかけた [1]. **つまり**, δ_r **は** "$r+1$ **層目のパラメータ** $\boldsymbol{w}_{r+1,j}$ **と誤差** δ_{r+1} **に** r **層目の勾配** $\rangle \nabla f(u_{r,j})$ **の積を適当な順番に並べたもの**"

- **畳み込みニューラルネットワークの場合も適当なサイズのパラメータ** $W_{r+1,ijk}$ **や** δ_{r+1}, $\nabla f(u_{r,ijk})$ **を用いて**

$$\delta_{r,ijk} = \langle W_{r+1,ijk}, \boldsymbol{\delta}_{r+1} \rangle \nabla f(u_{r,ijk})$$

と書くことができる.

[1] $\tilde{w}_{r+1,j}$ **は** W_{r+1} **の第** j **列ベクトル**

逆伝播: 全結合層 I

全結合層: $Z_r \in \mathbb{R}^{d \times d \times K} \mapsto Z_{r+1} = f(U_r) \in \mathbb{R}^{d'}$

- **パラメータ** $W_r = (W_{r,1}, \dots, W_{r,d'}) \in \mathbb{R}^{d \times d \times K \times d'}$ **および** $\boldsymbol{b}_r = (b_{r,1}, \dots, b_{r,d'})^\top \in \mathbb{R}^{d'}$ **を用いて**,

$$u_{r,i} = \sum_{k=1}^{K} \sum_{i,j=1}^{d'} w_{r,ijk} z_{r,ijk} + b_{r,i} = \langle W_{r,i}, Z_r \rangle + b_{r,i}$$

- **これまで通り**, Z_{r+1} **が出力層なら**, **適当な誤差関数** E **と活性化関数** f **に対して**

$$\boldsymbol{\delta}_r = \frac{\partial E}{\partial \boldsymbol{u}_r} = f(\boldsymbol{u}_r) - \boldsymbol{y}$$

となる. **ただし**, $\boldsymbol{y}$ **は観測データの出力** (e.g., **多クラス分類なら** 1-of-K **表記のベクトル**)

逆伝播: 全結合層 II

- Z_{r+1} が出力層でない場合, 適当な大きさの行列 W_{r+1} に対して,

$$u_{r+1,i} = \boldsymbol{w}_{r+1,i}^{\top} Z_{r+1} + b_{r+1.i} = \boldsymbol{w}_{r+1,i}^{\top} f(U_r) + b_{r+1.i}$$

なので,

$$\begin{aligned}
\delta_{r,i} &= \frac{\partial E}{\partial u_{r,i}} = \sum_j \frac{\partial E}{\partial u_{r+1,j}} \frac{\partial u_{r+1,j}}{\partial u_{r,i}} = \sum_j \delta_{r+1,j} \frac{\partial \boldsymbol{w}_{r+1,j}^{\top} f(U_r)}{\partial u_{r,i}} \\
&= \sum_j \delta_{r+1,j} w_{r+1,ij} \nabla f(u_{r,i}) = \tilde{\boldsymbol{w}}_{r+1,i}^{\top} \boldsymbol{\delta}_{r+1} \nabla f(u_{r,i})
\end{aligned}$$

$$\Rightarrow \boldsymbol{\delta}_r = W_{r+1}^{\top} \boldsymbol{\delta}_{r+1} \odot \nabla f(\boldsymbol{u}_r)$$

逆伝播: プーリング層 I

プーリング層: $Z_r \in \mathbb{R}^{d\times d\times K} \mapsto Z_{r+1} \in \mathbb{R}^{d'\times d'\times K}$. **フィルタサイズ** s, **ストライド数** β **に対して**, $d' = (d-s)/\beta + 1$

- $P_{ij} = \{(\beta(i-1)+p, \beta(j-1)+q, k) \mid p, q = 1, \ldots, s\}$ **として**, $W_{r,ij} = (w_{r,ij,pqk})_{ij} \in \mathbb{R}^{d\times d}$ **を**

$$(\max \text{ プーリング}) \quad w_{r,ij,pqk} = \begin{cases} 1 & (p,q) = \arg\max_{(s,t)\in P_{ij}} z_{r,stk} \\ 0 & \text{その他} \end{cases}$$

$$(\text{average プーリング}) \quad w_{r,ij,pqk} = 1/s^2$$

とすれば, $i, j = 1, \ldots, d'$, $k = 1, \ldots, K$ **に対して**

$$Z_{r+1,ijk} = u_{r,ijk} = \langle W_{r,ij}, Z_{r,k} \rangle$$

- **このとき**, **プーリングの写像** g **は恒等写像と解釈する**.

$$\nabla g(u_{r,ijk}) = \left.\frac{\partial g(u_{r,pqr})}{\partial u_{r,pqr}}\right|_{u=u_{r,ijk}} = \begin{cases} 1 & i=p, j=q, k=r \\ 0 & \text{その他} \end{cases}$$

逆伝播: プーリング層 II

適当なサイズの W_{r+1}, $\boldsymbol{b}_{r+1}$ に対して,

$$u_{r+1,ijm} = \langle W_{r+1,m}, Z_{r+1,ij} \rangle + b_{r+1,m} = \langle W_{r+1,m}, U_{r,ij} \rangle + b_{r+1,m}$$

