



سیگنال‌ها و سیستم‌ها - دکتر بهروزی

Face Detection With CNN

سینا کریمی

محمد رضا علی محمدی

تیر و مرداد ۱۳۹۹

## خلاصه:

### اهداف:

در این پژوهش می‌خواهیم به کمک شبکه‌ی عصبی پیچشی (CNN) که از قشر بینایی الهام‌گرفته شده است صورت انسان را در تصاویر داده شده تشخیص دهیم.

### بخش‌های مختلف پژوهش:

- شبکه‌های عصبی و شبکه‌های عصبی پیچشی
- اجزای شبکه‌ی عصبی
- استفاده از شبکه‌ی عصبی پیچشی برای تشخیص صورت
- پیش‌پردازش و تعلیم دادن (train) شبکه
- نتایج

## شبکه‌های عصبی و شبکه‌های عصبی پیچشی

ابتدا توضیحاتی در مورد شبکه‌های عصبی بیان می‌کنیم و سپس به صورت خاص شبکه‌های عصبی پیچشی را بررسی می‌کنیم.

### شبکه‌های عصبی:

شبکه‌های عصبی مصنوعی (Artificial Neural Networks) با الهام گرفتن از عملکرد مغز و در راستای شبیه‌سازی عملکرد مغز و حل مسائل متفاوت به وجود آمده‌اند. مدل‌های محاسباتی بسیار متفاوتی برای شبکه‌ی عصبی معرفی شده‌اند که هر کدام از بخش مشخصی الهام گرفته‌اند و کاربرد خاصی دارند. در این پژوهه تمرکز ما بر روی شبکه‌های عصبی پیچشی (Convolutional Neural Networks) و استفاده از آن‌ها برای تشخیص صورت در تصاویر خواهد بود.

### شبکه‌های عصبی پیچشی (ConvNet یا CNN):

در این پژوهه می‌خواهیم به کمک شبکه‌ی عصبی پیچشی (CNN)، که از قشر بینایی الهام گرفته شده‌است، صورت انسان را در تصاویر داده شده تشخیص دهیم. در واقع این شبکه طوری طراحی شده است که می‌توان الگوها را در تصاویر تشخیص دهد و به همین دلیل در پردازش تصاویر بسیار مورد استفاده قرار می‌گیرد. این تشخیص الگو به لایه‌های پیچشی انجام می‌شود؛ هر نورون در این لایه‌ها به کمک کانوالو کردن یک ماتریس ورودی و به دست آوردن ماتریسی جدید می‌تواند وجود الگوی متناظر با ماتریسیش را در تصویر اولیه تشخیص دهد. در واقع هر نورون به این صورت عمل می‌کند که تصویر را پیمایش می‌کند و در هر درایه از ماتریس جدید میزان شباهت آن قسمت از تصویر با ماتریسیش را قرار می‌دهد. به عنوان

مثال فیلتر زیر نوعی تشخیص دهنده‌ی مرز عمودی است:

$$\begin{array}{c}
 \begin{array}{|c|c|c|c|c|c|} \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 \end{array} \\
 6 \times 6
 \end{array}
 *
 \begin{array}{c}
 \begin{array}{|c|c|c|} \hline
 1 & 0 & -1 \\ \hline
 1 & 0 & -1 \\ \hline
 1 & 0 & -1 \\ \hline
 \end{array} \\
 3 \times 3
 \end{array}
 =
 \begin{array}{c}
 \begin{array}{|c|c|c|c|} \hline
 -0 & 30 & 30 & 0 \\ \hline
 0 & 30 & 30 & 0 \\ \hline
 0 & 30 & 30 & 0 \\ \hline
 0 & 30 & 30 & 0 \\ \hline
 \end{array} \\
 4 \times 4
 \end{array}$$

همان‌طور که می‌بینید در وسط تصویر ورودی مرزی وجود دارد. و این باعث شده است که درایه‌های ستون‌های میانی در خروجی مقادیر زیادی داشته باشند که نشان‌دهنده‌ی وجود مرز در آن نقاط است.

