EECS 182 Deep Neural Networks
Fall 2022 Anant Sahai

Homework 6

This homework is due on Saturday, Nov 5, 2022, at 10:59PM.

Deliverables: Please submit the code/notebooks to the code gradescope assignment. Submit your written answers in the written gradescope assignment, and attach a pdf printout of the notebook.

1. Redo the Midterm

Please redo the midterm exam (linked at https://static.us.edusercontent.com/files/hQ7lNcAn75o6lyD0KMtVssl7) as homework. Attach your answers to the written part of this homework.

Solution: Solutions for these will be released separately.

2. Implementing Simple Transformers

(a) Implementing a simple transformer. Fill out Section (a) in the notebook. No written portion.

(b) Designing a transformer that selects by content. Fill out Section (b) in the notebook and **comment on the similarities and differences between the weights and intermediate outputs of the learned and hand-coded model.**

Solution: Answers may vary, but weights, keys, and queries are likely more evenly distributed than those the student implemented. However, Vm, attention scores and values should be roughly the same.

Explanation: this occurs since the network is simple enough that there is only one correct attention pattern (only attend to matching positions) and only one correct thing to do with attended positions (use them to copy the original content into the output). In more complicated problems it's rare to be able to predict precisely what features the transformer learns.

(c) Designing a transformer that selects by position. Fill out Section (b) in the notebook and **comment on the similarities and differences between the weights and intermediate outputs of the learned and hand-coded model.**

Solution: Answers may vary, but weights, keys, and queries are likely more evenly distributed than those the student implemented. However, attention scores and outputs should be roughly the same, since there is only one correct attention pattern (only attend to the firs position) and only one correct thing to do with attended positions (use them to copy the first element into the output).

(d) (optional) Designing a transformer that selects by position. Fill out Section (b) in the notebook and **comment on the similarities and differences between the weights and intermediate outputs of the learned and hand-coded model.**

Solution: Answers may vary, but all transformer outputs may look substantially different between the two models except the sign of the final output should be the same. This shows us that transformers can solve tasks in ways which are unintuitive to humans.

Please submit your notebook and your written answers to Gradescope.

 All Rights Reserved. This may not be publicly shared without explicit permission.

3. Regularization and Dropout [Optional]

(This is an extension on Q11 on the midterm.)

You saw one perspective on the implicit regularization of dropout in HW, and here, you will see another one. Recall that linear regression optimizes the following learning objective:

$$\mathcal{L}(\mathbf{w}) = ||\mathbf{y} - X\mathbf{w}||_2^2 \tag{1}$$

One way of using *dropout* during SGD on the d-dimensional input features $\mathbf{x}_i$ involves keeping each feature at random $\sim_{i.i.d}$ $Bernoulli(p)$ (and zeroing it out if not kept) and then performing a traditional SGD step.

It turns out that such dropout makes our learning objective effectively become

$$\mathcal{L}(\check{\mathbf{w}}) = E_{R\sim Bernoulli(p)}\left[||\mathbf{y} - (R \odot X)\check{\mathbf{w}}||_2^2\right] \tag{2}$$

where $\odot$ is the element-wise product and the random binary matrix $R \in \{0,1\}^{n\times d}$ is such that $R_{i,j} \sim_{i.i.d}$ $Bernoulli(p)$. We use $\check{\mathbf{w}}$ to remind you that this is learned by dropout.

Recalling how Tikhonov-regularized (generalized ridge-regression) least-squares problems involve solving:

$$\mathcal{L}(\mathbf{w}) = ||\mathbf{y} - X\mathbf{w}||_2^2 + ||\Gamma\mathbf{w}||_2^2 \tag{3}$$

for some suitable matrix Γ,

show that we can manipulate (2) to eliminate the expectations and get:

$$\mathcal{L}(\check{\mathbf{w}}) = ||\mathbf{y} - pX\check{\mathbf{w}}||_2^2 + p(1-p)||\check{\Gamma}\check{\mathbf{w}}||_2^2 \tag{4}$$

with $\check{\Gamma}$ being a diagonal matrix whose j-th diagonal entry is the norm of the j-th column of the training matrix X.

 All Rights Reserved. This may not be publicly shared without explicit permission.