とする[2]. $U_{r,ijk} = (u_{r,\beta'(i-1)+p,\beta'(j-1)+q,k})_{p,q}$ に対して, $U_{r,ij} = (U_{r,ijk})_k$ を 3 次元配列とすれば, 誤差 $\delta_{r,ijk}$ は

$$\begin{aligned}
\delta_{r,ijk} = \frac{\partial E}{\partial u_{r,ijk}} &= \sum_{s,t,l} \frac{\partial E}{\partial u_{r+1,stl}} \frac{\partial u_{r+1,stl}}{\partial u_{r,ijk}} \\
&= \sum_{s,t,l} \delta_{r+1,stl} \frac{\partial \langle W_{r+1,l}, U_{r,st} \rangle}{\partial u_{r,ijk}} = \sum_{s,t,l} \delta_{r+1,stl} w_{r+1,i-\beta'(s-1),j-\beta'(t-1),k,l}
\end{aligned}$$

[2] $r+2$ 層目が全結合層ならば, $u_{r+1,m} = \langle W_{r+1,m}, U_r \rangle + b_{r+1,m}$ となる.

逆伝播: プーリング層 III

$$\delta_{r,ijk} = \sum_{s,t,l} \delta_{r+1,stl} w_{r+1,i-\beta'(s-1),j-\beta'(t-1),k,l}$$

は, $\boldsymbol{\delta}_{r+1} = (\delta_{r+1,stl})_{s,t,l}$, $W_{r+1,ijk} = \left(w_{r+1,i-\beta'(s-1),j-\beta'(t-1),k,l}\right)_{s,t,l}$ とすれば,

$$\delta_{r,ijk} = \langle \boldsymbol{\delta}_{r+1}, W_{r+1,ijk} \rangle$$

とかける. したがって, $\delta = (\delta_{r,ijk})_{i,j,k}$ とすれば, これがプーリング層の誤差となる.

畳み込み層: **パディング数** α **で畳み込んだ後の入力を** $Z_r \in \mathbb{R}^{d\times d\times K}$ **として** [3], $Z_r \mapsto Z_{r+1} \in \mathbb{R}^{d'\times d'\times M}$. **ただし**, **フィルタサイズ** H, **ストライド数** β **に対して**, $d' = (d-H)/\beta + 1$.

- **パラメータ** $W_r = (W_{r,1}, \dots, W_{r,d'}) \in \mathbb{R}^{d\times d\times K\times M}$ **および** $\boldsymbol{b}_r = (b_{r,1}, \dots, b_{r,M})^\top \in \mathbb{R}^M$ **を用いて**,

$$
\begin{aligned}
u_{r,ijm} &= \sum_{k=1}^{K}\sum_{p,q=1}^{d'} w_{r,pqkm} z_{r,\beta(i-1)+p,\beta(j-1)+q,k} + b_{r,m} \\
&= \langle W_{r,m}, Z_{r,ij}\rangle + b_{r,m}
\end{aligned}
$$

ただし, $Z_{r,ijk} = (z_{r,ij,\beta(i-1)+p,\beta(j-1)+q,k})_{p,q}$ **に対して**, $Z_{r,ij} = (Z_{r,ij1}, \dots, Z_{r,ijK}) \in \mathbb{R}^{H\times H\times K}$

[3] $X \in \mathbb{R}^{p\times p}$ **をパディング数** α **で畳み込むと** $d = p + 2\alpha$ **となる**.

誤差の更新

$$u_{r,ijm} = \langle W_{r,m}, Z_{r,ij} \rangle + b_{r,m}$$

が, プーリング層の順伝播と一致することに注意すれば, 適当な $\boldsymbol{\delta}_{r+1} = (\delta_{r+1,stl})_{s,t,l}$, $W_{r+1,ijk} = (w_{r+1,i-\beta'(s-1),j-\beta'(t-1),k,l})_{s,t,l}$ を用いて,

$$\delta_{r,ijk} = \langle \boldsymbol{\delta}_{r+1}, W_{r+1,ijk} \rangle$$

とかける. したがって, $\delta = (\delta_{r,ijk})_{i,j,k}$ とすれば, これが畳み込み層の誤差となる.

パラメータの更新ルール

δ_r が計算できると，パラメータの更新ルールを求めることができる：つまり，合成関数の微分を用いて，

$$\frac{\partial E}{\partial w_{r,pqkm}} = \sum_{s,t=1}^{H} \frac{\partial E}{\partial u_{r,stm}} \frac{\partial u_{r,stm}}{\partial w_{r,pqkm}} = \sum_{s,t=1}^{H} \delta_{r,stm} z_{r,\beta(s-1)+p,\beta(t-1)+q,k}$$

$$\frac{\partial E}{\partial b_{r,m}} = \sum_{s,t=1}^{H} \frac{\partial E}{\partial u_{r,stm}} \frac{\partial u_{r,stm}}{\partial b_{r,m}} = \sum_{s,t=1}^{H} \delta_{r,stm}$$

より，

$$\frac{\partial E}{\partial W_{r,m}} = \sum_{s,t=1}^{H} \delta_{r,stm} Z_{r,st} \quad \text{および，} \quad \frac{\partial E}{\partial \boldsymbol{b}_r} = \sum_{s,t=1}^{H} \boldsymbol{\delta}_{r,st}$$

となる．これを用いて，確率的勾配降下法や Adam のアルゴリズムを実装すれば良い．