در ادامه قسمت‌های مختلف شبکه‌ی عصبی پیچشی را توضیح می‌دهیم و سپس شبکه‌ی عصبی مورد استفاده برای تشخیص وجود صورت در را شرح می‌دهیم.

### لایه‌ی پیچشی (Convolutional layer)

هر نورون در این لایه عملی مشابه توضیح صفحه‌ی قبل انجام می‌دهد. این لایه برای تشخیص الگوها استفاده می‌شود.

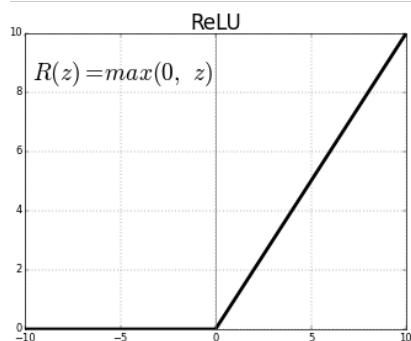
### لایه‌ی ادغام (Pooling Layer)

Image Matrix			
2	1	3	1
1	0	1	4
0	6	9	5
7	1	4	1

Max Pool	
2	4
7	9

پس از عبور از این لایه اندازه‌ی تصویر کوچک می‌شود به این صورت که به طور مثال تصویر به قسمت‌های  $2 \times 2$  تقسیم می‌شود و از هر کدام از این قسمت‌ها یک درایه انتخاب می‌شود. برای کاربرد مورد استفاده‌ی ما به دنبال تشابه الگو و ورودی می‌گردیم و از این رو بزرگترین درایه‌ی هر قسمت انتخاب می‌شود و به همین دلیل به آن لایه‌ی ادغام بیشینه (Maxpool Layer) می‌گوییم.



### واحد یکسوساز خطی (Rectified Linear Unit)

درایه‌های کوچک‌تر از صفر هر تصویر بعد از عبور از این لایه تبدیل به صفر می‌شوند و درایه‌های بزرگ‌تر از صفر بدون تغییر باقی می‌مانند. این عمل باعث می‌شود ماتریس‌ها تُنک (Sparse) بشوند و عملیات ماتریسی سریع‌تر انجام شود.

### لایه‌ی کاملاً متصل (Fully Connected Layer)

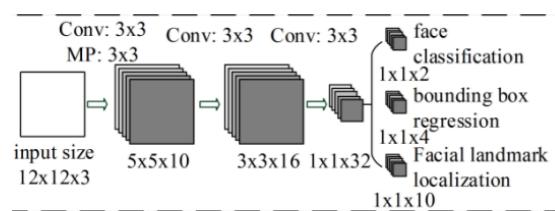
بعد از تمام لایه‌های پیچشی یک لایه کاملاً متصل قرار می‌گیرد که به ازای هر خروجی و هر کدام از درایه‌های باقی‌مانده وزنی نسبت داده می‌شود. با جمع زدن وزن‌دار درایه‌ها برای هر خروجی عددی به دست می‌آید. با مقایسه‌ی این عدددها مشخص می‌شود که کدام خروجی باید انتخاب شود. این لایه برای طبقه‌بندی (Classification) قرار داده می‌شود.

حال شبکه‌ی استفاده شده برای تشخیص صورت را توضیح می‌دهیم.



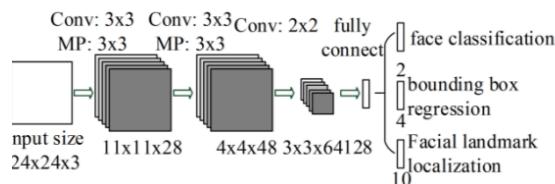
### شبکه‌ی استفاده شده در Pnet:

این شبکه از ۴ لایه‌ی پیچشی و ۱ لایه‌ی تماماً متصل تشکیل شده است. یک بار به کمک لایه‌ی ادغام اندازه‌ی تصویر کوچک می‌شود و پس از هر لایه‌ی پیچشی یکسوساز قرار دارد. این شبکه با دقیقی حدود ۸۰ درصد می‌تواند صورت در عکس را تشخیص دهد. ما در این پروژه خروجی این بخش را به که شبکه‌ی بسیار بزرگ‌تر و پیچیده‌تر است می‌دهیم تا با دقت بیشتری وجود یا عدم وجود صورت در بخش خاصی از تصویر که Pnet به عنوان صورت معرفی کرده را تشخیص دهد. در این حالت دقیق شبکه تا بیش از ۹۹ درصد بالا می‌رود.



