

Homework 2 Part 1

An Introduction to Convolutional Neural Networks

11-785: INTRODUCTION TO DEEP LEARNING (SPRING 2023)

OUT: **Feb 17, 2023**

EARLY SUBMISSION BONUS: **Mar 4, 2023, 11:59 PM, Eastern Time**

DUE: **Mar 17, 2023, 11:59 PM, Eastern Time**

Start Here

- **Collaboration policy:**

- You are expected to comply with the [University Policy on Academic Integrity and Plagiarism](#).
- You are allowed to talk with / work with other students on homework assignments
- You can share ideas but not code, you must submit your own code. All submitted code will be compared against all code submitted this semester and in previous semesters using [MOSS](#).

- **Directions:**

- Do not use any auto-differentiation toolboxes (PyTorch, TensorFlow, Keras, etc) - you are only permitted and recommended to vectorize your computation using the Numpy library.
- We recommend that you look through all of the problems before attempting the first problem. However we do recommend you complete the problems in order, as the difficulty increases, and questions often rely on the completion of previous questions.

Homework objectives

If you complete this homework successfully, you would ideally have learned:

- How to write code to implement a CNN from scratch
 - How to implement convolutional layers
 - How to implement pooling layers
 - How to implement downsampling and upsampling layers
 - How to chain these up, along with components you have already implemented in HW1P1 to compose a CNN of any size
- Your code will be able to perform forward inference through the CNNs
- How to write code to implement *training* of your CNN
 - How to perform a forward pass through your network
 - How to implement backpropagation through the convolutional layers
 - How to implement backpropagation through the pooling layers
 - How to implement backpropagation through resampling layers
 - How to combine these to perform backpropagation through an entire CNN to compute gradients (to train all parameters of the network)

Checklist

Here is a checklist page that you can use to keep track of your progress as you go through the write-up and implement the corresponding sections in your starter notebook. As you complete each function in the notebook, you can check the corresponding boxes aligned with each section.

1. Getting Started

Download code handout and extract the file

Activate the conda environment created in HW1P1

Copy your `linear.py`, `activation.py` and `loss.py` from HW1P1 to `mytorch` fol.

Read the whole assignment write-up for an overview

2. Complete the Components of a CNN

Complete Resampling for 1D and 2D data [10 Points]

Complete the following Convolutional layers

 Complete `Conv1d_stride1` and use it to build `Conv1d` [15 Points]

 Complete `Conv2d_stride1` and use it to build `Conv2d` [15 Points]

 Implement `ConvTranspose1d` and `ConvTranspose2d` [10 Points]

Complete `MaxPool2d` and `MeanPool2d` [30 Points]

Complete `Flatten` layer

3. Putting everything together to make CNN models

Convert Scanning MLPs to CNNs [5 Points]

Build a CNN model using the parts we coded in previous sections [5 Points]

4. Hand-in

Make sure you pass all test cases in the local autograder

Make the `handin.tar` file and submit to autolab

Contents

1 Setup and Submission	4
2 Notation	5
3 Introduction	6
3.1 Structure of a CNN	6
3.2 1D vs 2D	7
4 Multiple Choice [5 Points]	8
5 Resampling [10 Points]	12
5.1 Upsample1d	12
5.1.1 Upsample1d Forward	12
5.1.2 Upsample1d Backward	13
5.2 Downsampling1d	13
5.2.1 Downsample1d Forward	13
5.2.2 Downsample1d Backward	14
5.3 Upsampling2d	15
5.3.1 Upsampling2d Forward	15
5.3.2 Upsample2d Backward	15
5.4 Downsampling2d	16
5.4.1 Downsample2d Forward	16
5.4.2 Downsample2d Backward	17
6 Convolutional layer	18
6.1 Conv1d_stride1 [10 Points]	19
6.1.1 Conv1d_stride1 Forward	20
6.1.2 Conv_stride1 Backward	21
6.2 Conv1d [5 points]	23
6.3 Conv2d_stride1 [10 Points]	24
6.3.1 Conv2d_stride1 Forward	25
6.3.2 Conv2d_stride1 Backward	27
6.3.3 Conv2d [5 points]	29
6.4 Transposed Convolution [10 points]	32
6.4.1 ConvTranspose1d	32
6.4.2 ConvTranspose2d	34
7 Pooling	36
7.1 MaxPool2d_stride1 and MeanPool2d_stride1 [20 Points]	36
7.2 MaxPool2d and MeanPool2d [10 Points]	37
8 Flatten layer	38
9 Converting Scanning MLPs to CNNs	39
9.1 CNN as a Simple Scanning MLP [5 Points]	39
9.2 CNN as a Distributed Scanning MLP [5 Points]	40
10 Build a CNN model [5 Points]	43
11 Appendix	44
11.1 Scanning MLP : Illustration	44
11.2 Numpy TensorDot	47

1 Setup and Submission

In this assignment, you will continue to develop your own version of PyTorch, which is of course called *MyTorch* [©] (still a brilliant name; a master stroke. Well done!). In addition, you'll convert two scanning MLPs to CNNs and build a CNN model. For Homework 2, *MyTorch* [©] will have the following file structure:

```
HW2P1
└── mytorch
    ├── Conv1d.py
    ├── Conv2d.py
    ├── ConvTranspose.py
    ├── flatten.py
    ├── pool.py
    ├── resampling.py
    ├── linear.py (Copy your file from HW1P1)
    ├── activation.py (Copy your file from HW1P1)
    └── loss.py (Copy your file from HW1P1)
└── models
    ├── mlp.py
    ├── mlp_scan.py
    └── cnn.py
└── MCQ
    └── mcq.py
└── autograder
    ├── helpers.py
    ├── test.py
    ├── hw2_autograder.py
    └── runner.py
└── create_tarball.sh
```

- **Environment** Activate the environment created in HW1P1 for all necessary packages:

```
conda activate idls23
```

- **(IMPORTANT)** **Copy** your completed `linear.py`, `activation.py`, `loss.py` from HW1P1, ensure that you received full marks for these sections. (The files in your homework are structured in such a way that you can easily import and reuse modules of code for your subsequent homeworks :D)
- **Hand-in** your code by running the following command from the top level directory, then **SUBMIT** the created `handin.tar` file to autolab:

```
sh create_tarball.sh
```

- **Autograde** your code by running the following command from the top level directory:

```
python3 autograder/runner.py
```

- **Debugging Tool**

Use the interactive python debugger `pdb`¹ to debug your code effectively.

- In your code, set breakpoints by adding this line: `pdb.set_trace()`
- Run the following line from top level directory to turn on `pdb` with your local autograder:

```
python3 -m pdb autograder/runner.py
```

¹Useful commands: [pdb cheatsheet](#)

2 Notation

**Numpy Tips:

- Use $A * B$ for element-wise multiplication $A \odot B$.
- Use $A @ B$ for mATrix multiplication $A \cdot B$.
- Use A / B for element-wise division $A \oslash B$.

Linear Algebra Operations

A^T	Transpose of A
$A \odot B$	Element-wise (Hadamard) Product of A and B
$A \cdot B$	Matrix multiplication of A and B
$A \oslash B$	Element-wise division of A and B
$A \circledast B$	Matrix convolution of A and B

Functions and Operations

$\log(x)$	Natural logarithm of x
$\sigma(x)$	Sigmoid, $\frac{1}{(1 + \exp^{-x})}$
$\tanh(x)$	Hyperbolic tangent, $\frac{e^x - e^{-x}}{e^x + e^{-x}}$
$\max_{\mathbb{S}} f$	The operator $\max_{a \in \mathbb{S}} f(a)$ returns the highest value $f(a)$ for all elements in the set \mathbb{S}
$\arg\max_{\mathbb{S}} f$	The operator $\arg\max_{a \in \mathbb{S}} f(a)$ returns the element a of the set \mathbb{S} that maximizes $f(a)$
$\sigma(x)$	Softmax function, $\sigma : \mathbb{R}^K \rightarrow (0, 1)^K$ and $\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}}$ for $i = 1, \dots, K$

Calculus

$\frac{dy}{dx}$	Derivative of scalar y with respect to scalar x
$\frac{\partial y}{\partial x}$	Partial derivative of scalar y with respect to scalar x
$\frac{\partial f(Z)}{\partial Z}$	Jacobian matrix $\mathbf{J} \in \mathbb{R}^{N \times M}$ of $f : \mathbb{R}^M \rightarrow \mathbb{R}^N$

3 Introduction

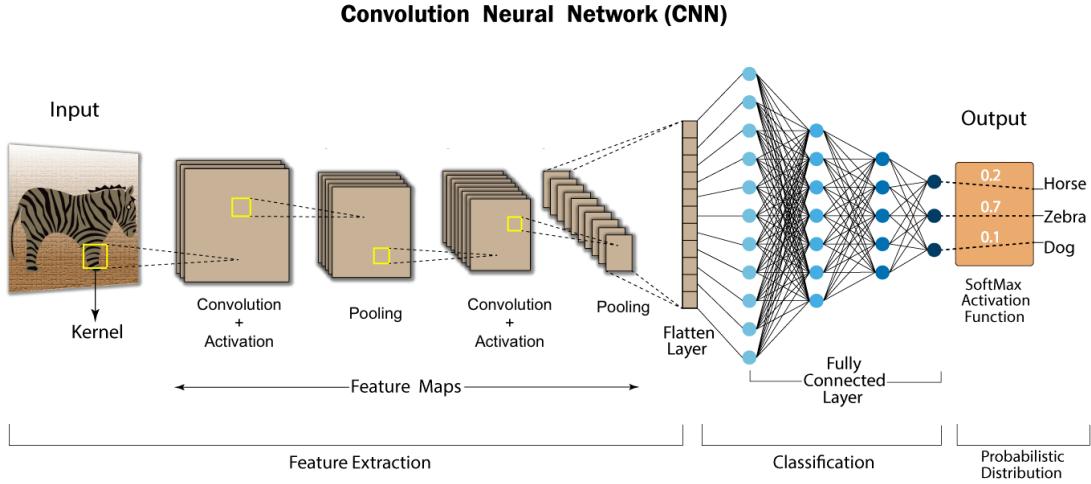


Figure 1: Standard 2d CNN for image classification

A Convolutional Neural Network is a **position-invariant** pattern detector and a special kind of FFNN that significantly reduces the number of parameters in a deep neural network with many units without losing too much in the quality of the model. CNNs have wide-ranging use-cases in image classification, object detection, semantic segmentation, image captioning, natural language processing, forecasting, and more. The foundational concept that it is based on is the concept of convolution. It is recommended to understand the convolution operation to fully appreciate the inspiration behind using CNNs. A good reference by 3B1B can be found here: [But what is a convolution?](#)

3.1 Structure of a CNN

A Convolutional Neural Network consists of the following parts:

- **Convolutional Layer**

This first layer is the Convolutional Neural Network's namesake. A convolutional layer consists of many filters, where each filter is responsible for capturing a different pattern. The filters in this layer are used to extract features from the input. For example, in images of cars, a convolutional layer might extract the "wheel" patterns. Another convolutional layer's filters might extract the patterns formed by the frame of the car images. Yet another convolutional layer's filters might extract the patterns formed by the bumper/license plate in the front of the car images. Check out [CNN Explainer](#) for a more visual explanation about CNNs!

The relationship between Convolutional Layer and CNN is the same as the relationship between Linear Layer and MLP.

- **Resampling (Upsampling and Downsampling)**

Resampling consists of upsampling and downsampling. To put it simply, upsampling is a way to "bloat" or fill up our input and make it bigger. Conversely, downsampling is a way to reduce the size of our input and make it smaller, while preserving the information that we need.

Why are these necessary? It is straight forward to implement Conv Layer with `stride = 1`². Let's say you have implemented `Conv_stride1` class and want to generalize your implementation to work for `stride != 1`. In that case, you can either modify your Conv Layer implementation to work for `stride != 1`, OR you can leave your implementation untouched, and tack on resampling to the output of the

²Stride is the number of pixel the filter passes in each step

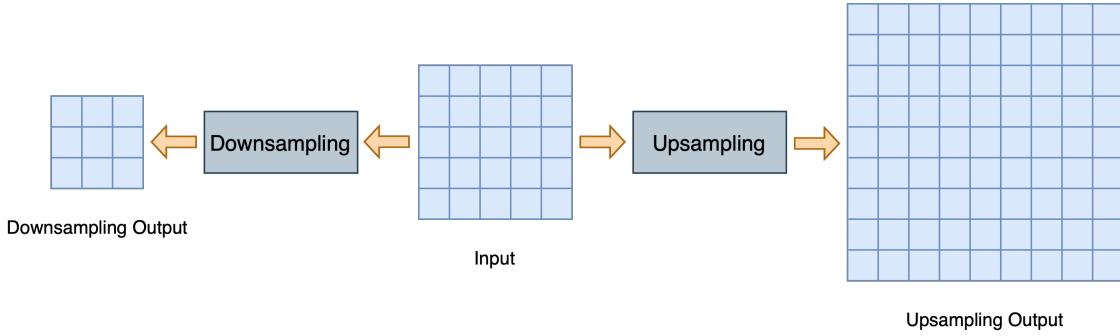


Figure 2: Resampling 2d input

`Conv_stride1` to “simulate” the stride. To understand resampling, let’s concretely break up the Conv Layer into 3 main cases:

- Stride = 1: In this case, after our `Conv_stride1` layer, we will simply proceed on to Pooling.
- Stride > 1: we need to DOWNSAMPLE the output of our `Conv_stride1` layer.
- Stride < 1: we need to UPSAMPLE the input and pass it to our `Conv_stride1` layer.

- **Activation**

As we’ve seen extensively in HW 1, we will now apply an activation to the output of our convolutional layer.

- **Pooling**

Next, we have a pooling layer. Pooling is used for jitter invariance. Conventionally, pooling is done with strides > 1 so that there is also a reduction in size of the output map. But why do we need to decrease the size? This reduces computational costs, and can help increase speed of the network. A decrease in size means less work for the machine, right? :)

- **Flatten** After obtaining all the feature maps, we want to use the features to do a classification task. Now we need to flatten our pooled feature maps before passing into the classification layers. This part is quite simple – just flatten our output so far into a single 1-dimensional array.