Solution: let P = R • X where • is the element-wise multiplication. Therefore, we have:

$$||y - Pw||_2^2 = y^T y - 2w^T P^T y + w^T P^T P w \tag{5}$$

That is:

$$\mathcal{E}_{R\sim Bernoulli(p)}[||y - R \bullet Xw||_2^2] = E_R[y^T y - 2w^T P^T y + w^T P^T P w] \tag{6}$$

Since the expected value of a matrix is the matrix of the expected value of its elements, we have that

$$(E_{R\sim Bernoulli(p)}[P])_{ij} = E_{R\sim Bernoulli(p)}[(R \bullet X)_{ij}] = X_{ij} E_{RBernoulli(p)}[R_{ij}] = pX_{ij} \tag{7}$$

It follows that:

$$2w^T P^T y = 2pw^T X^T y \tag{8}$$

and:

$$(P^T P)_{ij} = \Sigma_{k=1}^N E_{RBernoulli(p)}[R_{ki} R_{kj} X_{ki} X_{kj}] \tag{9}$$

where if i!=j then they are independent so the off-diagonal elements results in $p^2(X^T X)_{ij}$ and if i == j, we get:

$$(P^T P)_{ij} = \Sigma_{k=1}^N E_{R\sim Bernoulli(p)}[R_{ki}^2 X_{ki}^2] = p(X^T X)_{ij} \tag{10}$$

we now can put everything together as follow:

$$\begin{aligned}
\mathcal{L}(w) &= E_{R\sim Bernoulli(p)}[||y - R \bullet Xw||_2^2] && (11)\\
&= E_{R\sim Bernoulli(p)}[y^T y - 2w^T P^T y + w^T P^T P w] && (12)\\
&= E_{R\sim Bernoulli(p)}[y^T y - 2w^T P^T y + p^2 w^T X^T X w - p^2 w^T X^T X w + w^T P^T P w] && (13)\\
&= E_{R\sim Bernoulli(p)}[(y^T y - 2w^T P^T y + p^2 w^T X^T X w) - p^2 w^T X^T X w + w^T P^T P w] && (14)\\
&= ||y - pXw||_2^2 - p^2 w^T X^T X w + w^T E_{R\sim Bernoulli(p)}[P^T P] w && (15)\\
&= ||y - pXw||_2^2 + w^T E_{R\sim Bernoulli(p)}[(P^T P - p^2 X^T X)] w && (16)\\
&= ||y - pXw||_2^2 + w^T (E_{R\sim Bernoulli(p)}[P^T P] - p^2 X^T X) w && (17)\\
&= ||y - pXw||_2^2 + w^T (p(1-p) diag(X^T X) w && (18)\\
&= ||y - pXw||_2^2 + p(1-p)||\Gamma w||_2^2 && (19)\\
& && (20)
\end{aligned}$$

where $\Gamma = (diag(X^T X))^{1/2}$ **Solution:** We have for $w = \frac{w_{new}}{p}$, $\lambda = \frac{1-p}{p}$ and $\Gamma_{new} = \sqrt{\lambda}\Gamma$:

$$\begin{aligned}
\mathcal{L}(w) &= ||y - pXw||_2^2 + p(1-p)||\Gamma w||_2^2 && (21)\\
&= ||y - pX\frac{w_{new}}{p}||_2^2 + p(1-p)||\Gamma \frac{w_{new}}{p}||_2^2 && (22)\\
&= ||y - Xw_{new}||_2^2 + \frac{(1-p)}{p}||\Gamma w_{new}||_2^2 && (23)\\
&= ||y - Xw_{new}||_2^2 + \lambda||\Gamma w_{new}||_2^2 && (24)
\end{aligned}$$

 All Rights Reserved. This may not be publicly shared without explicit permission.

$$= ||y - Xw_{new}||_2^2 + ||\sqrt{\lambda}\Gamma w_{new}||_2^2 \quad (25)$$
$$= ||y - Xw_{new}||_2^2 + ||\Gamma_{new} w_{new}||_2^2 \quad (26)$$
$$(27)$$

where $\Gamma_{new} = \sqrt{\lambda}(diag(X^T X))^{1/2}$

4. Homework Process and Study Group

Citing sources and collaborators are an important part of life, including being a student!
We also want to understand what resources you find helpful and how much time homework is taking, so we can change things in the future if possible.

(a) **What sources (if any) did you use as you worked through the homework?**

(b) **If you worked with someone on this homework, who did you work with?**
List names and student ID's. (In case of homework party, you can also just describe the group.)

(c) **Roughly how many total hours did you work on this homework? Write it down here where you'll need to remember it for the self-grade form.**

Contributors:

- All CS182 staff.
- Olivia Watkins.
- Jake Austin.
- Anant Sahai.
- Saagar Sanghavi.
- Jerome Quenum.

 All Rights Reserved. This may not be publicly shared without explicit permission.