40014140111035

مهتاب نوروزقزوینی

موضوع مقاله Nonparallel Emotional Speech Conversion

چکیده

هنر تقلید صدای انسان با کامپیوتر، یکی از چالش برانگیزترین موضوعات پردازش گفتار در سالهای اخیر بوده است.یک سیستم تبدیل گفتار دارای دو سمت است. در یک سمت آن، گوینده مبدا قرار دارد که صدایش برای تقلید صدای گوینده

هدف که در سمت دیگر سیستم قرار دارد تغییر داده میشود. برای تبدیل گفتار فرد مبدا به فرد هدف از دو روش موازی و ناموازی استفاده میشود. در روش موازی گوینده مبدا و هدف جمالت یکسانی بیان کرده و در روش ناموازی جمالت متفاوتی

بیان میکنند. بیشتر محققین تبدیل گفتار برای آموزش تابع تبدیل از دادگان آموزشی موازی استفاده کردهاند. با این حال، در عمل همیشه امکان جمعآوری دادگان موازی وجود ندارد و بنابراین نیاز استفاده از روشهای ناموازی به وجود میآید.

مواد و روشها: گفتار گوینده مبدا و هدف ضبط شده و سپس مورد آنالیز قرار گرفت. با پردازش سیگنال، ویژگیهای گفتار هر دو نفر استخراج شد. سپس عمل هم ردیف سازی انجام شده و تابع تبدیل گفتار بدست آمد. برای تبدیل گفتار

مبدا به هدف، گفتار مبدا آنالیز شده و سپس عمل استخراج ویژگی انجام شد. تابع تبدیل گفتار بدست آمده از قسمت قبل، بر ویژگیهای استخراج شده اعمال شد. سپس عمل معکوس استخراج ویژگی انجام شده و در پایان سنتز گفتار صورت

گرفت. صدای سنتز شده، صدای فرد هدف میباشد.

یافتهها: نتایج آزمایشهای عددی و عینی مشخص کرد که روش پیشنهادی ما از روش آموزش موازی بهتر است. همچنین

در آزمایشها مشاهده شد که این برتری هم از لحاظ کیفیت و هم از لحاظ شباهت به گویندهی هدف، برای اندازههای مختلف دادگان آموزشی از پنج تا چهل جمله صادق است

.

بحث و نتیجه گیری: به نظر میرسد که روش پیشنهادی ما یک رقیب جدی برای روشهای آموزش موازی برای همردیف سازی فریم است.

کلمات کلیدی: تبدیل گفتار، آنالیز و سنتز صدا، سیستمهای آموزش ناموازی، الگوریتم INCA ،مدل مخلوط گاوسی، مدل پس زمینه سراسری، تبدیل گفتار بالدرنگ

در بسیاری از منابع، سیستم پردازش گفتار با پردازش صدا یکسان در نظر گرفته می‌شود. در حالی که این دو حوزه فناوری، اندکی با یکدیگر متفاوت هستند. تمرکز سیستم‌های پردازش گفتار، بیشتر بر کلمات و عبارات گفته‌شده می‌باشد که تبدیل گفتار از قالب کلامی به متن و انجام تجزیه و تحلیل بر روی آن یکی از برجسته‌ترین کاربردهای این حوزه است.

.1مقدمه

تبدیل گفتار، هنر تقلید صدای انسان با کامپیوتر، یکی از چالشی ترین موضوعات پردازش گفتار در سالهای اخیر بوده است. یک سیستم تبدیل گفتار دارای دو سمت است. در یک سمت آن، گوینده مبدأ

قرار دارد که صدایش برای تقلید صدای گوینده هدف که در سمت دیگر سیستم قرار دارد تغییر داده میشود. عملکرد یک سیستم تبدیل گفتار به کیفیت طبیعی بودن(و شباهت)به گویندهی هدف

صدای تبدیل شدهی آن بستگی دارد. بسته به اینکه گوینده های مبدأ و هدف جمالت آموزشی یکسان یا متفاوتی را ادا کرده باشند، روشهای تبدیل گفتار به ترتیب به روشهای با دادگان موازی یا

ناموازی تقسیم میشوند. بیشتر محققین این رشته ترجیح میدهند که از آموزش موازی استفاده کنند تا بتوانند روی دقت تابع نگاشت رگرسیون تمرکز کنند. به طور حتم، جمع آوری دادگان موازی

برای همهی سناریوهای عملی امکان پذیر نیست، بنابراین طراحی روشهایی که بتوانند با دادگان ناموازی کار کنند ضروری است.

در روشهای آموزش موازی، همردیف سازی فریم های گویندگان مبدأ و هدف، با اعمال الگوریتم پیچش زمانی پویا DTW)) به جفتهای متناظر آنها صورت میگیرد. در مرحله ی بعد،

یک تابع تبدیل دلخواه از روی زوج ویژگیهای جفت شده، تخمین زده میشود. چند نمونه از توابع تبدیل عبارتند از: نگاشت مبتنی بر مدل مخلوط گوسی ، GMMرگرسیون کمترین مربعات جزئی

با هسته ی پویا، DKPLSسیستمهای دینامیکی خطیLDS و شبکههای عصبی عمیق .بحث در مورد مزیتها و کمبودهای این روشها در این تحقیق نمیگنجد، چون که تحقیق ما روی روشهای آموزش ناموازی متمرکز است .

همچنین ویژگیهای طیفی مختلفی برای تشکیل تابع تبدیل مورد استفاده قرار میگیرند .