- **Classification Layer**

- **Linear Layer:** Now, we’ll insert this long vector of data (consisting of flattened, pooled feature maps) into a linear layer. This fully connected linear layer will allow our network to move towards performing a classification, i.e. map the representation between the input and the output.
- **Softmax Layer:** ³ Victory at last. We have reached the softmax layer, which will allow for us to output a probability distribution over our output classes. We will use this to allow our CNN to perform a classification.

3.2 1D vs 2D

1D Convolutional Neural Networks (`Conv1D`) are used for processing sequences of **one-dimensional signals**, such as **time-series data or text data represented as a sequence of word embeddings**.

2D Convolutional Neural Networks (`Conv2D`) are used for processing two-dimensional data, such as **images**.

For Conv Layer class and Resampling classes, we will implement both the 1D and 2D versions.

³Note that Softmax layer is not always required, for HW2P2, if you are using CrossEntropy loss, it expects logits instead of probability.

4 Multiple Choice [5 Points]

These questions are intended to give you major hints throughout the homework. Answer the questions by returning the correct letter as a string in the corresponding question function in **MCQ/mcq.py**. Each question has only a single correct answer. Verify your solutions by running the local **autograder**. To get credit (5 points), you must answer **all** questions correctly.

- (1) **Question 1:** Given the following architecture of a scanning MLP, what are the parameters of the equivalent Time Delay Neural Network which uses convolutional layers?

As you have seen in the lectures, a convolutional layer can be viewed as an MLP which scans the input. This question illustrates an example of how the parameters are shared between a scanning MLP and an equivalent convolutional nework for 1 dimensional input. (Help⁴)⁵[1 point]

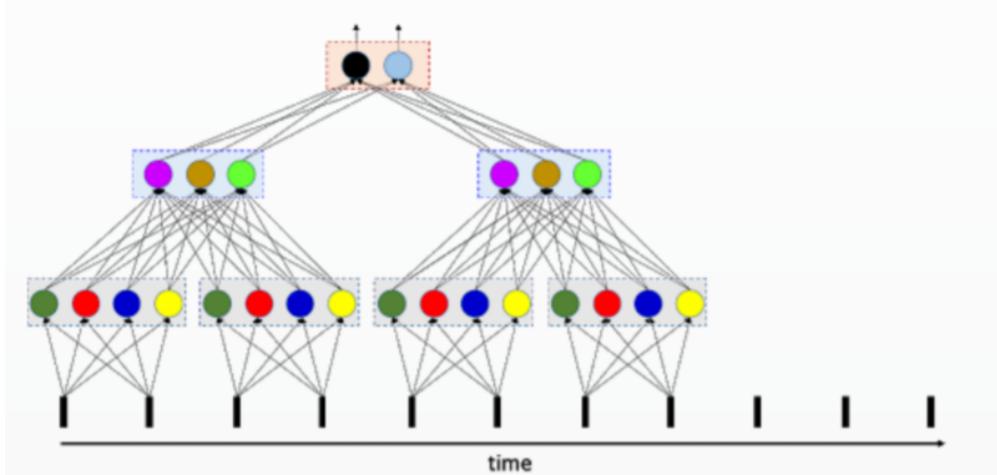


Figure 3: The architecture of a scanning MLP

- (A) The first hidden layer has 4 filters of kernel-width 2 and stride 2; the second layer has 3 filters of kernel-width 8 and stride 2; the third layer has 2 filters of kernel-width 6 and stride 2
 - (B) The first hidden layer has 4 filters of kernel-width 2 and stride 2; the second layer has 3 filters of kernel-width 2 and stride 2; the third layer has 2 filters of kernel-width 2 and stride 2
 - (C) The first hidden layer has 2 filters of kernel-width 4 and stride 2; the second layer has 3 filters of kernel-width 2 and stride 2; the third layer has 2 filters of kernel-width 2 and stride 2
- (2) **Question 2:** Considering padding and dilation, which equations below are equivalent for calculating the out dimension (width) for a 1D convolution (L_{out} at CONV1D)? [1 point]

```

eq 1: out_width = [in_width + 2 * padding - dilation * (kernel - 1) - 1] // stride + 1
eq 2: out_width = [(in_width_padded - kernel_dilated) // stride] + 1, where in_width_padded
      = in_width + 2 * padding, kernel_dilated = (kernel - 1) * (dilation - 1) + kernel
eq 3: great_baked_potato = (2*potato + 1*onion + celery//3 + love)**(sour cream)

```

- (A) eq 1 is the one and only true equation
- (B) eq 2 is the one and only true equation
- (C) eq 1, 2, and 3 are all true equations
- (D) eq 1 and 2 are both true equations

⁴Allow the input layer to be of an arbitrary length. The shared parameters should scan the entire length of the input with a certain repetition. In the first hidden layer, the horizontal gray boxes act as the black lines from the input layer. Why? Think...

⁵Understanding this question will help you with 3.3 and 3.4.

- (3) **Question 3:** In accordance with the image below, choose the correct values for the corresponding channels, width, and batch size given stride = 2 and padding = 0? [1 point]

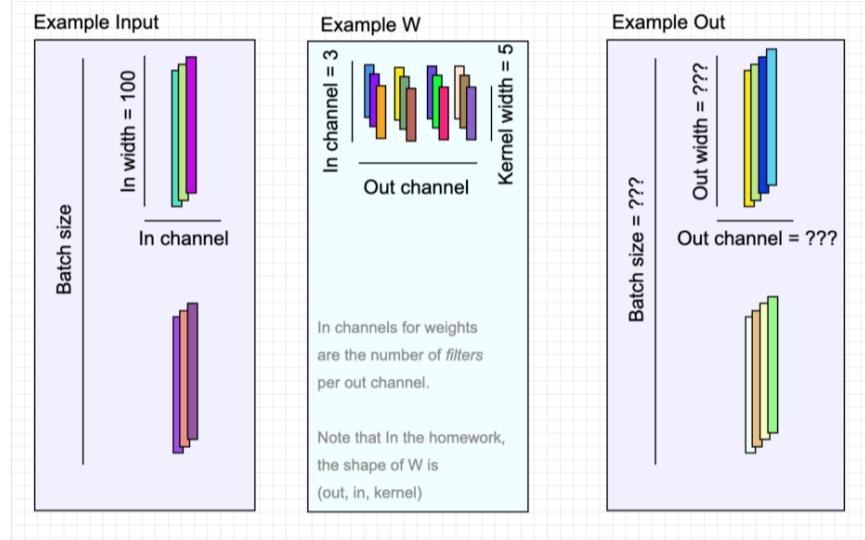


Figure 4: Example dimensions resulting from a 1d Convolutional layer

(A)

Example Input: Batch size = 2, In channel = 3, In width = 100
 Example W: Out channel = 4, In channel = 3, Kernel width = 5
 Example Out: Batch size = 2, Out channel = 4, Out width = 20

(B)

Example Input: Batch size = 2, In channel = 3, In width = 100
 Example W: Out channel = 4, In channel = 3, Kernel width = 5
 Example Out: Batch size = 2, Out channel = 4, Out width = 48

- (4) **Question 4:** Explore the usage of the command `numpy.tensordot`. Run the following example and verify how this command combines broadcasting and inner products efficiently. What is the shape of C and the value of C[0,0]? [1 point]

```
A = np.arange(30.).reshape(2,3,5)
B = np.arange(24.).reshape(3,4,2)
C = np.tensordot(A,B, axes = ([0,1],[2,0]))
```

- (A) [5,4] and 820
 (B) [4,5] and 1618

(5) **Question 5:** This question will help you understand why CNN is helpful and powerful. Please choose the listing properties of CNN.[1 Point]

(For more details, you can refer to <https://www.deeplearningbook.org/contents/convnets.html>)

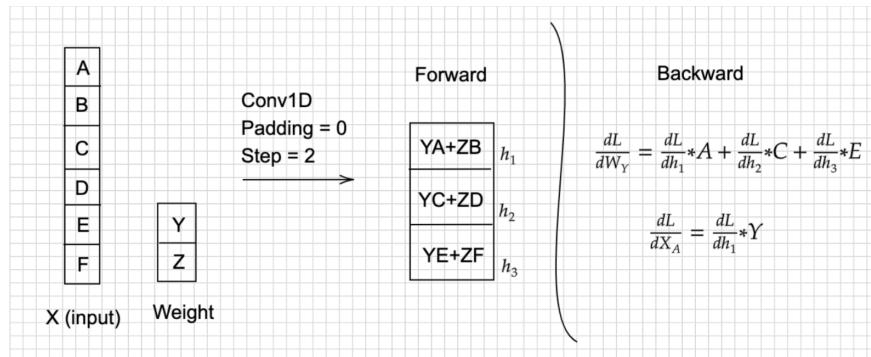
(1) Sparse Interactions: It means every output interacts with every input units. This is accomplished by making the kernel smaller than the input. So that we only needs fewer parameters and fewer operations. Like processing images, we only needs to detect small features, like the edges.

(2) Parameter Sharing: It refers to using the same parameters for more than one function in a model. One can say that the network has tied weights because the value of the weight applied to one input is tied to the value of a weight applied elsewhere

(3) Equivalent Representation: In the case of convolution, the parameter sharing causes the layer to have a property called equivalent representation to translation. Equivalent means if the inputs changes, the output changes in the same way.

- (A) (1) is correct
- (B) (1) and (2) are correct
- (C) (2) and (3) are correct
- (D) all of them are correct

(6) **Question 6:** Given the toy example below, what are the correct values for the gradients? If you are still confused about backpropagation in CNNs watch the lectures or Google backpropagation with CNNs. [0 Point ⁶]



- (A) I have read through the toy example and now I understand backpropagation with Convolutional Neural Networks.
- (B) This whole baked potato trend is really weird.
- (C) Seriously, who is coming up with this stuff?
- (D) Am I supposed to answer A for this question, I really don't understand what is going on anymore?

⁶Question 6 and 7 carries 0 marks, but feel free to ask TAs if you have any questions about them.

(7) **Question 7:** This question will help you visualize the CNN as a scanning MLP. Given a weight matrix for a scanning MLP, you want to determine the corresponding weights of the filter in its equivalent CNN. $W_{MLP}(input_size, output_size)$ is the weight matrix for a layer in a MLP and $W_{conv}(out_channel, in_channel, kernel_size)$ is the equivalent convolutional filter. As discussed in class, you want to use the convolutional layer instead of scanning the input with an entire MLP, then how will you modify W_{MLP} to find W_{conv} [0 Point]

(To help you visualize this better, you can refer to Figure 3 and the lecture slides)

- (A) $W_{conv} = W_{MLP}.reshape(input_channel, kernel_size, output_channel).transpose()$
where $input_channel * kernel_size = input_size$ and $output_size = output_channel$
- (B) $W_{conv} = W_{MLP}.reshape(kernel_size, input_channel, output_channel).transpose()$
where $input_channel * kernel_size = input_size$ and $output_size = output_channel$
- (C) $W_{conv} = W_{MLP}.reshape(output_channel, input_channel, kernel_size)$
where $input_channel * kernel_size = input_size$ and $output_size = output_channel$
- (D) Am I supposed to answer A for this question, I really don't understand what is going on anymore?

5 Resampling [10 Points]

Before moving to convolution, we will take a look at resampling operations which will make convolutions with strides $! = 1$ more intuitive.

5.1 Upsample1d

In this section, your task is to implement the `Upsample1d` class. Please consider the following class structure.

```
class Upsample1d:

    def __init__(self, upsampling_factor):
        self.upsampling_factor = upsampling_factor

    def forward(self, A):
        Z = # TODO

        return Z

    def backward(self, dLdZ):
        dLdA = # TODO

        return dLdA
```

Table 1: Upsample1d Class Components

Code Name	Math	Type	Shape	Meaning
<code>batch_size</code>	N	scalar	-	batch size
<code>in_channels</code>	C	scalar	-	Number of channels
<code>input_width</code>	W_{in}	scalar	-	Width of input channels
<code>output_width</code>	W_{out}	scalar	-	Width of output channels
<code>upsampling_factor</code>	k	scalar	-	$\text{upsampling_factor} \in \mathbb{Z}^+$
<code>A</code>	A	matrix	$N \times C \times W_{in}$	pre-upsampling values
<code>Z</code>	Z	matrix	$N \times C \times W_{out}$	post-upsampling values
<code>dLdZ</code>	$\partial L / \partial Z$	matrix	$N \times C \times W_{out}$	gradient of Loss wrt Z
<code>dLdA</code>	$\partial L / \partial A$	matrix	$N \times C \times W_{in}$	gradient of Loss wrt A

5.1.1 Upsample1d Forward

Upsampling as the name suggests is used to increase the size of input. In MyTorch implementation, we do this by a simple operation of adding $(k - 1)$ intermediate 0s. The Figure 5 shows how Upsampling is done.

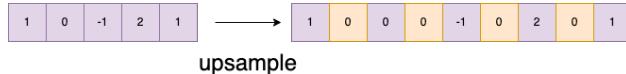


Figure 5: Upsampling1d Forward Example ($k=2$)

The output will be of width

$$W_{out} = k \times (W_{in} - 1) + 1$$

where k is the *upsampling_factor*. For this example, $k = 2$, so only 1 zero is inserted between the elements.

5.1.2 Upsample1d Backward

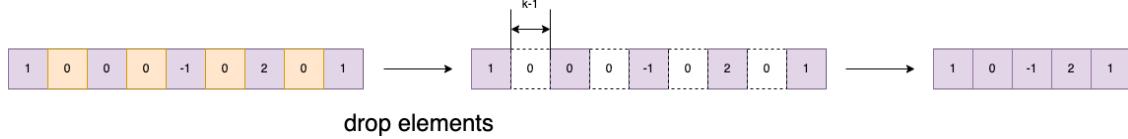


Figure 6: Upsampling1d Backward Example ($k=2$)

Refer to Figure 6, we do the reverse operation of forward. Only the $dLdZ$ entries corresponding to the original input data should be kept, and the $dLdZ$ entries corresponding to the padded 0s should have no effect on $dLdA$ and hence should be dropped.

