به نام خدا

دانشکده علوم و فنون نوین

پردازش گفتار (۱۴۸-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

تمرین شماره ۴

تاریخ تحویل: ۱۴۰۲/۰۳/۲۰

۱. (۳۵٪) [پیاده‌سازی: تشخیص اعداد با شبکه عصبی پرسپترون چندلایه (MLP)]

از شبکه عصبی پرسپترون چندلایه برای تشخیص اعداد ۰ تا ۹ انگلیسی بهره ببرید. برای این کار از دادگان AudioMNIST بهره ببرید. در این دادگان، ۳۰۰۰۰ رکورد در ۱۰ کلاس اعداد ۰ تا ۹ انگلیسی وجود دارد که ۶۰ گوینده گوناگون آن‌ها را گفته‌اند. شبکه MLP باید بصورت دستی پیاده‌سازی گردد و نباید از کتابخانه‌های آماده بهره گرفته شود. برای این پرسش:

گام ۱- فراخوانی دادگان و بخش‌بندی ۸۰/۲۰ آن:

- این داده، ۶۰ پوشه دارد که برای هر گوینده است و در هر پوشه، شماری فایل صدا است که گوینده آن پوشه، آن‌ها را گفته است. لازم است که تمامی ۳۰۰۰۰ صدا را در کنار هم داشته باشید و سپس ۸۰/۲۰ را اعمال کنید.

گام ۲- پیش‌پردازش‌های لازم روی داده:

- از روش MFCC با طول فریم ۲۰ میلی ثانیه، ۲۴ فلیتر مل و ۱۲ ویژگی با مشتق‌های مرتبه یک و دو بهره ببرید.

گام ۳- ساخت مدل MLP:

- از آنجایی که باید ۱۰ کلاس را دسته‌بندی کنیم، تعداد ۱۰ نرون باید در خروجی شبکه عصبی قرارداده شود. شمارگان نرون‌های ورودی شبکه نیز باید به اندازه ویژگی‌های گرفته‌شده از هر فایل صدا باشد.
- یک لایه میانی (مخفی) برای شبکه قرار دهید. برای شمارگان نرون‌ها لایه میانی، میانگین شمارگان نرون‌های لایه ورودی و خروجی درنظر بگیرید.
- تابع فعال‌سازی لایه[های] میانی و خروجی، سیگموید دوقطبی باشد. مقدار نرخ یادگیری را برابر با ۰.۰۰۱ قرار دهید.
- تابع هدف برای بهینه شدن را MSE در نظر بگیرید و از الگوریتم SGD در آموزش بهره ببرید.
- در پایان، از softmax برای یافت بهترین کلاس بهره ببرید.

h.veisi@ut.ac.ir

گام ۴- آموزش و آزمون مدل:

آ- شبکه را آموزش دهید. برای کار یک اندازه کردن همه فایل‌ها با هم، می‌توانید تا از روش zero padding بهره ببرید. چه راهی دیگری برای این کار می‌توانید پیشنهاد دهید. مقدار Accuracy را برای هر دسته و به صورت میانگین و نمودار خطا MSE را برحسب تکرارها گزارش کنید.

ب- تعداد لایه‌های میانی را به دولایه افزایش دهید و بخش آ را در حالت‌های زیر تکرار کنید:

- شمارگان نرون‌ها در هر لایه میانی برابر با آنچه در حالت تک لایه بوده، درنظر بگیرید.
- شمارگان نرون‌های لایه یکم میانی، دو-سوم و نیم شمارگان نرون‌های لایه ورودی باشد.
- شمارگان نرون‌های لایه یکم میانی، نیم و یک-سوم شمارگان نرون‌های لایه ورودی باشد.

پ- **(اضافی)**: از تابع فعال‌ساز ReLU برای انجام بندهای آ و ب بهره ببرید. نتیجه‌ها را باهم مقایسه کنید.

۲. (۳۵٪) [پیاده‌سازی: تشخیص ژانر موسیقی با شبکه عصبی پیچشی(CNN)]

در این تمرین بنا داریم تا ژانرهای موسیقیایی را با طراحی شبکه‌های عصبی پیچشی از هم تشخیص دهیم. دادگان این پرسش، GTZAN است که ۱۰ کلاس ژانر موسیقی در آن است. همه این صداها، طول ۳۰ ثانیه دارد. برای پیاده‌سازی شبکه CNN می‌توانید از کتابخانه‌ها و ابزارهای دلخواه بهره ببرید. برای این پرسش:

گام ۱- فراخوانی دادگان و بخش‌بندی ۸۰/۲۰ آن:

- این داده، ۱۰ پوشه دارد که برای هر کلاس است و در هر پوشه، ۱۰۰ فایل صدا است. لازم است که تمامی صداها را در کنار هم داشته باشید و سپس ۸۰/۲۰ را اعمال کنید.