روشهای تبدیل گفتار ناموازی، به دو دسته تقسیم میشوند. دسته اول روشهایی هستند که سعی نمیکنند فریمهای ناموازی مبدأ و هدف را همردیف کنند. آنها مستقیما از دادههای مبدأ و هدف برای

ساختن یا تطبیق تابع تبدیل استفاده میکنند.

نحوه کار سیستم‌های پردازش گفتار

سیستم‌های پردازش گفتار به طور معمول دارای یک فرایند چند مرحله‌ای هستند. ابتدا، ویژگی‌های مربوطه از سیگنال گفتار استخراج می‌شوند. سپس، مدل‌های مرجع با استفاده از این ویژگی‌ها طراحی می‌یابند. در مرحله سوم، بردارهای ویژگی استخراج شده از گفتار به مدل‌های مرجع ارسال می‌شوند.

لازم است مدل‌های مرجع برای هر واحد صدا (واج) ایجاد شوند. مدلی که بالاترین میزان اطمینان را ایجاد می‌کند، هویت واحد صدا را نشان می‌دهد. همچنین، توالی واحدهای صوتی شناسایی شده با استفاده از مدل‌های زبانی اعتبارسنجی می‌شود. به عبارت دیگر، از مدل‌های زبانی برای تبدیل دنباله واحدهای صوتی به متن استفاده می‌شود.

از نظر مفهومی، رویکردهای ایجاد سیستم‌های پردازش گفتار به دو نوع مبتنی بر الگو و یا مدل تقسیم‌بندی می‌شوند. در رویکرد مبتنی بر الگو، ابتدا سیستم با استفاده از الگوهای گفتاریِ شناخته شده آموزش داده می‌شود. سپس، با مقایسه سیگنال‌های گفتاری ناشناخته با الگوهای احتمالی آموخته شده در مرحله آموزش، پردازش انجام می‌شود.

توالی احتمالی کلمات که فاصله بین الگوهای ناشناخته و الگوی شناخته شده را به حداقل می‌رساند، به عنوان توالی بهینه انتخاب می‌شود. الگوریتم پیچش زمانی پویا (DTW) و کوانتیزاسیون برداری (VQ) از جمله روش‌های رایج در این زمینه هستند. در سیستم‌های مبتنی بر مدل، ویژگی‌های مناسب برای هر واحد صدا (واج) از داده‌های آموزش استخراج می‌شوند. لازم است مدل‌های مرجع برای هر واحد صدا ایجاد شوند.

از روش‌های متداول این نوع از مدل‌سازی می‌توان به مدل پنهان مارکوف(HMM) ، مدل مخلوط گوسی(GMM) ، شبکه عصبی (NN) و ماشین بردار پشتیبان (SVM) اشاره کرد. مدل‌های پنهان مارکوف و شبکه‌های عصبی از مدل‌های رایج پردازش گفتار هستند

مدل‌های پنهان مارکوف :(HMM) بسیاری از سیستم‌های بازشناسی گفتار براساس مدل‌های پنهان مارکوف بنا شده‌اند. روش HMM که بر اساس اصول احتمالات عمل می‌کند، پردازش گفتار را در سه سطح کلی انجام می‌دهد. در سطح نخست، شناسایی واج‌ها و یا واحدهای صدا انجام می‌گیرد.

در مرحله دوم، توالی واج‌ها و ساخت کلمات مورد بررسی قرار می‌گیرد. بدین منظور، واج‌هایی که در کنار هم بیشترین احتمال را دارند، انتخاب شده و کلمات را تشکیل می‌دهند. هدف مرحله سوم، ایجاد توالی بهینه کلمات و ایجاد جمله است. در این مرحله احتمال وجود فعل‌ها، اسم‌ها، قید و یا صفت در کنار هم ارزیابی می‌شود و ترکیبی که دارای بیشترین احتمال است به عنوان گزینه نهایی انتخاب می‌گردد. از مزایای این روش دقت بالای آن در شناسایی توالی کلمات است. با این حال در شناسایی واج‌ها با تلفظ‌ها و یا لهجه‌های مختلف دارای انعطاف کمی می‌باشد.

شبکه‌های عصبی :(NN) شبکه‌های عصبی همانطور که از نامش نیز مشخص است، شبکه‌هایی از گره‌های بهم پیوسته می‌باشد که نحوه عملکرد آن مشابه با مغز انسان است. ارتباطات بین این گره‌ها توسط شاخص وزن‌ها مشخص می‌شود که با آموزش شبکه، مقدار آن‌ها به طور بهینه تعیین می‌گردد. انعطاف‌پذیری بالا از مزیت‌های ارزشمند این روش است.

سیستم‌های تحلیل صوت

با توجه به عبارات و لحن استفاده شده در گفتار افراد در کنار سایر ویژگی‌های صوتی می‌توان ویژگی‌های گوینده و نوع گفتار او را تحلیل نمود. به طور کلی، برخی از کارکردهای مهم این نوع از سیستم‌ها به شرح زیر می‌باشند:

تشخیص احساس، سن و جنسیت

تشخیص زبان گفتار

تشخیص و تأیید گوینده

تعیین نوع بیان جمله

تشخیص میزان هوشیاری یا خواب‌آلودگی

سیستم‌های سنتز گفتار

پردازش گفتار، قابلیت ایجاد یک فایل صوتی سفارشی همراه با احساس مورد نظر را دارد. تبدیل متن به گفتار، یکی از برجسته‌ترین کاربردها در این زمینه است که می‌تواند در موارد مختلف از جمله خواندن اخبار و یا چت‌بات‌ها مورد استفاده قرار گیرد.