5.2 Downsample1d

Downsampling1d is just “dropping” off elements with a factor of k .

In this section, your task is to implement `Downsample1d` class. Please consider the following class structure.

```
class Downsample1d:

    def __init__(self, stride):
        self.downsampling_factor = downsampling_factor

    def forward(self, A):
        Z = # TODO

        return Z

    def backward(self, dLdZ):
        dLdA = # TODO

        return dLdA
```

Table 2: Downsample1d Class Components

Code Name	Math	Type	Shape	Meaning
downsampling_factor	k	scalar	-	downsampling factor $\in \mathbb{Z}^+$
A	A	matrix	$N \times C \times W_{in}$	pre-downsampling features
Z	Z	matrix	$N \times C \times W_{out}$	post-downsampling features
dLdZ	$\partial L / \partial Z$	matrix	$N \times C \times W_{out}$	gradient of Loss wrt Z
dLdA	$\partial L / \partial A$	matrix	$N \times C \times W_{in}$	gradient of Loss wrt A

5.2.1 Downsample1d Forward

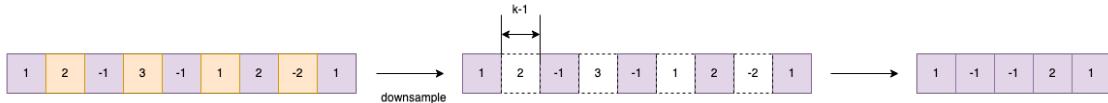


Figure 7: Downsampling1d Forward Example ($k=2$)

As you may have figured it out, implementation wise, Upsampling and Downsampling are inverse operations of one another. Upsampling forward is downsampling backward and vice-versa.

The output width will be:

$$W_{out} = W_{in} // k + 1$$

5.2.2 Downsample1d Backward

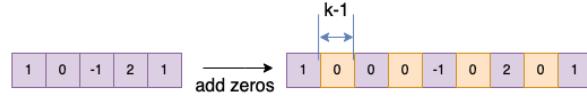


Figure 8: Downsampling1d Backward Example ($k=2$)

In `backward` function, one thing to note is that `dLdZ`, the gradient of input should be the **same size as the input**. We pad 0s in between because the gradient of Loss wrt the elements dropped in downsampling forward should be 0.

Hint: you might want to store W_{in} in `forward`. Think about what happens when the input size is even/odd.

5.3 Upsampling2d

2D upsampling is used for inputs like images where upsampling is performed in both the x and y direction. Please consider the following class structure.

```
class Upsample2d:

    def __init__(self, upsampling_factor):
        self.upsampling_factor = upsampling_factor

    def forward(self, A):
        Z = # TODO

        return Z

    def backward(self, dLdZ):
        dLdA = # TODO

        return dLdA
```

Table 3: Upsample2d Class Components

Code Name	Math	Type	Shape	Meaning
upsampling_factor	k	scalar	-	upsampling factor $\in \mathbb{Z}^+$
A	A	matrix	$N \times C \times H_{in} \times W_{in}$	pre-upsampling features
Z	Z	matrix	$N \times C \times H_{out} \times W_{out}$	post-upsampling features
dLdZ	$\partial L / \partial Z$	matrix	$N \times C \times H_{out} \times W_{out}$	gradient of Loss wrt Z
dLdA	$\partial L / \partial A$	matrix	$N \times C \times H_{in} \times W_{in}$	gradient of Loss wrt A

5.3.1 Upsampling2d Forward

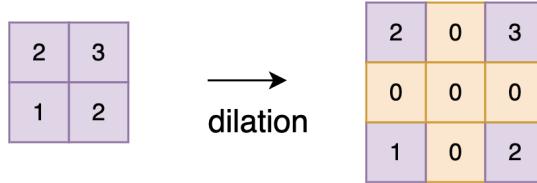


Figure 9: Upsampling 2d ($k=2$)

Upsampling an image is a simple operation where $(k - 1)$ zeros are added in between pixels of the input map. (Some may know this as dilation). The above diagram gives an intuitive explanation.

- Can you deduce H_{out} based on H_{in} and k ?
- Can you deduce W_{out} based on W_{in} and k ?
- Hint: same logic as `Upsample1d`

5.3.2 UpSample2d Backward

The exact opposite takes place in backward where intermediate $k - 1$ elements are dropped.

5.4 Downsample2d

In downsampling, the input features are reduced by a factor of k in both x and y dimension.

In this section, you will implement the `Downsample2d` class. Please consider the following class structure.

```
class Downsample2d:

    def __init__(self, downsampling_factor):
        self.downsampling_factor = downsampling_factor

    def forward(self, A):
        Z = # TODO

        return Z

    def backward(self, dLdZ):
        dLdA = # TODO

        return dLdA
```

Table 4: Downsample2d Class Components

Code Name	Math	Type	Shape	Meaning
downsampling_factor	k	scalar	-	downsampling factor $\in \mathbb{Z}^+$
A	A	matrix	$N \times C \times H_{in} \times W_{in}$	pre-downsampling features
Z	Z	matrix	$N \times C \times H_{out} \times W_{out}$	post-downsampling features
dLdZ	$\partial L / \partial Z$	matrix	$N \times C \times H_{out} \times W_{out}$	gradient of Loss wrt Z
dLdA	$\partial L / \partial A$	matrix	$N \times C \times H_{in} \times W_{in}$	gradient of Loss wrt A

5.4.1 Downsample2d Forward

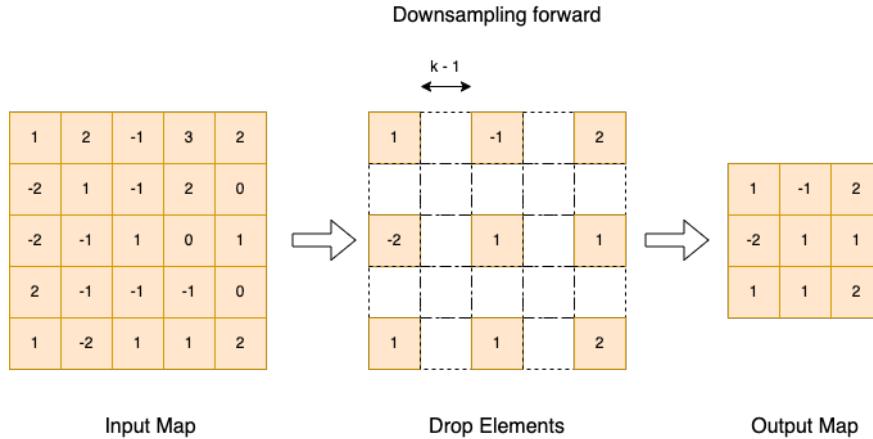


Figure 10: Downsample 2d Forward Example ($k=2$)

- Can you deduce H_{out} based on H_{in} and k ?
- Can you deduce W_{out} based on W_{in} and k ?
- Hint: same logic as `Downsample1d`

5.4.2 Downsample2d Backward

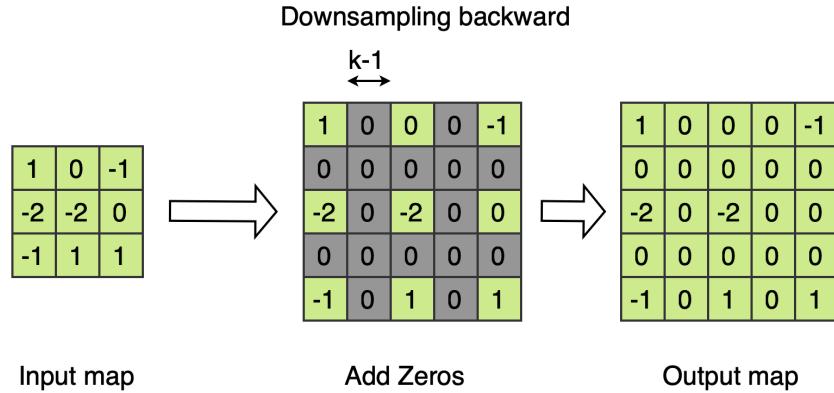


Figure 11: Downsample 2d Backward Example ($k=2$)

In `backward` function, we pad 0s in between because the gradient of Loss wrt the elements dropped in downsampling forward should be 0. Since the values of the dropped elements don't affect the downsampling output, they have zero gradients with respect to Loss.

6 Convolutional layer

Congrats on completing the Resampling layers. Now come the interesting portions of this homework. We will look into 1d and 2d convolutions and how to implement them from scratch.

But first of all, what is a convolution anyway and why does it help us capture the patterns?

Assume that we have a pattern which is 3 by 3 pixels that looks like this:

$$P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

The above matrix represents a pattern that looks like a cross. The small regression model that will detect such patterns would need to learn a 3 by 3 parameter matrix F where parameters at positions corresponding to the 1s in the input patch would be positive numbers, while the parameters in positions corresponding to 0s would be close to zero. If we take the elements of F as the weight parameters for corresponding elements of P , the weighted sum we obtain is higher the more similar F is to P . For instance, assume that F looks like this:

$$F = \begin{bmatrix} 0 & 2 & 1 \\ 2 & 4 & 3 \\ 0 & 3 & 0 \end{bmatrix}$$

Convolving P_1 and F , we will get

$$P_1 \circledast F = 0 \cdot 0 + 1 \cdot 2 + 0 \cdot 1 + 1 \cdot 2 + 1 \cdot 4 + 1 \cdot 3 + 0 \cdot 0 + 1 \cdot 3 + 0 \cdot 0 = 14$$

If our input patch had a different pattern, for example, that of a letter T:

$$P_2 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

The convolution result will be lower:

$$P_2 \circledast F = 10$$

So, you can see the more the patch “looks” like the filter, the higher the value of the convolution operation is.

Important note: For this section, please refer to Appendix 11.2 to learn about how to use tensordot in order to avoid multiple nested loops. It takes some time to learn, but it will save you a lot of time later.

6.1 Conv1d_stride1 [10 Points]

Convolution 1d involves convolving the input with a kernel in just 1 direction. We will first implement `Conv1d_stride1`⁷, and use the resampling functions which we just built to generalize it to work for `stride != 1`.

In this section, you will implement the `Conv1d_stride1` class. Please consider the following class structure.

```
class Conv1d_stride1:

    def __init__(self, in_channels, out_channels, kernel_size, weight_init_fn, bias_init_fn):

        self.in_channels = in_channels
        self.out_channels = out_channels
        self.kernel_size = kernel_size

        self.W = weight_init_fn(out_channels, in_channels, kernel_size)
        self.b = bias_init_fn(out_channels)

        self.dLdW = np.zeros(self.W.shape)
        self.dLdb = np.zeros(self.b.shape)

    def forward(self, A):

        self.A = A
        Z = # TODO

        return Z

    def backward(self, dLdZ):
        self.dLdW = #TODO
        self.dLdb = #TODO

        dLdA = # TODO

        return dLdA
```

Table 5: Conv1d_stride1 Class Components

Code Name	Math	Type	Shape	Meaning
<code>kernel_size</code>	K	scalar	-	kernel size
A	A	matrix	$N \times C_{in} \times W_{in}$	data input for convolution features after conv1d with stride 1
Z	Z	matrix	$N \times C_{out} \times W_{out}$	
W	W	matrix	$C_{out} \times C_{in} \times K$	weight parameters, also called kernels in CNN
b	b	vector	$C_{out} \times 1$	bias parameters
dLdZ	$\partial L / \partial Z$	matrix	$N \times C_{out} \times W_{out}$	how changes in outputs affect loss
dLdA	$\partial L / \partial A$	matrix	$N \times C_{in} \times W_{in}$	how changes in inputs affect loss
dLdW	$\partial L / \partial W$	matrix	$C_{out} \times C_{in} \times K$	how changes in weights affect loss
dLdb	$\partial L / \partial b$	vector	$C_{out} \times 1$	how changes in bias affect loss

⁷Stride is the number of pixels by which the kernel moves at each step.

6.1.1 Conv1d_stride1 Forward

When $C_{in} = 1$:

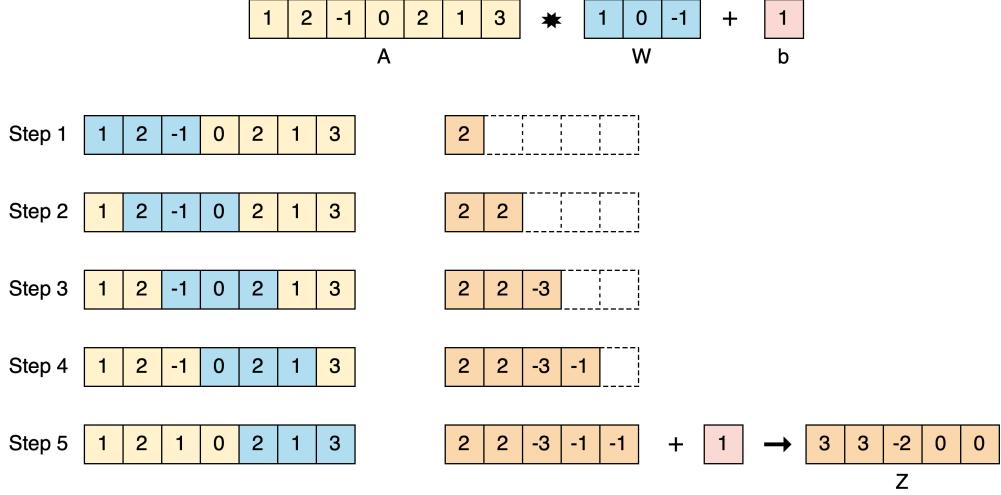


Figure 12: Conv1d with single channel input

An example illustrating the forward function is presented in Figure 12. A single channel input with 7 features is convolved with a single channel filter with kernel_size = 3. Convolution is basically an element wise multiplication and summation. As shown, when the filter scans through the input, at each step, there is an element wise multiplication between the patch of input elements and the filter elements. The output for a single convolutional step is a single scalar (orange). In each step, the filter shifts to the right by stride = 1. There is also a bias which is added to the output (broadcast addition). It is a single scalar per output channel that is added to all the elements of that channel.