### شبکه‌ی استفاده شده در Rnet:

همان‌طور که در بالا توضیح داده شد این لایه به نوعی نقش تصحیح لایه‌ی Pnet را دارد. شبکه‌ی Rnet مورد استفاده‌ی ما از ۱۴ لایه تشکیل شده است که تعداد نورون‌های لایه‌های آن هم از Pnet بیش‌تر است. توجه کنید تصویر ورودی Rnet دو برابر تصویر ورودی Pnet است. اما شبکه‌ی آن یک ادغام بیش‌تر دارد.

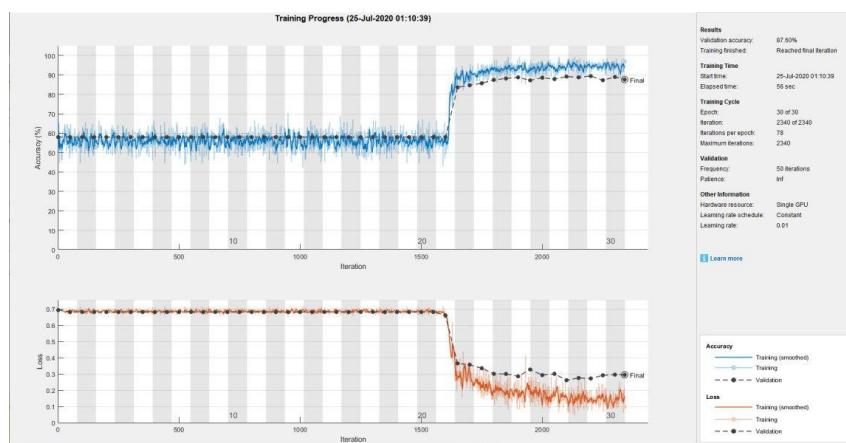


## پیش‌پردازش (Preprocess):

به کمک روش adaptive histogram equalizer تصاویر ورودی را پیش‌پردازش می‌کنیم به این صورت است که مقدار هر نقطه از تصویر را از میانگین نقاط اطراف آن کم می‌کنیم. در واقع هدف از این کار این است که تمایز صورت با زمینه بیشتر شود. این پیش‌پردازش باعث این می‌شود که شبکه دقت بهتری داشته باشد.

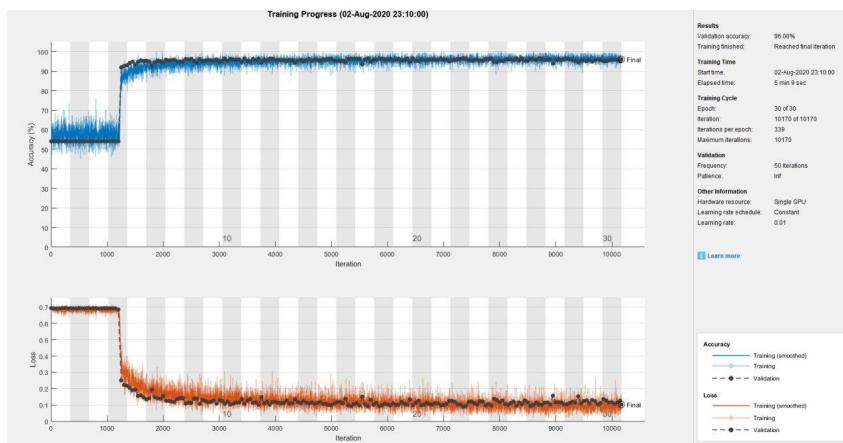
## تعلیم شبکه‌ی عصبی:

برای تعلیم شبکه از دیتابیس Wider\_images استفاده کردیم و هر عکسی که یک صورت در آن وجود داشت را در ۶ اندازه‌ی متفاوت (۷-۱۲) به شبکه داده‌ایم و مشابه این برای تصویرهایی که در آن‌ها صورت نیست عمل می‌کنیم. راهنمایی از الگوریتم پس‌انتشار (Backpropagation) برای اصلاح مقادیر اولیه استفاده کردیم. تصویر اول مربوط به زمانی است که داده‌ها بدون پیش‌پردازش به شبکه‌ی Pnet داده می‌شدند. دقت شبکه در این حالت حدود ۸۸ درصد است و نسبتاً دیر به

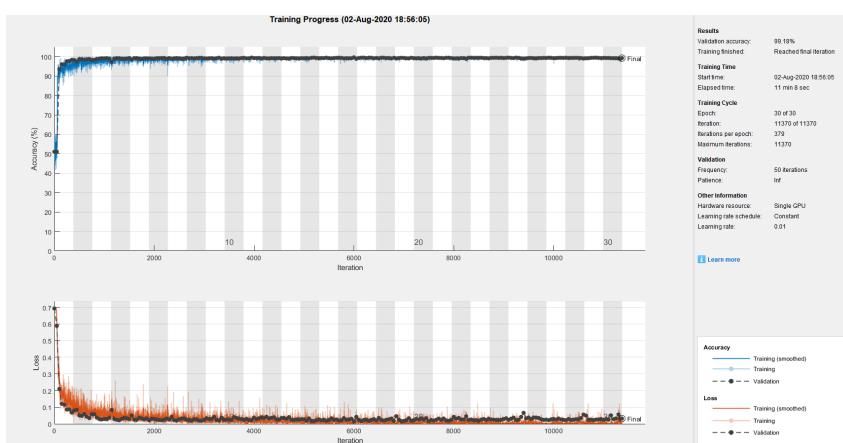


دقت مناسب می‌رسد.

حال اگر تصاویری که برای تعلیم شبکه می‌دهیم را پیش‌پردازش کنیم دقیق شبكه به ۹۶ درصد می‌رسد و با داده‌های کمی به دقیق مناسب می‌رسد.



برای تعلیم Rnet تصاویر را در ۵ اندازه‌ی متفاوت (۱۹-۲۴) به شبکه داده‌ایم.  
نمودار زیر مربوط به تعلیم شبکی Rnet است که بزرگ‌تر و پیچیده‌تر است و البته داده‌های بیشتری برای تعلیم نیاز دارد.  
همان‌طور که می‌بینید دقیق این شبکه به بیش از ۹۹ درصد می‌رسد!



عکسی را به عنوان ورودی می‌گیریم و پیش‌پردازش گفته شده را روی آن انجام می‌دهیم عکس بعد از پیش‌پردازش به شکل زیر خواهد بود:



حال این تصویر را به شبکه‌ی Pnet می‌دهیم این شبکه کادرهای قرمز را به عنوان صورت تشخیص می‌دهد:



قسمت‌هایی که Pnet به عنوان صورت تشخیص داده بود را به Rnet می‌دهیم تا درستی آن‌ها را بررسی کند این شبکه کادرهای قرمز را به عنوان صورت تشخیص می‌دهد:



با تار کردن قسمت‌هایی که شبکه‌ی Rnet هم درستی آن‌ها را تایید کرده به تصویر نهایی می‌رسیم؛ تار کردن با میانگیری در ناحیه‌ی مورد نظر اتفاق می‌افتد:



منابع:

[https://www.youtube.com/watch?v=YRhxdk\\_sIs&feature=youtu.be](https://www.youtube.com/watch?v=YRhxdk_sIs&feature=youtu.be)

<https://www.youtube.com/watch?v=ILsA4nyG7I0>

<https://www.youtube.com/watch?v=FmpDIaiMleA>

<https://towardsdatascience.com/face-detection-neural-network-structure-257b8f6f85d1>

---