همچنین تبدیل صوت افراد دارای اختلالات گفتاری، به شیوه‌ای قابل فهم، از دیگر قابلیت‌های پردازش گفتار در این زمینه است. به طور کلی، کارکرد سیستم‌های سنتز گفتار شامل موارد زیر می‌تواند باشد:

تبدیل متن به گفتار

تبدیل صوت

تغییر و یا افزودن احساس دلخواه به صوت

تولید گفتار سفارشی با صدای فرد مورد نظر

این روش , انتقال ویژگی های مرتبط با احساسات یک سیگنال گفتار را ممکن می سازد در حالی که هویت گوینده و محتوای زبانی را حفظ می کند.

داده های موازی و هم ترازی زمانی که در بسیاری از موارد در دسترس نیست.

کاربردهای واقعی: ما براساس a به آموزش های بی نظیری دست پیدا می کنیم.

تکنیک انتقال سبک بدون نظارت , که یک مدل ترجمه بین دو توزیع را به جای یک قطعی یاد می گیرد .

نقشه برداری تک به تک بین نمونه های جفتی . تبدیل مدل شامل یک رمزگشا و یک رمزگشا برای هر احساس است.

دامنه : فرض می کنیم که سیگنال گفتار می تواند تجزیه شود.

به صورت یک کد محتوایی متغیر و مرتبط با احساسات

کد سبک در فضای نهفته تبدیل احساسات انجام می شود.

با استخراج و بازترکیب کد منبع نتایج ارزیابی اثربخشی رویکرد ما را نشان می دهد.

اصطلاحات شاخص : تبدیل گفتار احساسی , بی نظیر

آموزش , انتقال سبک , اتوکدر,GAN

تبدیل صدا (VT) تکنیکی برای اصلاح برخی ویژگی های گفتار انسان در عین حفظ اطلاعات زبانی آن است.

تبدیل صدا (VVC) یا تغییر سبک صحبت کردن از جمله, تبدیل احساس و لهجه به یکدیگر.

هدف , تغییر ویژگی های مرتبط با احساسات یک سیگنال گفتاری و در عین حال حفظ محتوای زبانی و هویت گوینده است.

تکنیک های تبدیل احساسات را می توان برای کاره ای مختلف به کار برد ,مانند تقویت گفتار تولید شده توسط کامپیوتر , پنهان کردن نگاتیو احساسات مردم , کمک به دوبله فیلم و خلق آثار بیشتر پیام های صوتی گویا در شبکه های اجتماعی

رویکردهای سنتی VC را نمی توان مستقیما به کار برد زیرا آن ها هویت گوینده را با فرض تلفظ و لحن به عنوان بخشی از اطلاعات وابسته به گوینده تغییر می دهند.

برخی از مطالعات بر روی مدل سازی متمرکز شده اند.

ویژگی های پروداکشن مانند گام , ضرب اهنگ و حجم در یک سیستم تبدیل صدای احساسی مبتنی بر قاعده پیشنهاد شد . این سیستم ویژگی های آکوستیک مرتبط با پروپسودی گفتار خنثی را برای تولید انواع مختلف احساسات اصلاح می کند . یک گفتار ابزار آنالیز - سنتز برای استخراج فرکانس پایه (F0) و پوشش توان از صوت خام مورد استفاده قرار گرفت.

در مدل فوجیتسو و مدل پیش بینی هدف با این حال , این روش نیاز به داده های موازی هم تراز زمانی دارد که به دست آوردن آن ها در کاربردهای واقعی و زمان دقیق دشوار است.

هم ترازی نیاز به بخش بندی دستی سیگنال گفتار در سطح بندی ولوم که بسیار زمان بر است.

برای پرداختن به این مسائل , ما یک روش آموزشی بی نظیر را پیشنهاد می کنیم.

فرض می کنیم که هر سیگنال گفتار xi را می توان به یک کد محتوایی C تجزیه کرد که نشان دهنده اطلاعات متغیر و یک کد سبک Si است.

که نشان دهنده اطلاعات وابسته به احساسات است C . به اشتراک گذاشته می شود.

سی دامنه است و شامل اطلاعاتی است که می خواهیم آن ها را حفظ کنیم.

در مرحله تبدیل , کد محتوا را استخراج می کنیم.

از گفتار منبع و بازگویی آن با کد سبک.

احساسات هدف : یک شبکه خصمانه مولد (GAN)

برای بهبود کیفیت گفتار تبدیل شده اضافه می شود.

هر گونه عملیات دستی ما رویکرد خود در IEMOCAP را برای چهار مورد ارزیابی کردیم.

احساسات : عصبانی , خوشحال , خنثی , غمگین ; که به طور گسترده در ادبیات تشخیص گفتار احساسی به دانش ما ,این اولین تلاش برای تبدیل احساسات بی نظیر در این زمینه است.

مجموعه داده ها , هر چند بازنمایی های مصنوعی از احساسات ما توانایی تبدیل مدل ها را با درصد تغییر از احساس منبع ارزیابی می کنیم.

ارزیابی ذهنی در آمازونMturk با صدها شنونده انجام شد ; الگوی ما را نشان می دهد

می تواند به طور موثری احساسات را تغییر دهد و هویت گوینده را حفظ کند.

بخش 2 :بقیه مقاله به شرح زیر سازماندهی شده است ارایه رابطه با کار قبلی .