Based on the example, can you deduce the value of W_{out} based on W_{in} and kernel_size?

When $C_{in} > 1$:

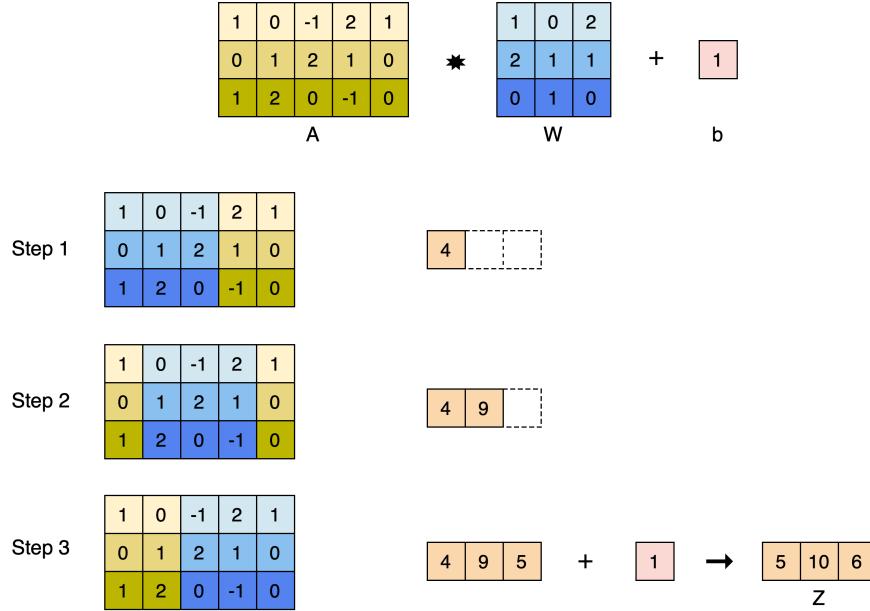


Figure 13: Conv1d with multichannel input

Figure 13 explains multi-channel convolutions. The input has 3 channels with 5 features and the kernel has 3 channels (same as input) with $\text{kernel_size} = 3$. Similar to single channel, the filter convolves the input and performs an element-wise multiplication and addition. It should be noted that in the multi-channel case, **output of element wise multiplication and addition from all the 3 channels are added to produce a single scalar for a convolution step.**

It can be observed that convolution of a single filter produces a single channel output. The weight W has size $C_{out} \times C_{in} \times K$, meaning that we have C_{out} filters of size $C_{in} \times K$, each producing 1 different outputs with width W_{out} , hence producing an output of shape $C_{out} \times W_{out}$. We recommend you to understand the process and try the convolutions by hand before proceeding to code.

6.1.2 Conv_stride1 Backward

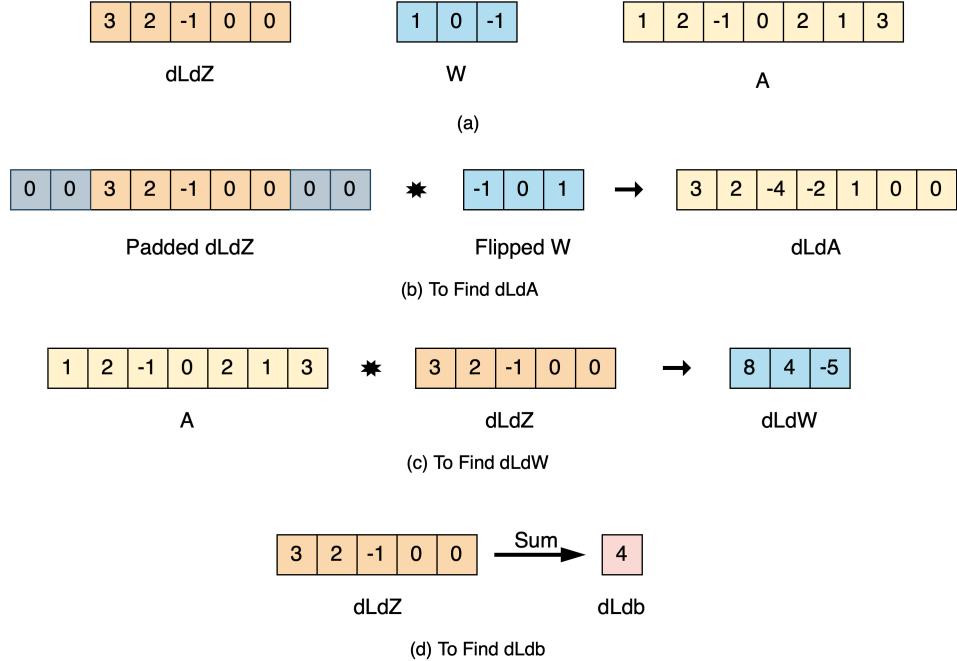


Figure 14: Conv1d backward with single channel input

Convolution backward is almost exactly the same operation done in the reverse order. Consider the figure shown in 14 for single channel backward and 15 for multi-channel backward. For the backward operation, we use the gradient of loss wrt to the output of convolution $dLdZ$. With this, we need to find the the gradient of the loss wrt to the kernels $dLdW$ and gradient of loss wrt to the bias $dLdb$.

$\begin{matrix} 2 & 1 & 1 \end{matrix}$ → sum → 4
 $dLdZ$ $dLdb$

$\begin{matrix} 2 & 1 & 1 \end{matrix}$ → broadcast → $\begin{matrix} 2 & 1 & 1 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{matrix}$
 $dLdZ$ $dLdZ$

$\begin{matrix} 1 & 0 & -1 & 2 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 1 & 2 & 0 & -1 & 0 \end{matrix}$ * $\begin{matrix} 2 & 1 & 1 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{matrix}$ = $\begin{matrix} 1 & 1 & 1 \\ 3 & 5 & 5 \\ 4 & 3 & -1 \end{matrix}$
 A $dLdZ$ $dLdW$

$\xrightarrow{\text{padding}}$ $\begin{matrix} 0 & 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 & 0 \end{matrix}$ $\begin{matrix} 1 & 0 & 2 \\ 2 & 1 & 1 \\ 0 & 1 & 0 \end{matrix}$ $\begin{matrix} 2 & 0 & 1 \\ 1 & 1 & 2 \\ 0 & 1 & 0 \end{matrix}$
 $\xrightarrow{\text{horizontal flip}}$
 $\begin{matrix} 0 & 0 & 2 & 1 & 1 & 0 & 0 \end{matrix} * \begin{matrix} 2 & 0 & 1 \end{matrix} = \begin{matrix} 2 & 1 & 5 & 2 & 2 \end{matrix}$
 $\begin{matrix} 0 & 0 & 2 & 1 & 1 & 0 & 0 \end{matrix} * \begin{matrix} 1 & 1 & 2 \end{matrix} = \begin{matrix} 4 & 4 & 5 & 2 & 1 \end{matrix}$
 $\begin{matrix} 0 & 0 & 2 & 1 & 1 & 0 & 0 \end{matrix} * \begin{matrix} 0 & 1 & 0 \end{matrix} = \begin{matrix} 0 & 2 & 1 & 1 & 0 \end{matrix}$
 $dLdZ$ flipped weight $dLdA$

Figure 15: Conv1d_stride1 Multichannel Backward Example

Finding dLdb: As you know from the forward, bias is a single scalar for each output channel. Hence, we get $dLdb$ by just summing the elements of $dLdZ$ channel wise to produce a vector of shape equal to the number of output channels.

Finding dLdW: Do a row-wise convolution between $dLdZ$ and the input A to get $dLdW$. Since one output map is produced from one filter, we can obtain derivatives w.r.t. all filters by doing the same process with different output maps. In addition, we can get derivatives for multiple channels of the same filter by convoluting $dLdZ$ with the corresponding channel in input.

Finding dLdA:

- Broadcast⁸ $dLdZ C_{in}$ times as shown. This is done because, no matter how many channels the input has, a single filter produces only one output channel in forward. For single channel case 14, it is not required as $C_{in} = 1$.
- Pad this with `kernel_size - 1` zeros on both sides (This is done as we would require the output to

⁸You don't need to explicitly broadcast, `tensordot` would save you the trouble!

- be larger than that of the input as convolution reduces the size in forward)
- Flip each channel of the filter left to right
 - Convolve each flipped channel of the filter with the broadcasted and padded $dLdZ$ to get $dLdA$

6.2 Conv1d [5 points]

In this section, we would be implementing `Conv1d` which works for any stride > 1 . We will reuse the code from `conv1d_stride1` 6.1 implementation to make it work for any stride > 1 .

Specifically, a `conv1d_stride1` layer followed by a `downsample1d` layer with factor $k=2$ is equivalent to a `conv1d_stride2`. More generally, a `conv1d_stride1` layer followed by a `downsample1d` layer with factor k is equivalent to a `conv1d` with stride k .

That's it! Now we can implement a `conv1d` operation for any value of stride with using just a combination of `conv1d_stride1` and `downsample` layer with factor k (= stride).

Your task is to implement forward and backward of the `Conv1d` class. Please consider the following class structure.

```
class Conv1d:

    def __init__(self, in_channels, out_channels, kernel_size, stride,
                 weight_init_fn, bias_init_fn):

        self.stride = stride
        self.conv1d_stride1 = # TODO
        self.downsample1d = # TODO

    def forward(self, A):

        Z = # TODO (<= 2 lines of code)
        # Line 1: Conv1d forward
        # Line 2: Downsample1d forward

        return Z

    def backward(self, dLdZ):

        dLdA = # TODO (<= 2 lines of code)
        # Line 1: Downsample1d backward
        # Line 2: Conv1d backward

        return dLdA
```

Table 6: Conv1d Class Components

Code Name	Math	Type	Shape	Meaning
<code>stride</code>	$stride$	scalar	-	equivalent to downsampling factor
<code>A</code>	A	matrix	$N \times C_{in} \times W_{in}$	data input for convolution
<code>Z</code>	Z	matrix	$N \times C_{out} \times W_{out}$	features after conv1d with stride
<code>dLdZ</code>	$\partial L / \partial Z$	matrix	$N \times C_{out} \times W_{out}$	how changes in outputs affect loss
<code>dLdA</code>	$\partial L / \partial A$	matrix	$N \times C_{in} \times W_{in}$	how changes in inputs affect loss

6.3 Conv2d_stride1 [10 Points]

2D convolution layers play a major role in today's image based intelligence tasks. From the inception of CNNs in the 80s by Yann LeCun, this model has undergone various transformations in a wide range of applications such as Image Classification, Image Segmentation, Object Detection and so on.

In this section, you will implement the `Conv2d_stride1` class. Similar to `Conv1d_stride1`, this class will be further generalized to work for `stride != 1` in the following sections.

Please consider the following class structure.

```
class Conv2d_stride1:

    def __init__(self, in_channels, out_channels, kernel_size, weight_init_fn, bias_init_fn):

        self.in_channels = in_channels
        self.out_channels = out_channels
        self.kernel_size = kernel_size

        self.W = weight_init_fn(out_channels, in_channels, kernel_size)
        self.b = bias_init_fn(out_channels)

        self.dLdW = np.zeros(self.W.shape)
        self.dLdb = np.zeros(self.b.shape)

    def forward(self, A):

        self.A = A
        Z = # TODO

        return Z

    def backward(self, dLdZ):
        self.dLdW = #TODO
        self.dLdb = #TODO

        dLdA = # TODO

        return dLdA
```

Table 7: Conv2d_stride1 Class Components

Code Name	Math	Type	Shape	Meaning
<code>kernel_size</code>	K	scalar	$N \times C_{in} \times H_{in} \times W_{in}$	kernel size
<code>A</code>	A	matrix	$N \times C_{in} \times H_{in} \times W_{in}$	data input for convolution
<code>Z</code>	Z	matrix	$N \times C_{out} \times H_{out} \times W_{out}$	features after conv2d with stride 1
<code>W</code>	W	matrix	$C_{out} \times C_{in} \times K \times K$	weight parameters, i.e. kernels
<code>b</code>	b	vector	$C_{out} \times 1$	bias parameters
<code>dLdZ</code>	$\partial L / \partial Z$	matrix	$N \times C_{out} \times H_{out} \times W_{out}$	how changes in outputs affect loss
<code>dLdA</code>	$\partial L / \partial A$	matrix	$N \times C_{in} \times H_{in} \times W_{in}$	how changes in inputs affect loss
<code>dLdW</code>	$\partial L / \partial W$	matrix	$C_{out} \times C_{in} \times K \times K$	how changes in weights affect loss
<code>dLdb</code>	$\partial L / \partial b$	vector	$C_{out} \times 1$	how changes in bias affect loss

Pay special attention to the order of the dimension of A and Z : batch size x channels x height x weight.