گام ۲- ورودی شبکه:

- از دادگان، MFCCها را برای ورودی شبکه برگزینید.

گام ۳- ساخت و آموزش مدل:

h.veisi@ut.ac.ir

- بعد از لایه ورودی، دو لایه میانی (مخفی) پیچش و سپس یک لایه flatten بگذارید.
- ساختار همه لایه‌های پیچش، به صورت زیر است:
  - یک لایه پیچش دو بعدی با تعداد ۳۲ کانال و اندازه کرنل ۳*۳
  - سپس یک لایه max pooling با اندازه کرنل ۳*۳ و stride برابر با ۲*۲ و padding مناسب
  - در پایان، یک لایه batch normalization بگذارید.
- در لایه خروجی، یک لایه dense بگذارید. از آنجایی که باید ۱۰ کلاس را دسته‌بندی کنیم، تعداد ۱۰ نرون باید در خروجی شبکه عصبی قرارداده شود.
- در آموزش مدل، از بهینه‌ساز adam بهره ببرید. مقدار نرخ یادگیری را برابر با ۰.۰۰۰۱ قرار دهید. همچنین batch size را برابر با ۳۲ و تعداد epoch را برابر ۳۰ در نظر بگیرید.
- تابع هدف برای بهینه شدن را sparse categorical cross entropy در نظر بگیرید.

گام ۴- آزمون مدل:

همچون پرسش یکم، Accuracy را برای این پرسش نیز گزارش کنید.

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

پردازش گفتار (۱۴۸-۰۵-۸۳)

نیم‌سال دوم ۱۴۰۲-۱۴۰۱

تمرین شماره ۴

تاریخ تحویل: ۱۴۰۲/۰۳/۲۰

۳. (۲۰٪+۳۰٪) [پیاده‌سازی: تشخیص احساس با Autoencoder، CNN]

در این پرسش می‌خواهیم یک برنامه‌ای بنویسید که بتواند صداها را با استخراج ویژگی، تشخیص احساس شادی یا غمگینی صدا را بدهد. در واقع با یک سوال کلاس‌بندی و تشخیص صدا مبتنی بر احساس شادی یا غم روبرو هستیم. دادگان این پرسش از این لینک قابل دریافت است. در این دادگان، ۲۴ فایل صدا ۱۲ فایل صدای مرد و ۱۲ فایل صدای زن با نرخ نمونه برداری ۴۸KHz می‌باشد. هر فایل با منطق زیر از سمت چپ نام گذاری شده است.

- ✓ Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- ✓ Vocal channel (01 = speech, 02 = song).
- ✓ Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- ✓ Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- ✓ Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- ✓ Repetition (01 = 1st repetition, 02 = 2nd repetition).
- ✓ Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

به طور مثال فایلی با اسم ***03-01-06-01-02-01-12.wav*** منطق آن به شکل زیر است.

1. Audio-only (03)
2. Speech (01)
3. Fearful (06)
4. Normal intensity (01)
5. Statement "dogs" (02)
6. 1st Repetition (01)
7. 12th Actor (12)
   Female, as the actor ID number is even.

در این تمرین خواسته شده تمرکز دانشجو روی احساس شادی یا غمگینی باشد یعنی سومین شماره ۳ یا ۴ باشد.

مسئله دسته‌بندی و تشخیص صداها همانطور که در تمرین‌ها و کلاس‌های درس گذشته گفته شد بیشتر با ویژگی از سیگنال گفتار بنام MFCC , Mel Spectrograms انجام می‌گیرد. در این تمرین تلاش می‌کنیم بیشتر با این ویژگی در تشخیص و کلاس بندی احساس آشنا شویم. برای بدست آوردن این

h.veisi@ut.ac.ir

ویژگی، کتابخانه librosa می‌تواند کمک‌کننده باشد (در مورد این ویژگی، سر کلاس TA، صحبت شد).

آ) (۲۰٪) کلاس‌بندی احساس با شبکه پیچشی (نمره اضافی):

در این بخش، می‌خواهیم از داده‌های خوانده شده دو ویژگی ذکر شده استخراج شود و شبکه پیچشی را طراحی و آموزش دهیم گام های آن به شرح زیر است:

گام ۱- دادگان از تک تک فایل‌ها خوانده شود (توجه شود فقط احساس شادی و غمگینی مد نظر است).

گام ۲- دو تابع برای استخراج دو ویژگی فایل‌های صدا بسازید و دادگان را فراهم کنید و ویژگی‌ها را استخراج کنید.