بخش 3 جزئیاتی را ارایه می دهد.

در بخش 4توصیف مدل ما . آزمایش ها و نتایج ارزیابی.

در نهایت در بخش 5 نتیجه گیری می کنیم.

۲ -کار مرتبط

۲/۱ .ویژگی های مرتبط با احساسات

روش های تبدیل احساسات قبلی به طور مستقیم ویژگی های مرتبط با پارامتر را که احساسات را منتقل می کنند، اصلاح می کنند.

مدل های مخلوط گاوسی(GMM) برای تبدیل طیف برای اولین بارپیشنهاد شد.

انواع ویژگی های آکوستیک: توالی طیفی، مدت زمان و توان پوشش، و بررسی تاثیر آن ها بر سنتز گفتار احساسی نویسندگان دریافتند کهF ۰ و طیفی

توالی عوامل غالب در تبدیل احساسات هستند، در حالی که

قدرت و مدت زمان به تنهایی تاثیر چندانی ندارد.

علاوه بر این ادعا شد که می توان تمام احساسات را با اصلاح توالی طیفی سنتز کرد، اما روشی برای انجام این کار ارائه نداد.

در این مقاله بر یادگیری مدل های تبدیل تمرکز می کنیم.

F۰ و دنباله طیفی.

یک مدل مخفی مارکف) HMM) برای گویندهی هدف و یک GMM برای گویندهی مبدأ ساخته میشود. تابع تبدیل به صورت مخلوط تبدیالت خطی ساده فرض میشود و با بیشینه

کردن احتمال بردارهای تبدیل شدهی مبدأ نسبت به HMM هدف، آموزش داده میشود. این روش جذاب است اما در عین حال کمبودهایی دارد. محدودیت اصلی آن، نیاز به اطالعات آوایی

برای دادگان گفتاری است. مشکل دیگر وقتی به وجود میآید که الگوریتم بیشینه سازی انتظار EM))اعمال میشود. در هر حالت از HMM هدف، یک GMM وجود دارد که برای تعیین احتمالات

پسین اولیهی اینGMM ها برای هر بردار تبدیل شده، بردار مبدأ به آنها داده میشود. این آغازسازی، دقت تخمین تابع تبدیل را کاهش میدهد. دلیل این است که وقتی فضای آکوستیکی مبدأ از

فضای آکوستیکی هدف دور است برای مثال در مورد تبدیل مرد به زن، این آغازسازی فضاهای آکوستیکی را با هم مخلوط میکند.

در مرحلهی تبدیل، ابتدا دنباله ی فریم های تست مبدأ با این تابع، تبدیل میشوند. سپس، این فریمهای نیمه تبدیل یافته، با بهترین دنبالهی فریمهای منطبق خود در دادگان آموزشی گویندهی هدف، جایگزین میشوند .

با این کار، فریمهای تبدیل شدهی نهایی به دست میآیند. روشی که برای انتخاب بهترین دنبالهی فریمهای منطبق استفاده میشود، انتخاب واحد(Selection Unit) است که از TTS قرض گرفته شده است.

رویکردهای آموزشی بی نظیر

داده موازی به معنای جملاتی با محتوای زبانی یک سان است.

از آنجا که داده های موازی سخت هستند،

برای جمع آوری، روش های بی نظیری توسعه داده شده است.

ایده ها را از ترجمه تصویر به تصویر قرض بگیرید و خلق کنید.

مدل های GAN مناسب برای گفتار، مانندVC-VAW-GAN

VC-StarGAN ،VC-CycleGAN ،VC-StarGAN

روند دیگر مبتنی بر مدل های خودرگرسیو مانند

WaveNetاگرچه می تواند به طور مستقیم و بدون استخراج ویژگی ها، بار محاسباتی سنگین و حجم زیاد آموزش ببیند.

میزان داده های آموزشی مورد نیاز برای اکثر کاربران مقرون به صرفه نیست.

۲/۳ .یادگیری نمایش درهم تنیده

کار ما از مطالعات اخیر در سبک تصویر الهام می گیرد.

انتقال: یک ایده اساسی پیدا کردن نمایش های درهم تنیده است که می توان به طور مستقل محتوا و سبک تصویر را مدل سازی کرد.

که یک شبکه عصبی کانولوشن(CNN) یک ایده آل است.

بازنمایی برای فاکتوربندی محتوای معنایی و سبک هنری آن ها روشی برای جداسازی و ترکیب مجدد محتوا معرفی کردند.

و سبک تصاویر طبیعی با تطبیق همبستگی ویژگی ها درلایه های کانولوشن مختلف برای ما، وظیفه یافتن بازنمایی های گسسته برای سیگنال گفتار است که بتواند احساسات را تقسیم کند.

از هویت گوینده و محتوای زبانی

روش ها

پژوهش در مورد بیان احساسات و ادراک انسان

دو نتیجه گیری عمده: نخست، درک احساسات انسان است.

فرآیند چند لایه.

مدل سه لایه و یادگیری اتصالات توسط یک سیستم استنتاج فازی. برخی از محققان دریافتند که اضافه کردن لایه های میانی

بر این اساس می توان دقت تشخیص احساسات را بهبود بخشید.

استفاده از پرسپترون های چندلایه را پیشنهاد می کنیم.

استخراج اطلاعات مرتبط با احساسات در سیگنال های گفتاری

دوم، فرآیند تولید احساسات گفتار انسان

جهت مخالف ادراک احساسات را دنبال می کند.