6.3.1 Conv2d_stride1 Forward

When $C_{in} = 1$

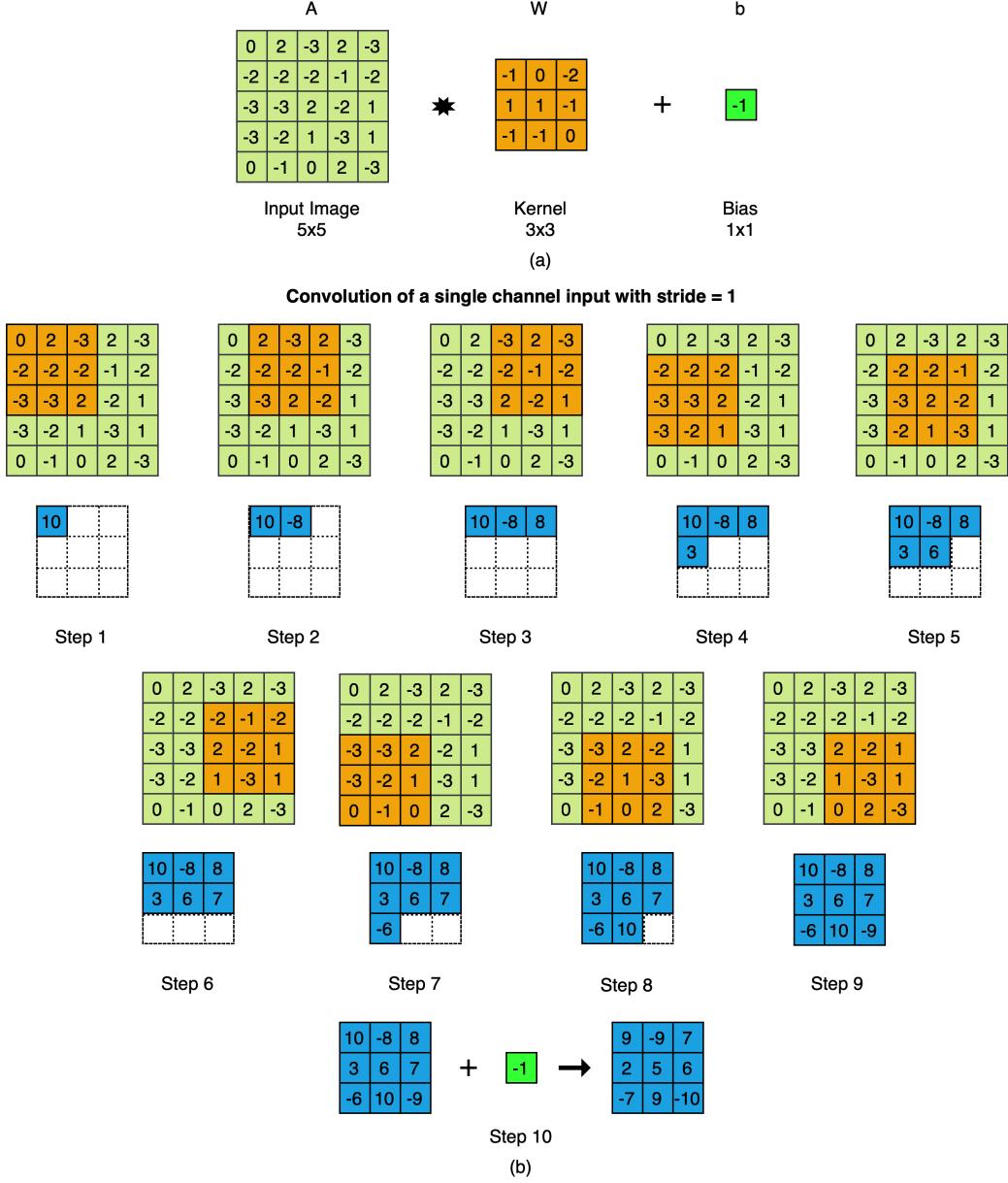


Figure 16: 2d Convolution Example stride=1 $C_{in} = 1$

Let us understand 2d convolutions with an example. Figure 16 shows the convolution of a 3x3 kernel on a 5x5 input image. This example is for a single channel input ($C_{in} = 1$).

Each step of convolution is depicted in the figure. Similar to the case of 1d, in a 2d convolution, an element wise multiplication of the filter with the image patch is done and then a sum is taken to produce a single output element (Image patch here means the portion of the image below the kernel). In step one, the filter starts from the top left corner of the image. After performing an element wise multiplication and a summation, we get the result to be 3 as shown in the output (blue colored). In step 2, the kernel moves

towards the right by a distance of 1 pixel. Therefore, the stride of this convolution is 1. As the filter scans through the image, the output map gets formed. It is intuitively observed that the position of the output element is influenced by where the kernel is placed on the image. The final step is to add a bias per channel for the output. We recommend to perform the convolution operation by hand to get a good grasp on the concept.

When $C_{in} = 2$

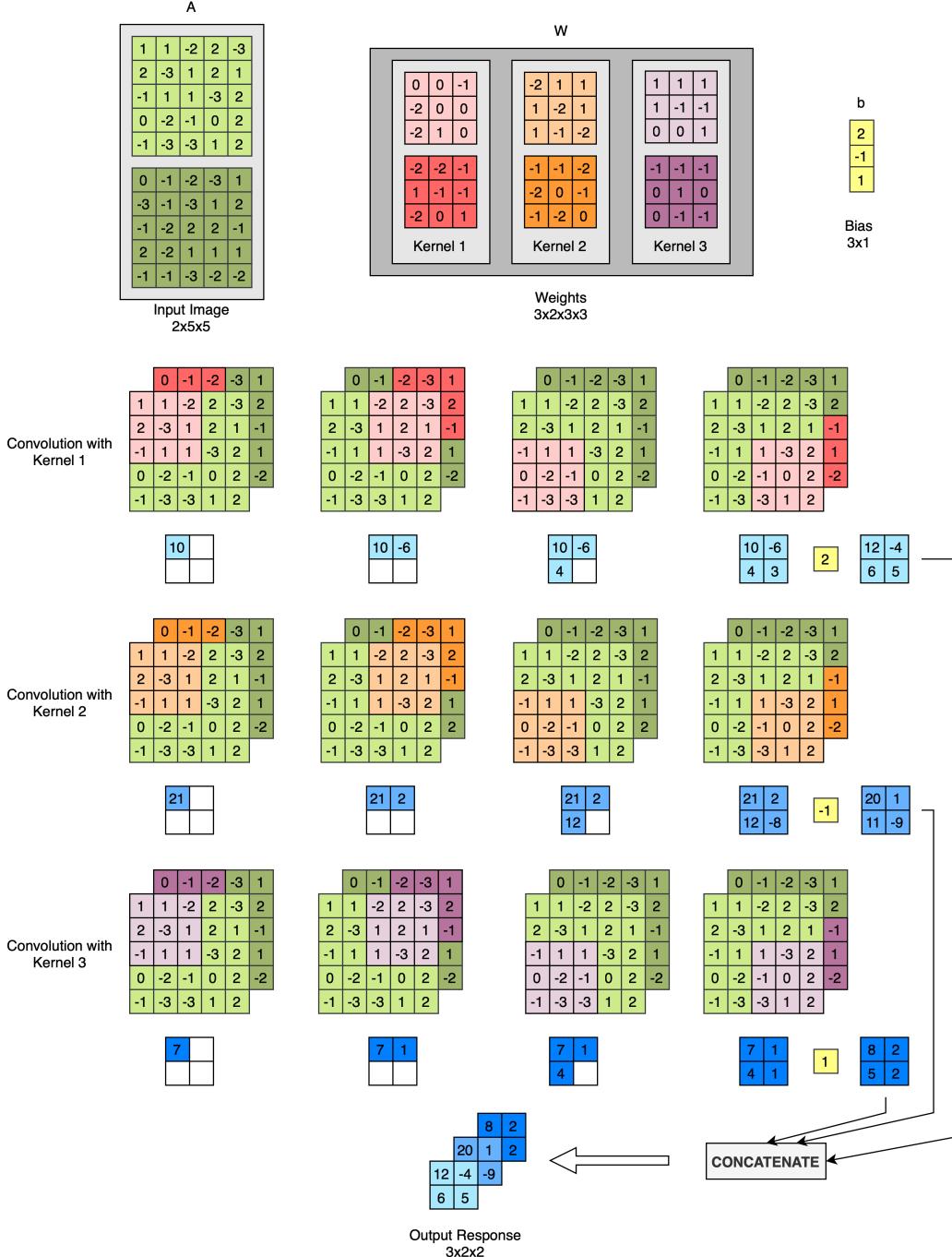


Figure 17: Multi channel convolution example $C_{in} = 2$

We can extend the same idea stated above for a single channel image to a multi-channel image. Since a single channel filter convolves with a single channel image as explained previously, we would require a multi channel kernel to convolve a multi channel image. Therefore, the number of channels of the convolving kernel should be **equal** to the number of channels of the input image.

As shown in figure 17, A has 2 channels and so do all the kernels in W . Consider convolution with Kernel 1. Starting from the top left, for channel 1 and channel 2 we have,

$$(1 \times 0) + (1 \times 0) + (-2 \times -1) + (2 \times -2) + (-3 \times 0) + (1 \times 0) + (-1 \times -2) + (1 \times 1) + (1 \times 0) = 1$$

$$(0 \times -2) + (-1 \times -2) + (-2 \times -1) + (-3 \times 1) + (-1 \times -1) + (-3 \times -1) + (-1 \times -2) + (-2 \times 0) + (2 \times 1) = 9$$

respectively. This sums to 10.

Applying the same idea from the previous paragraphs, 2 input patches with 2 kernel channels would produce 2 output scalars. Then the outputs are summed to get a final single output scalar. The filter then takes 2 pixel steps (in this example stride = 2) and computes the same. After the 4th step, we get a single channel output. **The take away is that, convoluting a filter with one input, produces a single channel output irrespective of the number of input/kernel channels.** When we use a different filter, we get a different output channel. That's what happens in convolution with Kernel 2 and Kernel 3. From 3 filters convolving with the input image, we get 3 output channels as shown. These channels are concatenated to produce a single 3 channel output image. The bias for this 3 channel output will be a 3×1 vector with each element as a bias for each channel.

The summary is that:

- no. input channels = no. kernel channels
- no. output channels = no. kernels

6.3.2 Conv2d_stride1 Backward

Now we will see how backward in convolution is performed. Backpropagation in 2d Convolution is not as hard at it sounds. Turns out, it employs that same process of convolution. Given an output map Z , in backprop, we get the gradient of the output map Z wrt the loss L $dLdZ$ from the previous layers. This gradient map will be of the same shape as Z . With $dLdZ$, we need to find the gradient of the Loss wrt to weights ($dLdW$), bias ($dLdb$) and input ($dLdA$).

To calculate self.dLdA:

- Pad the $dLdZ$ map with $K - 1$ zeros as shown in Figure 18. We do this because, in forward, convolution reduces the size
- Flip the filter top to bottom and left to right.
- Convolve the flipped kernel over the padded ($dLdZ$) to get ($dLdA$) as the convolution output.

The above process gives us self.dLdA for one input channel. To get the same for other input channels, we flip the filter channel corresponding to the input channel and convolve with the padded ($dLdZ$).

To calculate self.dLdW: Simply convolve the $dLdZ$ with the input map A to get $dLdW$. Since one output map is produced from one filter, we can obtain gradients w.r.t. all filters by doing the same process with different output maps. In addition, we can get gradients for multiple channels of the same filter by convoluting ($dLdZ$) with the corresponding channel in input.

To calculate self.dLdb: As you know from the forward, bias is a single scalar for each output channel. Hence, we get $dLdb$ by just summing the elements of $dLdZ$ channelwise to produce a vector of shape equal to the number of output channels.

self.dLdW and self.dLdb represent the unaveraged gradients of the loss w.r.t self.W and self.b . Their shapes are the same as the weight self.W and the bias self.b .

$$\begin{array}{|c|c|c|} \hline 1 & 0 & 1 \\ \hline 1 & 2 & 3 \\ \hline 0 & 1 & 2 \\ \hline \end{array} \xrightarrow{\text{sum}} 11 \\
 \text{dLdZ} \qquad \qquad \qquad \text{dLdB}$$

$$\begin{array}{|c|c|c|c|c|} \hline 0 & 2 & -3 & 2 & -3 \\ \hline -2 & -2 & -2 & -1 & -2 \\ \hline -3 & -3 & 2 & -2 & 1 \\ \hline -3 & -2 & 1 & -3 & 1 \\ \hline 0 & -1 & 0 & 2 & -3 \\ \hline \end{array} * \begin{array}{|c|c|c|} \hline 1 & 0 & 1 \\ \hline 1 & 2 & 3 \\ \hline 0 & 1 & 2 \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline -14 & -7 & -16 \\ \hline -7 & -13 & -4 \\ \hline -6 & -10 & -3 \\ \hline \end{array} \\
 \text{A} \qquad \qquad \qquad \text{dLdZ} \qquad \qquad \qquad \text{dLdW}$$

$$\begin{array}{|c|c|c|} \hline 1 & 0 & 1 \\ \hline 1 & 2 & 3 \\ \hline 0 & 1 & 2 \\ \hline \end{array} \xrightarrow{\text{padding}} \text{padding} \qquad \qquad \qquad \text{kernel_size - 1} \\
 \begin{array}{|c|c|c|c|c|c|c|c|} \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 2 & 3 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|} \hline -1 & 0 & -2 \\ \hline 1 & 1 & -1 \\ \hline -1 & -1 & 0 \\ \hline \end{array} \xrightarrow{\text{weight}} \begin{array}{|c|c|c|} \hline -2 & 0 & -1 \\ \hline -1 & 1 & 1 \\ \hline 0 & -1 & -1 \\ \hline \end{array} \xrightarrow{\text{horizontal flip}} \begin{array}{|c|c|c|} \hline 0 & -1 & -1 \\ \hline -1 & 1 & 1 \\ \hline -2 & 0 & -1 \\ \hline \end{array} \xrightarrow{\text{vertical flip}} \begin{array}{|c|c|c|} \hline 0 & -1 & -1 \\ \hline -1 & 1 & 1 \\ \hline -2 & 0 & -1 \\ \hline \end{array} \\
 \text{weight} \qquad \qquad \qquad \text{horizontal flip} \qquad \qquad \qquad \text{vertical flip}$$

$$\begin{array}{|c|c|c|c|c|c|} \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 2 & 3 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array} * \begin{array}{|c|c|c|} \hline 0 & -1 & -1 \\ \hline -1 & 1 & 1 \\ \hline -2 & 0 & -1 \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline -1 & 0 & -3 & 0 & -2 \\ \hline 0 & -1 & -5 & -3 & -7 \\ \hline 0 & 1 & 1 & -2 & -7 \\ \hline -1 & -2 & -2 & -2 & -2 \\ \hline 0 & -1 & -3 & -2 & 0 \\ \hline \end{array} \\
 \text{dLdZ} \qquad \qquad \qquad \text{flipped weight} \qquad \qquad \qquad \text{dLdA}$$

Figure 18: Conv2d_stride1 Backward: Convolution

6.3.3 Conv2d [5 points]

In this section, we would be implementing `Conv2d` which works for any stride > 1. We will reuse the code from `conv1d_stride1` 6.3 implementation to make it work for any stride > 1.