گام ۳- به صورت تصادفی، ۷۰٪ از دادگان برای آموزش و ۱۰٪ برای اعتبارسنجی و ۲۰٪ باقیمانده برای آزمون مدل تقسیم شود. دقت کنید که مدل در آموزش خود نباید از این ۲۰٪ داده را ببیند و این فایل‌ها لطفا ذخیره شود.

گام ۴- شبکه پیچشی در شکل ۱ را با تابع‌های فعال‌ساز: لایه‌های اول و میانی "relu" ، لایه‌ی خروجی از "sigmoid" با کتابخانه‌های آماده طراحی کنید و برای بهینه‌ساز، از "adam" و برای تابع هزینه از "binary_cross_entropy" و برای نرخ یادگیری از ۰.۰۱ استفاده شود و پروسه آموزش را برای هر کدام از دو ویژگی ذکر شده به صورت جداگانه انجام دهید.

h.veisi@ut.ac.ir

Layer (type)	Output Shape
conv2d (Conv2D)	(None, 16, 8, 32)
max_pooling2d (MaxPooling2D)	(None, 8, 4, 32)
conv2d_1 (Conv2D)	(None, 8, 4, 64)
max_pooling2d_1 (MaxPooling2	(None, 4, 2, 64)
flatten (Flatten)	(None, 512)
dense (Dense)	(None, 32)
dense_1 (Dense)	(None, 1)

شکل ۱ شبکه پیچشی مد نظر

در ادامه تعداد دورها را به شکلی قرار دهید که مدل همگرا شود. حداقل تعداد دورها برای هر کدام از دو ویژگی ذکر شده باید چقدر باشد؟ پس از آموزش مدل نمودار هزینه-دور برای آموزش و اعتبارسنجی برای هر کدام از دو ویژگی یادشده رسم و گزارش شود.

در پایان بر روی دادگان آزمون برای هر کدام از دو ویژگی یادشده سنجه‌های recall، precision ، accuracy، f1 score گزارش شود.

h.veisi@ut.ac.ir

ب) (۳۰٪) تشخیص احساس با خود رمزنگارها:

در این بخش، می‌خواهیم از دادههای ذخیره شده‌ی آموزش، اعتبارسنجی و آزمون یک خود رمزنگار را طراحی و آموزش دهیم گام های آن به شرح زیر است:

گام ۱- دادگان از آغاز خوانده شود سپس داده‌های شادی و غمگین جدا شود.

گام 2- دو ویژگی یادشده استخراج شود و برای آماده‌سازی دادگان از "min_max_scaler" استفاده شود و برای هر کدام از داده‌های شادی و غمگین انجام شود.

گام ۳- شبکه خود رمزنگار پیچشی که در شکل ۲ را با تابع‌های فعال‌ساز: لایه‌های اول و میانی "ReLU" ، لایه ی خروجی از "sigmoid" با کتابخانه‌های آماده طراحی کنید و برای بهینه‌ساز، از بهینه‌ساز "Adam" و برای تابع هزینه از "mean_squared" و برای نرخ یادگیری از ۰.۰۱ استفاده شود. سپس مدل برای هر کدام از دو کلاس شادی و غمگینی و برای هرکدام از دو ویژگی یادشده به صورت جداگانه آموزش داده شود .

h.veisi@ut.ac.ir

Layer (type)	Output Shape
conv2d_4 (Conv2D)	(None, 16, 8, 32)
max_pooling2d_4 (MaxPooling2	(None, 8, 4, 32)
conv2d_5 (Conv2D)	(None, 8, 4, 64)
max_pooling2d_5 (MaxPooling2	(None, 4, 2, 64)
conv2d_6 (Conv2D)	(None, 4, 2, 64)
up_sampling2d (UpSampling2D)	(None, 8, 4, 64)
conv2d_7 (Conv2D)	(None, 8, 4, 32)
up_sampling2d_1 (UpSampling2	(None, 16, 8, 32)
conv2d_8 (Conv2D)	(None, 16, 8, 1)

شکل ۲ شبکه خود رمز نگار مد نظر

در ادامه تعداد دورها را به شکلی قرار دهید که مدل همگرا شود. دست کم تعداد دورها برای هر کدام از دو ویژگی و دو داده شادی و غمگین ذکر شده باید چقدر باشد؟ پس از آموزش مدل نمودار هزینه – دور برای آموزش هر کدام از دو ویژگی و دادهی شادی و غمگین ذکر شده رسم و گزارش شود.

در پایان بر روی دادهی آموزش برای هر کدام از دو ویژگی و دادهی شادی و غمگین یادشده مقدار loss گزارش شود.

پیروز باشید

h.veisi@ut.ac.ir