یعنی فرآیند رمزگذاری گوینده، عملیات معکوس فرآیند رمزگشایی شنونده است.

تولید و ادراک گفتاری احساسی، روش بازنمایی یکسانی دارند، یعنی کدکننده و رمزگشا عبارتند از:

عملیات معکوس با ساختارهای آینه ای دسته بندی های احساسی مختلف: هدف ما یادگیری یک نقشه است.

مدل

مدل مولد گفتار را به صورت جزئی نشان می دهد.

فرض کنیم یک کد نهفته مشترک و مرتبط با احساسات داشته باشیم.

دارای یک رمزگشا قطعی و معکوس آن است.

برای تبدیل احساسات، ما فقط کد منبع را استخراج و دوباره کامپایل کنید

راه اندازی آزمایشی

ما رویکرد پیشنهادی را در پایگاه داده ثبت حرکت احساسی تعاملی (IEMOCAP) ارزیابی کردیم.

که در 5 جلسه برگزار می شود پر کردن بلندگوها در سناریوهای نوشته شده و بداهه، که در آن.

احساسات به طور طبیعی برانگیخته می شوند . در این مقاله ما تنها چهار دسته احساسی را در نظر می گیریم : 1) خشم , 2) شادی , 3) خنثی,

(4از آنجایی که مدل برای تغییر گوینده طراحی نشده است

برای هر سخنران به طور مستقل آزمایش هایی انجام می شود.

برای مثال در جلسه 1 , 420 نفر حضور دارند.

نمونه های آموزشی و 108 نمونه آزمون برای گوینده زن

نمونه های آموزشی با طول ثابت 128 فریم به طور تصادفی از توالی های صوتی خام انتخاب می شوند .

استفاده از وکودر WORLD برای استخراج فرکانس های اساسی ,توالی های طیفی (sps) و تناوبی (saps) از صوت خام شکل های موج نمونه برداری شده در 16 .KHzطول چارچوب 5 msاست.

کدینگ , 24 ضریب اول مل - استرال (MCEPs) را در نظر می گیریم.

به عنوان بردارهای ویژگی . میانگین و واریانس کل مجموعه آموزشی برای نرمال سازی ویژگی محاسبه می شوند . نمونه های تست می توانند طول زمانی دل خواه دارند و در زمان واقعی تبدیل می شوند.

معماری شبکه

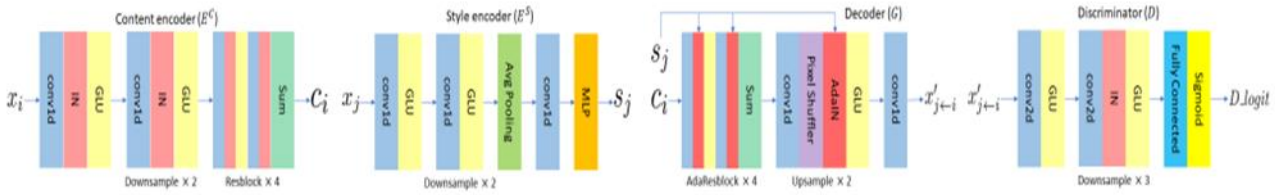


Figure 4: *The network structure of content encoder, style encoder, decoder, and GAN discriminator.*

به عنوان ورودی و یادگیری نمایش های درهم تنیده محتوا و سبک : در رمزگذاری محتوا , نرمال سازی نمونه حذف میانگین و واریانس مشخصه اصلی که نشان می دهد

اطلاعات سبک احساسی : در کدکننده سبک , ویژگی های احساسی توسط یک MLP سه لایه کدگذاری می شوند که میانگین کانال و واریانس را تولید می کند.

احساسات از طریق نرمال سازی نمونه تطبیقی اضافه می شود.

(AdaIN)لایه قبل از فعال سازی : این مکانیزم مشابه مدل تبدیل F0 در معادله است.

ما گفتار تولید شده را براساس سه معیار ارزیابی می کنیم : کیفیت صدا , شباهت گوینده و توانایی تبدیل احساسات.

ارزیابی ذهنی ما تست های ادراک را روی ترک مکانیکی آمازون 1 انجام می دهیم.

بازشناسی یا تشخیص چهره (Face Recognition) چیست؟

سیستم تشخیص چهره (Face Recognition) یک فناوری است که می‌تواند شخص را ازطریق یک تصویر دیجیتال یا یک فریم ویدئو از یک منبع ویدیویی شناسایی یا تأیید کند. تشخیص چهره تکنیک شناسایی چهره‌ی افرادی است که در مجموعه‌ای داده وجود دارند. بااینکه تشخیص چهره، در مقایسه با دیگر انواع تکنیک‌های تشخیص، دشوارتری است، به‌دلیل اینکه انسان‌ها معمولاً افراد را با چهره‌شان شناسایی می‌کنند، این حوزه همواره تمرکز اصلی محققان بوده است.

در حوزه‌ی بینایی ماشین (Computer Vision) تشخیص چهره (Face Recognition) رشته‌ی تحقیقاتی است که به ماشین‌ها این امکان را می‌دهد تا بتوانند چهره‌ی افراد را شناسایی کنند. نرم‌افزار تشخیص چهره در بازارهای مصرفی و همچنین صنایع امنیتی و نظارتی کاربردهای بی‌شماری دارد. محققان با کار در این حوزه قصد دارند تا با توسعه‌ی فناوری تشخیص چهره (Face Recognition) زندگی ما را راحت‌تر و تجارت را بهبود دهند.