Specifically, a `conv2d_stride1` layer followed by a `downsample2d` layer with factor $k=2$ is equivalent to a `conv2d_stride2`. More generally, a `conv2d_stride1` layer followed by a `downsample2d` layer with factor k is equivalent to a `conv1d_stride` with stride k . Figure 19 shows this.

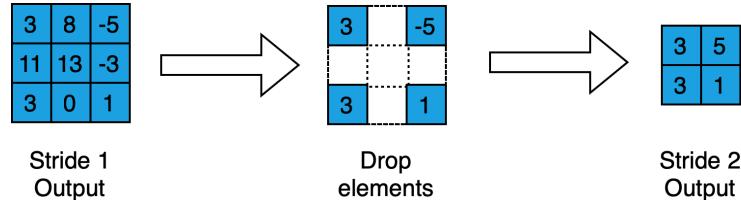


Figure 19: Stride 2 convolution as stride 1 convolution and downsampling with factor = 2

That's it! Now we can implement a conv2d operation for any value of stride with using just a combination of conv2d_stride1 and downsample layer with factor k (= stride).

For backward, the steps in forward are reversed in the same order. For stride > 1, downsampling backward is called as shown in Figure 20 and then convolution stride 1 backward is called as shown in Figure 21.

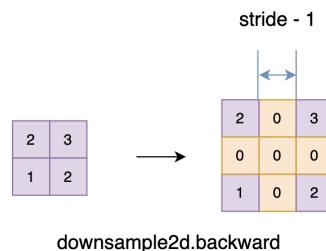
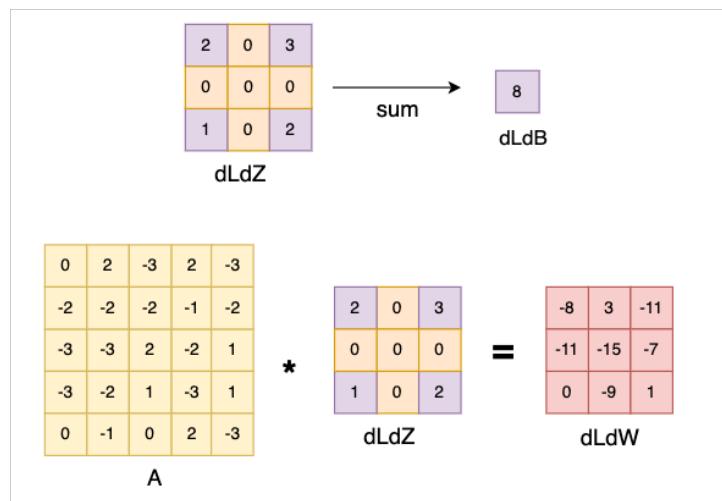


Figure 20: Downsample2d Backward: Conv2d Example



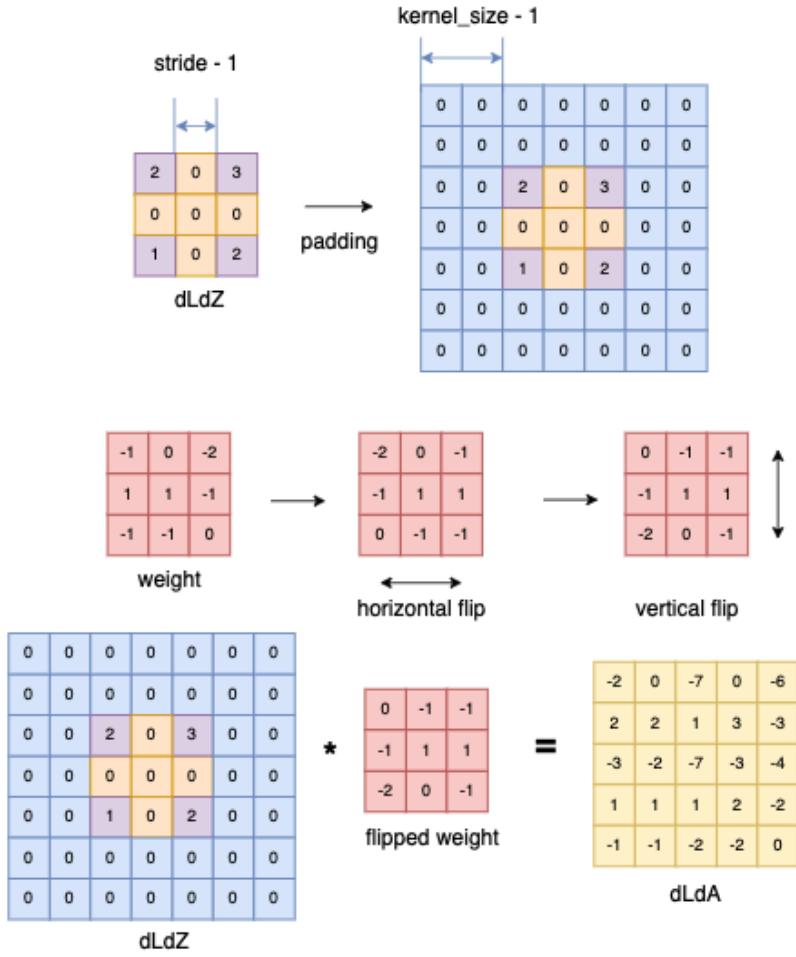


Figure 21: Conv2d Backward Example (stride > 1)

Make sure that you have understood that strided convolutions are just a combination of convolution with stride 1 and downsampling. In this section, your task is to implement the `forward` and `backward` attributes of the `Conv2d` class. Please consider the following class structure.

```
class Conv2d():
    def __init__(self, in_channels, out_channels, kernel_size, stride,
                 weight_init_fn, bias_init_fn):
        self.stride = stride
        self.conv2d_stride1 = None # TODO
        self.downsample2d = None # TODO

    def forward(self, A):
        # Call Conv2d_stride1 forward
        # TODO

        # Call downsample2d forward
        Z = # TODO

    return Z
```

```

def backward(self, dLdZ):
    # Call downsample2d backward
    # TODO

    # Call Conv2d_stride1 backward
    dLdA = # TODO

    return dLdA

```

Table 8: Conv2d Class Components

Code Name	Math	Type	Shape	Meaning
stride	$stride$	scalar	-	Stride = downsampling factor
A	A	matrix	$N \times C_{in} \times H_{in} \times W_{in}$	data input for convolution
Z	Z	matrix	$N \times C_{out} \times H_{out} \times W_{out}$	features after conv2d with stride
dLdZ	$\partial L / \partial Z$	matrix	$N \times C_{out} \times H_{out} \times W_{out}$	how changes in outputs affect loss
dLdA	$\partial L / \partial A$	matrix	$N \times C_{in} \times H_{in} \times W_{in}$	how changes in inputs affect loss

6.4 Transposed Convolution [10 points]

6.4.1 ConvTranspose1d

In regular convolutions, the affine value Z for a layer are obtained by a direct convolution on the input values A in the previous layer. This causes a reduction of the output size. However, in an Transposed Convolution layer, the input values A are upsampled and then convolved, to be “pushed” to the next layer Z . Thus the output map’s size is increased.

The primary operation of ConvTranspose1d is to upsample the input and then convolve (with stride 1) to get the output. In your implementation, you will pass `upsampling_factor k` to the ConvTranspose1d class, which is the same in definition of convolution with a fractional stride ($\frac{1}{k}$). Please consider the Figure 22 for better intuition.

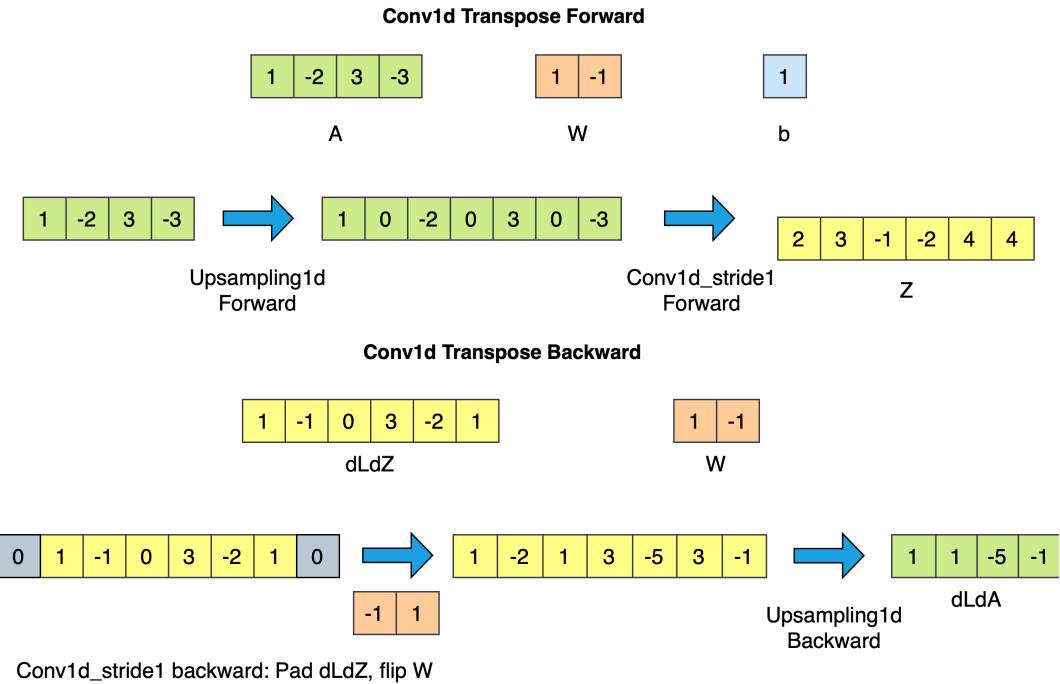


Figure 22: 1d Transpose Convolution

In this section, your task is to implement the `forward` and `backward` attributes of the `ConvTransposed1d` class. Please consider the following class structure.

```
class ConvTranspose1d():

    def __init__(self, in_channels, out_channels, kernel_size, upsampling_factor,
                 weight_init_fn, bias_init_fn):

        self.upsampling_factor = upsampling_factor
        self.upsample1d = # TODO
        self.conv1d_stride1 = # TODO

    def forward(self, A):

        Z = # TODO (<= 2 lines of code)
        # Line 1: Upsampling1d forward
        # Line 2: Conv1d_stride1 forward

        return Z

    def backward(self, dLdZ):

        dLdA = # TODO (<= 2 lines of code)
        # Line 1: Conv1d_stride1 backward
        # Line 2: Upsampling1d backward

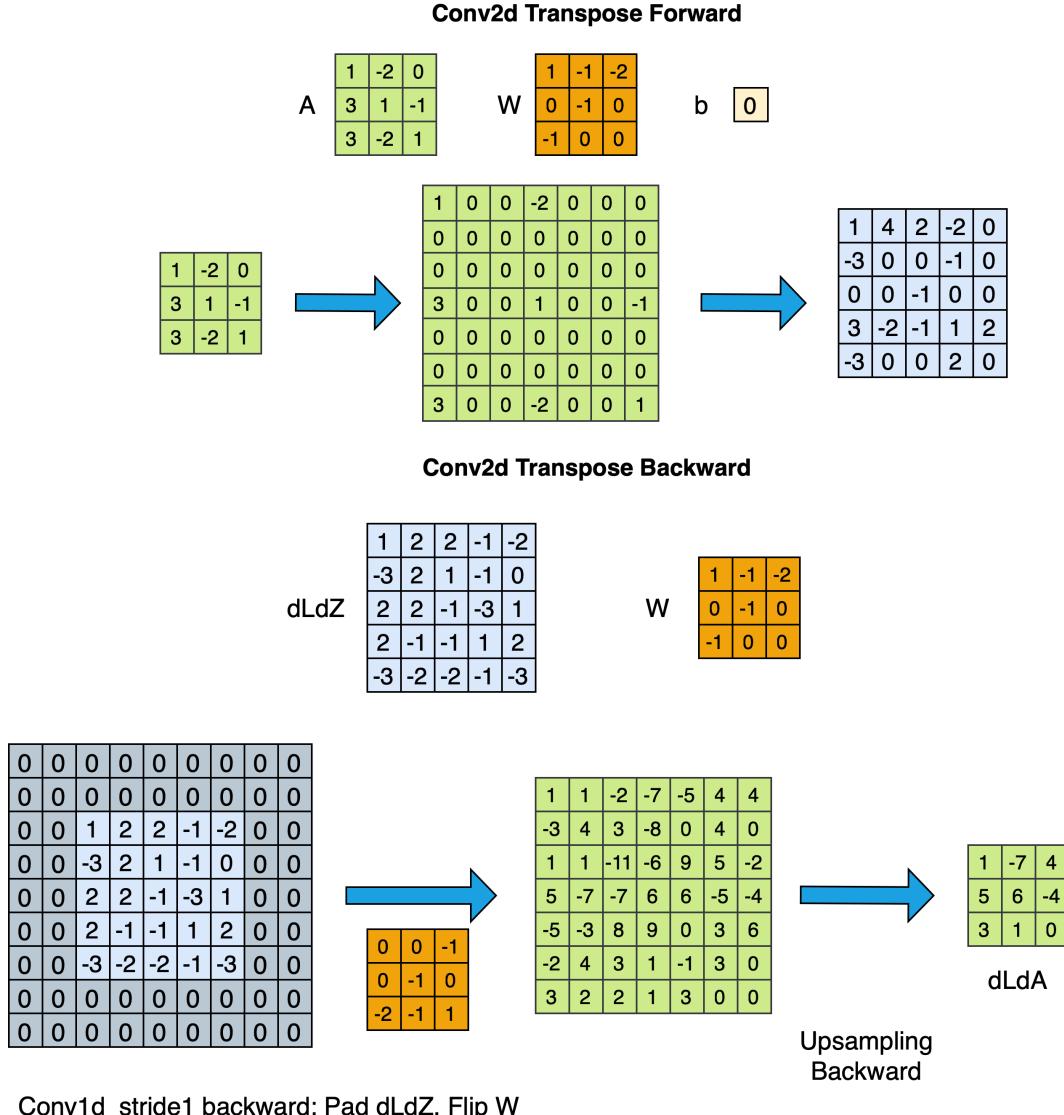
        return dLdA
```

Table 9: ConvTransposed1d Class Components

Code Name	Math	Type	Shape	Meaning
<code>upsampling_factor</code>	$upsampling_factor$	scalar	-	upsampling factor
A	A	matrix	$N \times C_{in} \times W_{in}$	data input for ConvTransposed1d features after ConvTransposed1d
Z	Z	matrix	$N \times C_{out} \times W_{out}$	
<code>dLdZ</code>	$\partial L / \partial Z$	matrix	$N \times C_{out} \times W_{out}$	how changes in outputs affect loss
<code>dLdA</code>	$\partial L / \partial A$	matrix	$N \times C_{in} \times W_{in}$	how changes in inputs affect loss

6.4.2 ConvTranspose2d

Similar to ConvTranspose1d, ConvTranspose2d is a combination of Upsample2d and convolution with stride 1. Please consider the Figure 23 for better intuition. The example uses as convolution with stride = 1/3. The downsampling factor k = 3 in the forward example. In the backward example, we have used an example from stride = 1/2 convolution where the downsampling factor = 2.