فناوری تشخیص چهره در حال حاضر برای بهبود پروتکل‌های امنیتی و روش‌های پرداخت در چین استفاده می‌شود و این احتمال وجود دارد که باقی جهان نیز از این روش پیروی کنند. در ادامه‌ی این مطلب از این فناوری درک بهتری و واضح‌تری به دست می‌آورید و درمی‌یابید که تشخیص چهره چیست، چه وظایفی دارد، چطور کار می‌کند و در چه موارد کاربردی دارد.

تشخیص چهره (Face Recognition) چیست؟

تشخیص چهره روشی است برای شناسایی یا تأیید هویت فرد با استفاده از چهره‌ی او در عکس، فیلم یا به‌صورت بلادرنگ.(Real-time)

به‌طور کلی، دو وظیفه‌ی اصلی وجود دارد که مدل‌های تشخیص چهره انجام می‌دهند. اولین وظیفه‌ی آن‌ها تأیید (Verification) است که در آن یک چهره‌ی ورودی جدید با یک هویت شناخته‌شده مقایسه می‌شود. مثالی خوب در این مورد بازکردن قفل تلفن‌های هوشمند با شناسایی چهره است. هنگام راه‌اندازی سیستم تلفن چهره‌ی شما را به‌عنوان مالک تلفن ثبت می‌کند؛ بنابراین تنها کار هنگام بازکردن قفل این است که چهره‌های ورودی جدید را با چهره‌ی ثبت‌شده خود در دستگاه مقایسه کنید.

وظیفه‌ی دوم آن شناسایی (Recognition) یا به‌عبارت دیگر، مقایسه‌ی یک چهره‌ی ورودی با یک پایگاه داده از چندین هویت یا چهره است. این وظیفه اغلب برای سیستم‌های امنیتی و نظارتی استفاده می‌شود. مثال خوب در این مورد تشخیص چهره در اجرای قانون است. در وب‌سایت INTERPOL بخش پزشکی قانونی وجود دارد که توضیح می‌دهد چگونه از شناسایی چهره برای شناسایی افراد مدنظر در فرودگاه‌ها و گذرگاه‌های مرزی استفاده می‌کنند.

تشخیص چهره چطور کار می‌کند؟

دانشمندان داده، به‌دلیل علاقه‌ی زیادی که به این زمینه دارند، هر سال رویکردهای جدیدی برای تشخیص چهره ایجاد می‌کنند.

در این بخش به‌طور خلاصه درباره‌ی مبانی نحوه‌ی کار مدل‌های تشخیص چهره بحث می‌کنیم. به‌طور کلی، مدل‌های تشخیص چهره این مراحل را دنبال می‌کنند:

شناسایی چهره(Face Detection)

دوربین تصویر یک چهره را به‌تنهایی یا در میان جمعیت شناسایی و مکان‌یابی می‌کند. تصویر ممکن است شخصی را نشان دهد که مستقیم به جلو یا به زوایای مختلفی نگاه می‌کند.

تحلیل چهره(Face Analysis)

در مرحله‌ی بعد تصویری از چهره گرفته و تحلیل می‌شود. بیشتر فناوری‌های تشخیص چهره، به‌جای استفاده از تصاویر سه‌بعدی، از تصاویر دوبعدی استفاده می‌کنند؛ زیرا انطباق یک تصویر دوبعدی با عکس‌های یک پایگاه داده راحت‌تر است. درواقع نرم‌افزار هندسه‌ی صورت شما را بررسی می‌کند. ازجمله مواردی که بررسی می‌شود می‌توان به فاصله‌ی چشم، عمق حفره‌های چشم، فاصله‌ی پیشانی تا چانه، فرم استخوان گونه و خط لب، گوش و چانه اشاره کرد. هدف این است که نشانه‌های صورت (Facial Landmarks)شناسایی شود که برای تشخیص چهره کلیدی هستند. در شکل زیر تصویری از ۶۸ نشانه (Landmarks) چهره است که به‌عنوان نقاط کلیدی صورت نیز شناخته می‌شود.

۶۸ نشانه‌ی (Landmark) چهره

تبدیل عکس به داده

در این مرحله اطلاعات آنالوگی (صورت) براساس خصوصیات چهره هر فرد به اطلاعات دیجیتالی (داده) تبدیل می‌شوند. درواقع اطلاعات آنالیزشده چهره به فرمول‌های ریاضی تبدیل می‌شوند. این کدهای عددی اثر چهره (Faceprint) نامیده می‌شوند. دقیقاً مانند اثر انگشت (Fingerprint) که برای هر شخص منحصربه‌فرد است، هر فرد اثر چهره‌ی منحصربه‌فرد خود را دارد.

یافتن عکس منطبق

در این مرحله اثر چهره با باقی چهره‌های موجود در پایگاه داده مقایسه می‌شود. این پایگاه داده شامل تعداد زیادی عکس است؛ برای مثال، اف‌بی‌آی پایگاه داده‌ای با ۶۵۰میلیون عکس دارد یا در فیس‌بوک

(Facebook)هر عکسی که با اسم یک شخص تگ می‌شود جزو پایگاه داده فیس‌بوک محسوب می‌شود که می‌توان از آن برای تشخیص چهره استفاده کرد.