In your implementation, you are going to pass a *upsampling_factor* to the ConvTranspose2d class, which is the inverse of the stride(in this case, stride is fractional). Your task is to implement the **forward** and **backward** attributes of the ConvTranspose2d class. Please consider the following class structure.

In this section, your task is to implement the `forward` and `backward` attributes of the `ConvTranspose2d` class. Please consider the following class structure.

```
class ConvTranspose2d():

    def __init__(self, in_channels, out_channels, kernel_size, upsampling_factor,
                 weight_init_fn, bias_init_fn):

        self.upsampling_factor = upsampling_factor
        self.upsample2d = # TODO
        self.conv2d_stride1 = # TODO

    def forward(self, A):

        Z = # TODO (<= 2 lines of code)
        # Line 1: Upsampling2d forward
        # Line 2: Conv2d_stride1 forward

        return Z

    def backward(self, dLdZ):

        dLdA = # TODO (<= 2 lines of code)
        # Line 1: Conv2d_stride1 backward
        # Line 2: Upsampling2d backward

        return dLdA
```

Table 10: ConvTranspose2d Class Components

Code Name	Math	Type	Shape	Meaning
<code>upsampling_factor</code>	k	scalar	-	upsampling factor
<code>A</code>	A	matrix	$N \times C_{in} \times H_{in} \times W_{in}$	data input for ConvTranspose2d
<code>Z</code>	Z	matrix	$N \times C_{out} \times H_{out} \times W_{out}$	
<code>dLdZ</code>	$\partial L / \partial Z$	matrix	$N \times C_{out} \times H_{out} \times W_{out}$	how changes in outputs affect loss
<code>dLdA</code>	$\partial L / \partial A$	matrix	$N \times C_{in} \times H_{in} \times W_{in}$	how changes in inputs affect loss

7 Pooling

Different from convolution, the operation in pooling is fixed and no parameters to learn.

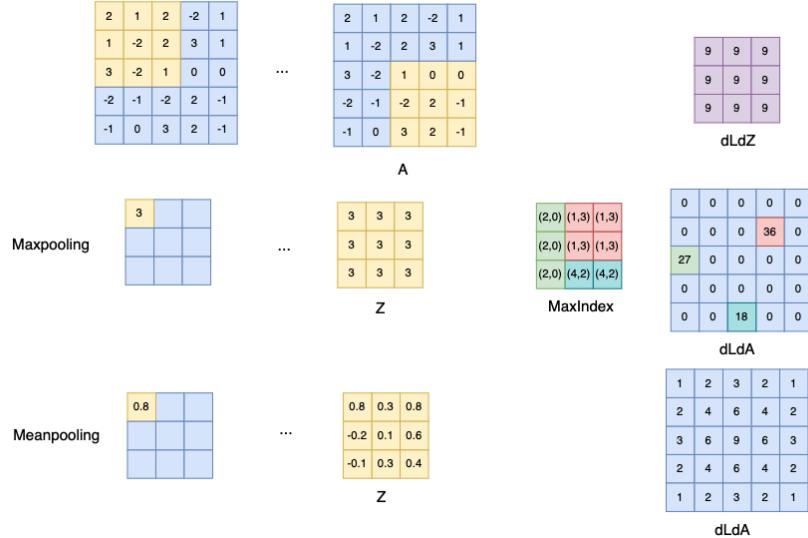


Figure 24: Pooling2d Example

Please complete the following classes in `mytorch/pool.py`

Max pooling selects the largest from a pool of elements and is performed by “scanning” the input. Implement both forward and backward methods for 1d and 2d max pooling. In case you find two max values in a kernel patch, you can use the first occurrence (consider row major ordering) of max value as the output for that kernel patch. This also can be seen from Figure 24.

Mean pooling takes the arithmetic mean of elements and is performed by “scanning” the input. Implement both forward and backward methods for 1d and 2d mean pooling.

7.1 MaxPool2d_stride1 and MeanPool2d_stride1 [20 Points]

Similar to the previous layers, you will implement a stride 1 pooling and then a combination of pooling and downsampling for higher strides. The pooling operation itself is just a jitter invariant operation. Pooling typically uses a $stride > 1$, which is the same as convolution followed by downsampling. You can take maxpooling as a normal convolution with standard filter and max activation. As to the meanpooling, it can be viewed as a convolution with a special filter, where each element in the filter is $\frac{1}{K}$ where K is `filter_size`. Consider the class structure given below.

```

class MaxPool2d_stride1():

    def __init__(self, kernel):
        self.kernel = kernel

    def forward(self, A):
        Z = # TODO
        return Z

    def backward(self, dLdZ):
        dLdA = # TODO
        return dLdA

class MeanPool2d_stride1():

    def __init__(self, kernel):
        self.kernel = kernel

    def forward(self, A):
        Z = # TODO
        return Z

    def backward(self, dLdZ):
        dLdA = # TODO
        return dLdA

```

7.2 MaxPool2d and MeanPool2d [10 Points]

You are expected to complete pooling operation in **MaxPool2d_stride1** and **MeanPool2d_stride1**. Use **MaxPool2d** and **MeanPool2d** as a wrapper class to implement downsampling after the max or mean operations.

```

class MaxPool2d():

    def __init__(self, kernel, stride):
        self.kernel = kernel
        self.stride = stride

        #Create an instance of MaxPool2d_stride1
        self.maxpool2d_stride1 =
            MaxPool2d_stride1(kernel)
        self.downsample2d = Downsample2d(stride)

    def forward(self, A):
        Z = # TODO
        return Z

    def backward(self, dLdZ):
        dLdA = # TODO
        return dLdA

class MeanPool2d():

    def __init__(self, kernel, stride):
        self.kernel = kernel
        self.stride = stride

    def forward(self, A):
        Z = # TODO
        return Z

    def backward(self, dLdZ):
        dLdA = # TODO
        return dLdA

```

Table 11: Pooling Class Components

Code Name	Math	Type	Shape	Meaning
stride	$stride$	scalar	-	stride
kernel_size	K	scalar	-	kernel_size
A	A	matrix	$N \times C_{in} \times H_{in} \times W_{in}$	data input for pooling
Z	Z	matrix	$N \times C_{in} \times H_{out} \times W_{out}$	features after pooling
dLdZ	$\partial L / \partial Z$	matrix	$N \times C_{in} \times H_{out} \times W_{out}$	how changes in outputs affect loss
dLdA	$\partial L / \partial A$	matrix	$N \times C_{in} \times H_{in} \times W_{in}$	how changes in inputs affect loss

8 Flatten layer

In `flatten.py`, complete `Flatten()`⁹.

This layer is often used between `Conv` and `Linear` layers, in order to squish the high-dim convolutional outputs into a lower-dim shape for the linear layer. In forward, a multi-dimensional input is reshaped into a single dimensional output. In backward, a single dimensional input is reshaped into a multi-dimensional output.



Figure 25: Flatten Forward Example

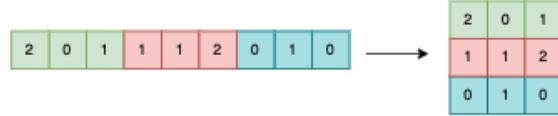


Figure 26: Flatten Backward Example

```
class Flatten():
    def forward(self, A):
        #TODO save shape of A
        Z = # TODO # reshape

    def backward(self, dLdZ):
        dLdA = # TODO # reshape
```

Table 12: Flatten Class Components

Code Name	Math	Type	Shape	Meaning
A	A	matrix	$N \times C_{in} \times W_{in}$	data input for Flatten
Z	Z	matrix	$N \times C_{in} \times W_{in}$	features after Flatten
dLdZ	$\partial L / \partial Z$	matrix	$N \times (C_{in} * W_{in})$	how changes in outputs affect loss
dLdA	$\partial L / \partial A$	matrix	$N \times C_{in} \times W_{in}$	how changes in inputs affect loss

Hint: This can be done in one line of code, with no new operations or (horrible, evil) broadcasting needed.

Bigger Hint: Flattening is a just subcase of reshaping.

⁹No mark is allocated in this section, but you will need to implement it correctly for the next part.

9 Converting Scanning MLPs to CNNs

9.1 CNN as a Simple Scanning MLP [5 Points]

In `models/mlp_scan.py` for `CNN_SimpleScanningMLP` compose a CNN that will perform the same computation as scanning a given input with a given **multi-layer perceptron**.

- You are given a 128×24 input (128 time steps, with a 24-dimensional vector at each time). You are required to **compute the result of scanning it with the given MLP**.
- The MLP evaluates **8 contiguous input vectors** at a time after flattening¹⁰ them (so it is effectively scanning for 8-time-instant wide patterns).
- The MLP “**strides**” **forward 4 time instants** after each evaluation, during its scan. It only scans until the end of the input (so it does not pad the end of the input with zeros).
- The MLP itself has three layers, with 8 neurons in the first layer (closest to the input), 16 in the second and 4 in the third. Each neuron uses a `ReLU()` activation except the final output neurons. All bias values are 0. Architecture is as follows:

```
[Flatten(), Linear(8 * 24, 8), ReLU(), Linear(8, 16), ReLU(), Linear(16, 4)]
```

- The **Multi-layer Perceptron is composed of three layers** and the architecture of the model is given in `models/mlp.py` included in the handout. You do not need to modify the code in this file, it is only for your reference.
- Since the network has **4 neurons in the final layer and scans with a stride of 4, it produces one 4-component output every 4 time instants**. Since there are 128 time instants in the inputs and no zero-padding is done, the network produces 31 outputs in all, one every 4 time instants. When **flattened**, this output will have **124 (4 × 31) values**.

For this problem you are required to implement the above scan, but you must do so using a Convolutional Neural Network. You must use the implementation of your Convolutional layers in the above sections to compose a Convolution Neural Network which will behave identically to scanning the input with the given MLP as explained above.

Your task is merely to understand how the architecture (and operation) of the given MLP translates to a CNN. You will have to determine how many layers your CNN must have, how many filters in each layer, the kernel size of the filters, and their strides and their activations. The final output (after flattening) must have 124 components.

Your tasks include:

- Designing the CNN architecture to correspond to a Scanning MLP
 - Create `Conv1d` objects defined as `self.conv1`, `self.conv2`, `self.conv3`, in the `init` method of `CNN_SimpleScanningMLP`.
 - For the `Conv1d` instances, you must specify the: `in_channels`, `out_channels`, `kernel_size`, and `stride`.
 - Add those layers along with the activation functions/ flatten layer to a class attribute you create called `self.layers`.
 - Initialize the weights for each convolutional layer, using the `init_weights` method. You must discover the orientation of the initial weight matrix(of the MLP) and convert it for the weights of the `Conv1d_stride1` layers.
 - This will involve (1) reshaping a transposed weight matrix into `out_channels`, `kernel_size`, `in_channels` and (2) transposing the weights back into the correct shape for the weights of a

¹⁰Remember that MLP only has linear layers, which are 1D.

Conv1d_stride1 instance, `out_channels`, `in_channels`, `kernel_size`.

- Use pdb to help you debug, by printing out what the initial input to this method is. Each index into the given weight's list corresponds to the Conv1d_stride1 layer. i.e. `weights[0]` are the weights for the first Conv1d_stride1 instance.

```

class CNN_SimpleScanningMLP():
    def __init__(self):
        self.conv1 = # TODO
        self.conv2 = # TODO
        self.conv3 = # TODO
        self.layers = []

    def init_weights(self, weights):
        w1,w2,w3 = weights
        self.conv1.conv1d_stride1.W = # TODO
        self.conv2.conv1d_stride1.W = # TODO
        self.conv3.conv1d_stride1.W = # TODO

    def forward(self, A):
        Z = A
        for layer in self.layers:
            Z = layer(Z)
        return Z

    def backward(self, dLdZ):
        dLdA = dLdZ
        for layer in self.layers[::-1]:
            dLdA = layer.backward(dLdA)
        return dLdA

```

The paths have been appended such that you can create layers with the calls to the class themselves, i.e. to make a ReLU layer, just use `ReLU()`. You have a weights file which will be used to autograde your network locally. We will not give you tables of variables in this section. Please figure out the shape and complete this section.

For more of a theoretical understanding of Simple Scanning MLPs, please refer to the Appendix section.

9.2 CNN as a Distributed Scanning MLP [5 Points]

Complete 9.2 in `models/mlp_scan.py` in the class `CNN_DistributedScanningMLP`. This section of the homework is very similar to 9.1, except that the MLP provided to you is a shared-parameter network that captures a distributed representation of the input.

You must compose a CNN that will perform the same computation as scanning a given input with a MLP.

Architecture details:

- The network has **8 first-layer neurons, 16 second-layer neurons and 4 third-layer neurons**. However, many of the neurons have **identical parameters**.
- As before, the MLP scans the input with a **stride of 4 time instants**.
- The parameter-sharing pattern of the MLP is illustrated in Figure 27. As mentioned, the **MLP is a 3 layer network with 28 neurons**.
- Neurons with the same color in a layer share the same weights. You may find it useful to visualize the

weights matrices to see how this symmetry translates to weights.