تشخیص چهره کجا استفاده می‌شود؟

مجریان قانون و توسعه‌دهندگان گوشی‌های هوشمند برای بهبود امنیت از تشخیص چهره استفاده می‌کنند. بااین‌حال، این موارد یگانه موارد استفاده‌ی تشخیص چهره نیست. کاربردهای این فناوری بسیار گسترده و متنوع است. نمونه‌های زیر صرفاً چند مورد از جالب‌ترین روش‌هایی است که امروزه در بسیاری از مشاغل از تشخیص چهره استفاده می‌شود.

واقعیت افزوده(AR / Augmented Reality)

بسیاری از برنامه‌های محبوب تلفن‌های هوشمند به تشخیص چهره (Face Recognition) متکی هستند. برخی از نمونه‌های معروف می‌توانند فیلترهای صورت در اینستاگرام(Instagram) ، اسنپ‌چت (Snapchat)و لاین (LINE) باشند. با قراردادن نشانه‌های چهره‌ی کاربر، برنامه AR می‌تواند فیلترهای تصویر را به‌طور دقیق و بلادرنگ روی صورت کاربر قرار دهد.

پرداخت غیرنقدی(Cashless Payment)

گرچه هنوز در اکثر کشورها این امکان در دسترس نیست، فروشگاه‌های زیادی وجود دارند که اکنون پرداخت ازطریق شناسایی چهره در چین را می‌پذیرند؛ علاوه‌براین، در شانزدهم اکتبر ۲۰۱۹ اسنپ‌پی (SnapPay)از راه‌اندازی فناوری پرداخت تشخیص چهره در امریکای شمالی خبر داد.

گیت‌های امنیتی(Security Gates)

یکی دیگر از کاربردهای این فناوری گیت‌های امنیتی است. چه ورودی مجتمع آپارتمانی باشد، لابی جلوی دفتر یا حتی ورودی‌های بلیط ایستگاه قطار، از فناوری تشخیص چهره می‌توان برای اجازه‌دادن یا ندادن ورود استفاده کرد. گرچه این فناوری هنوز در اکثر کشورها رایج نیست، به نظر می‌رسد بسیاری از مشاغل در چین خیلی‌سریع با این فناوری کنار آمده‌اند.

همان‌طور که می‌بینید، کاربردهای مفید بی‌شماری برای تشخیص چهره (Face Recognition) وجود دارد. با افزایش دقت مدل‌ها، کشورهای بیشتری احتمالاً فناوری تشخیص چهره را در زیرساخت‌های خود به کار می‌گیرند.

پردازش تصویر چیست و کاربردهای آن کجاست!

تشخیص چهره روشی برای شناسایی یا تایید هویت فرد، مبتنی بر چهره او می‌باشد. سیستم‌های بازشناسی چهره می‌توانند برای شناسایی افراد در تصاویر، فیلم‌ها و واقعیت استفاده شوند.

بازنشانی چهره در دسته امنیت بیومتریک قرار می‌گیرد. انواع دیگر نرم‌افزارهای بیومتریک شامل شناسایی صدا، اثر انگشت و حدقه چشم می‌باشد. این فناوری بیشتر در موارد قانونی و امنیتی به کار می‌رود. اما در سایر حیطه‌ها نیز روز به روز پرتقاضاتر می‌شود.

بازشناسی یا تشخیص چهره چگونه کار می‌کند؟

اکثر افراد به وسیله FaceID که برای باز کردن قفل گوشی‌های آیفون استفاده می‌شود، با فناوری بازنشانی چهره آشنا هستند. به طور معمول، بازنشانی چهره برای شناسایی هویت فرد، از پایگاه داده عظیمی از تصاویر استفاده نمی‌کند. بلکه تنها مالک دستگاه را شناسایی کرده و مانع دسترسی سایرین می‌شود.

فراتر از باز کردن قفل گوشی‌های هوشمند، بازشناسی چهره توسط انطباق تصویر افرادی که از مقابل یک دوربین می‌گذرند با چهره‌های موجود در یک لیست، کار می‌کند. این لیست می‌تواند شامل تصویر هر کسی باشد، از جمله افرادی که مرتکب جرم خاصی نشده‌اند، و این تصاویر می‌توانند از هر منبعی به پایگاه داده بیایند، حتی از حساب‌های شبکه‌های اجتماعی. فناوری‌های چهره‌ای می‌توانند با هم متفاوت باشند، اما طبق مراحل زیر عمل می‌کنند:

مرحله ۱: یافتن چهره

دوربین چهره را در جمعیت یا تنها، شناسایی و موقعیت‌یابی می‌کند. تصویر نمایان شده از شخص ممکن است به صورت تمام رخ یا نیم رخ باشد.

مرحله ۲: تحلیل چهره

اکنون تصویری از چهره گرفته شده و تحلیل می‌شود. بیشتر فناوری‌های بازنشانی چهره به جای تصاویر سه بعدی، بر تصاویر دو بعدی تکیه می‌کنند. زیرا تطابق دادن تصویر دوبعدی با تصاویر موجود در پایگاه داده یا تصاویر عمومی، آسان‌تر است.

نرم‌افزار مختصات چهره شما را می‌خواند. موارد مهم در اینجا فاصله میان چشم‌ها و عمق آنها، فاصله پیشانی تا چانه، شکل گونه‌ها و خطوط لب‌ها، گوش‌ها و چانه است. در اینجا هدف شناسایی ویژگی‌های خاص یک چهره است که سبب شناسایی آن می‌شود.