You are required to identify the symmetry in this MLP and use that to come up with the architecture of the CNN (number of layers, number of filters in each layer, their kernel width, stride and the activation in each layer).

The **aim** of this task is to understand how scanning with this **distributed-representation MLP** (with shared parameters) can be represented as a **Convolutional Neural Network**.

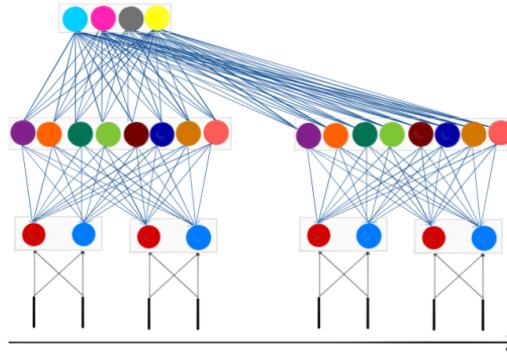


Figure 27: The Distributed Scanning MLP network architecture

```
class CNN_DistributedScanningMLP():
    def __init__(self):
        self.conv1 = # TODO
        self.conv2 = # TODO
        self.conv3 = # TODO
        self.layers = []

    def init_weights(self, weights):
        w1, w2, w3 = weights
        self.conv1.conv1d_stride1.W = # TODO
        self.conv2.conv1d_stride1.W = # TODO
        self.conv3.conv1d_stride1.W = # TODO

    def forward(self, A):
        Z = A
        for layer in self.layers:
            Z = layer(Z)
        return Z

    def backward(self, dLdZ):
        dLdA = dLdZ
        for layer in self.layers[::-1]:
            dLdA = layer.backward(dLdA)
        return dLdA
```

Your tasks include:

- Designing the CNN architecture to correspond to a Distributed Scanning MLP
 - Create Conv1d objects defined as `self.conv1`, `self.conv2`, `self.conv3`, in the init method of `CNN_DistributedScanningMLP` by defining the: `in_channels`, `out_channels`, `kernel_size`, and `stride` for each of the instances.

- Then add those layers along with the activation functions/ flatten layer to a class attribute you create called `self.layers`.
- Initialize the weights for each convolutional layer, using the `init_weights` method. Your job is to discover the orientation of the initial weight matrix and convert it for the weights of the `Conv1d_stride1` layers. This will involve:
 - (1) Reshaping the transposed weight matrix into `out_channels`, `kernel_size`, `in_channels`
 - (2) Transposing the weights again into the correct shape for the weights of a `Conv1d_stride1` instance. You must `slice` the initial weight matrix to account for the shared weights.

The autograder will run your CNN with a different set of weights, on a different input (of the same size as the sample input provided to you). The MLP employed by the autograder will have the `same` parameter sharing structure as your sample MLP. The weights, however, will be different.

For more of a theoretical understanding of Distributed Scanning MLPs, please refer to the Appendix section.

10 Build a CNN model [5 Points]

Finally, in `models/cnn.py`, implement a CNN model.

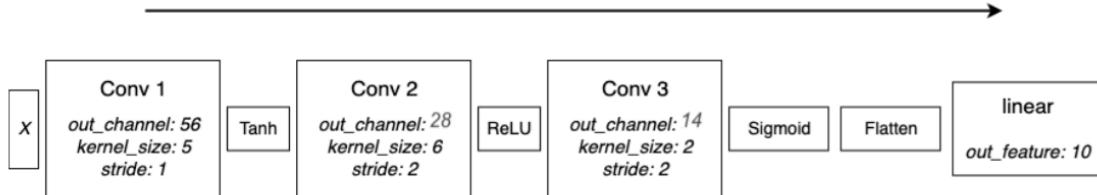


Figure 28: CNN Architecture to implement.

- First, initialize your `convolutional_layers` in the `init` function using `Conv1d` instances.
 - Then initialize your flatten and linear layers.
 - You have to calculate the `out_width` of the final CNN layer and use it to correctly give the linear layer the correct input shape. You can use some of the formulas referenced in the previous sections to calculate the output size of a layer.
- Now, implement the `forward` method, which is extremely similar to the MLP code from HW1.
- There are no batch norm layers.
- Remember to add the `Flatten` and `Linear` layer after the convolutional layers and activation functions.
- Next, implement the `backward` method which is extremely similar to what you did in HW1.
- The `step` function and `zero_gradient` function are already implemented for you.
- Remember that we provided you the `Tanh` and `Sigmoid` code; if you haven't already, see earlier instructions for copying and pasting them in.
- Please refer to the lecture slides for pseudocodes.

We ask you to implement this because you may want to modify it and use it for HW2P2.

Great work as usual!! All the best for HW2P2!!

11 Appendix

11.1 Scanning MLP : Illustration

Consider a 1-D CNN model (This explanation generalizes to any number of dimensions). Specifically consider a CNN with the following architecture:

- Layer 1: 2 filters of kernel width 2, with stride 2
- Layer 2: 3 filters of kernel width 3, with stride 3
- Layer 3: 2 filters of kernel width 3, with stride 3
- Finally a single softmax unit which combines all the outputs of the final layer.

This is a regular, if simple, 1-D CNN. The computation performed can be visualized by the figure below.

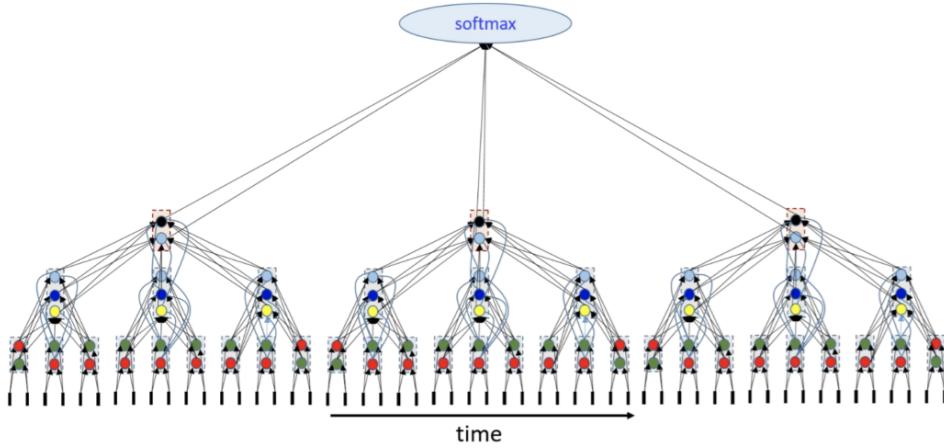


Figure 29: Scanning MLP Figure 1

Input: The little black bars at the bottom represent the sequence of input vectors. There are two layer-1 filters of width 2.

Layer 1: The red and green circles just above the input represent these filters (each filter has one arrow going to each of the input vectors it analyzes, in the illustration; a more complete illustration would have as many arrows as the number of components in the vector).

Each filter (of width 2, stride 2) analyzes 2 inputs (that's the kernel width), then strides forward by 2 to the next step. The output is a sequence of output vectors, each with 2 components (one from each level-1 filter).

In the figure the little vertical rectangular bars shows the sequence of outputs computed by the two layer-1 filters.

The layer-1 outputs now form the sequence of output vectors that the second-layer filters operate on.

Layer-2: Layer 2 has 3 filters (shown by the dark and light blue circles and the yellow circle). Each of them gets inputs from three (kernel width) of the layer-1 bars. The figure shows the complete set of connections. The three filters compute 3 outputs, which can be viewed as one three-component output illustrated by the vertical second-level rectangles in the figure.

The layer-2 filters then skip 3 layer-1 vectors (stride 3) and then repeat the computation. Thus we get one 3-component layer-2 output for every three layer-1 outputs.

Layer 3 works on the outputs of layer 2.

Layer-3: Layer 3 consists of two filters (the black and grey circles) that get inputs from three layer-2 vectors (kernel width 3). Each of the layer 3 filters computes an output, so we get one 2-component output (shown by the orange boxes). The layer-3 units then stride 3 steps over the layer-2 outputs and the compute the next output.

Softmax: The outputs of the layer-3 units are all jointly then sent on to the softmax unit for the final classification.

Now note that this computation is identical to "scanning" the input sequence of vectors with the MLP below, where the MLP strides by 18 steps of input after each analysis. The outputs from all the individual evaluations by the scanning MLP are sent to a final softmax.

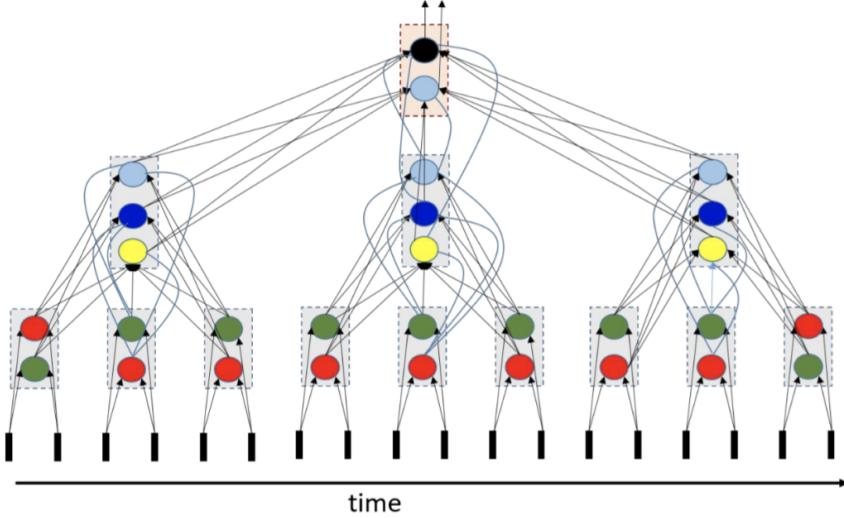


Figure 30: Scanning MLP Figure 2

The "scanning" MLP here has three layers. The first layer has 18 neurons, the second has 9 and the third has 2. So the CNN is actually equivalent to scanning with an MLP with 29 neurons.

Notably, since this is a distributed representation, and although the MLP has 29 neurons, it only has 7 unique neuron types. The 29 neurons share 7 shared sets of parameters.

It's sometimes more intuitive to use a horizontal representation of the arrangement of neurons, e.g.

Note that this figure is identical to the second figure shown. But it also leads to more intuitive questions such as "do the individual groups of neurons (shown in each rectangular bar) have to have scalar activations, or could they be grouped for vector activations. Such intuitions lead to other architectures.

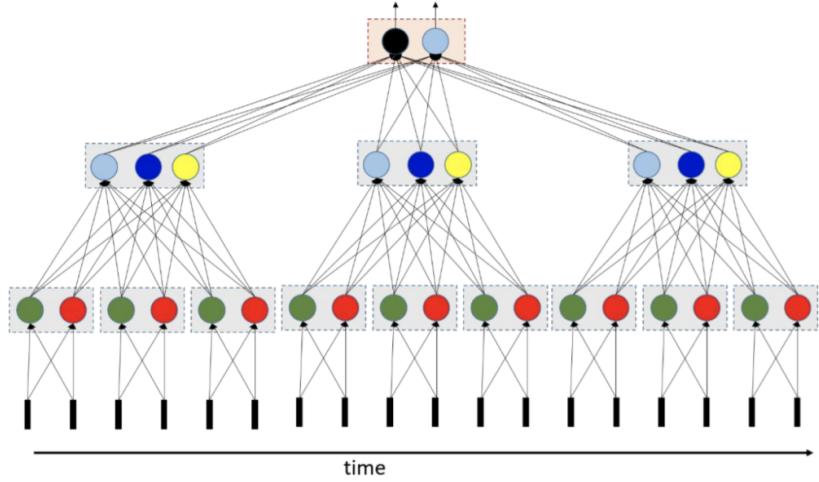


Figure 31: Scanning MLP Figure 3

11.2 Numpy TensorDot

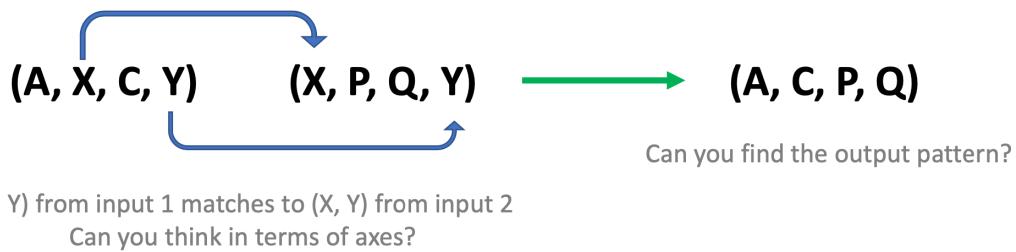
[TensorDot](#) is an amazing tool by numpy which allows you to simplify convolutions. It helps you to reduce the number of loops in your code.

So as the name suggests, tensordot is a function which helps you do a dot product. In the most basic level, it helps you do an elementwise multiplication and summation. It does not do the whole convolution, rather just a step in convolution. Where do you think there is an element wise multiplication and summation in convolution? When the kernel is over a patch/segment of the input image, we do an element-wise multiplication and summation to get a single number as output. After this operation, the kernel strides and does the same operation again. So tensordot does not make you escape a loop. It just reduces the number of loops.

So what constrain do you need 2 arrays to satisfy inorder to do a tensordot? Since its first step is an element-wise multiplication, the shapes of the 2 arrays should be the same. For example, you cannot do an element wise multiplication between a 3×3 array and a 4×4 array.

This begs the question, should the complete shape of the 2 arrays which we use in tensordot be the same? Not required. This is why we specify the axes parameter in tensordot. You can think of a big array as an array of small subarrays. The subarrays can have "matching shapes" to do a tensordot (or you have to make it match in convolution between a big input and a small kernel. Are you really using a big input in each step? Think about it). After doing a tensordot operation, the output which you get will be of shape consisting the unmatched axes in the same order. The below figure tries to explain this.

Tensordot



Should match all the axes that you think needs to be matched. Not restricted to 2 axes

You could also refer to [this blog](#) which has a more detailed explanation of tensordot as well.