مرحله ۳: تبدیل تصویر به داده

فرآیند ثبت تصویر، اطلاعات آنالوگ را به دسته‌ای از اطلاعات دیجیتال (داده)، بر اساس ویژگی‌های چهره شخص تبدیل می‌کند. به عبارتی، نتیجه تحلیل چهره شما به یک فرمول ریاضی تبدیل می‌شود. کد عددی آن "اثر چهره "(faceprint) نام دارد. همانگونه که هر فرد اثر انگشت مختص به خود دارد، اثر چهره او نیز یکتا است.

مرحله ۴: یافتن تطابق

اثر چهره شما با پایگاه داده‌ای از سایر چهره‌ها مقایسه می‌شود. برای مثال، FBIبه بیش از ۶۵۰ میلیون تصویر از پایگاه داده‌های مختلف دسترسی دارد. در فیسبوک، هر تصویر که با نام فرد برچسب گذاری شده است. به جزوی از پایگاه داده فیسبوک تبدیل می‌شود که می‌تواند برای بازنشانی چهره نیز استفاده شود. اگر اثر چهره شما با یکی از تصاویر موجود در پایگاه داده‌های بازنشانی چهره منطبق باشد، تصمیم‌گیری انجام می‌شود.

از میان تمام روش‌های بیومتریکی، بازنشانی چهره، به عنوان طبیعی ترین روش شناخته شده است. این امری طبیعی است. زیرا ما نیز خود و دیگران را با نگاه کردن به چهره‌ها می‌شناسیم و نه اثر انگشت و حدقه چشم. محاسبه شده است که تاکنون نیمی از جمعیت جهان مرتبا با روش بازنشانی چهره سروکار پیدا کرده‌اند.

بازشناسی چهره یا احراز هویت از طریق شناسایی چهره در زمینه‌های مختلفی به کار می‌رود. از جمله این زمینه‌ها می‌توان به موارد زیر اشاره کرد:

۱ (باز کردن قفل گوشی‌های هوشمند

گوشی‌های مختلفی، از جمله سری‌های جدید آیفون از بازنشانی چهره برای باز کردن قفل گوشی هوشمند استفاده می‌کنند. این روشی قدرتمند برای محافظت از اطلاعات شخصی و کسب اطمینان از عدم دسترسی به اطلاعات حساس در صورت سرقت گوشی است. شرکت اپل ادعا دارد که احتمال اینکه قفل گوشی با یک چهره شانسی (که متعلق به مالک نیست) باز شود، یک در میلیون است.

۲ (اجرای قانون

بازنشانی چهره در موارد قانونی کاربرد زیادی دارد. طبق گزارشNBC ، استفاده از این فناوری در نهادهای قانونی در ایالات متحده آمریکایی و سایر کشورها، روز به روز افزایش می‌یابد. پلیس عکس افرادی که دستگیر شده‌اند را گرفته و آن را با پایگاه داده‌های بازنشانی چهره محلی، ایالتی و فدرال مقایسه می‌کند. پس از اینکه از فرد دستگیر شده عکس گرفته شد، این عکس به پایگاه داده اضافه می‌شود تا پلیس بتواند آن را در صورت انجام بررسی های کشف جرم، تحلیل کند.

به علاوه، بازشناسی چهره گوشی‌های هوشمند سبب می‌شود افسران پلیس بتوانند از گوشی، تبلت یا سایر دستگاه‌های قابل حمل از رانندگان یا عابران پیاده عکس گرفته و آن را با یک یا چند تصویر موجود در پایگاه داده بازنشانی چهره مقایسه کنند تا هویت وی را شناسایی کنند.

۳ (فرودگاه‌ها و کنترل مرزها

فناوری بازنشانی چهره به فرودگاه‌های زیادی در سراسر جهان راه پیدا کرده است. تعداد مسافرانی که دارای پاسپورت بیومتریک هستند رو به افزایش است، که سبب می‌شود از صف‌های طولانی رها شده و تنها با گذر از یک درگاه کنترل پاسپورت الکترونیک، سریع تر به گِیت خروج برسند.

بازشناسی چهره نه تنها باعث کاهش زمان انتظار می‌شود، بلکه امنیت فرودگاه‌ها را نیز افزایش می‌دهد. دپارتمان امنیت ایالات متحده پیش‌بینی کرده است که فناوری بازنشانی چهره تا سال ۲۰۲۳، توسط ۹۷ درصد از مسافران مورد استفاده قرار گیرد. علاوه بر فرودگاه‌ها و عبور از مرزها، این فناوری باعث افزایش امنیت رویدادهای بزرگ، مانند المپیک نیز می‌شود.

۴ (یافتن افراد گمشده

بازنشانی چهره می‌تواند برای پیدا کردن افراد گمشده و قربانیان قاچاق انسانی به کار رود. فرض کنید اطلاعات افراد گمشده در یک پایگاه داده نگهداری شوند. در این صورت، به محض بازنشانی چهره آنها

توسط این فناوری، مراجع قانونی می‌توانند مورد اطلاع قرار بگیرند؛ فرقی نمی‌کند که فرد گمشده در فروشگاه، فرودگاه یا هر مکان عمومی دیگری باشد.

۵ (کاهش سرقت از فروشگاه‌ها

بازنشانی چهره می‌تواند برای شناسایی ورود سارقان یا افراد با سابقه سرقت به فروشگاه‌ها استفاده شود. تصویر افراد می‌تواند با تصاویر موجود در پایگاه داده تطابق یافته و در صورت وجود فردی با امکان خطر احتمالی در یک فروشگاه، مسئولان از آن باخبر شوند.

۶ (بهبود تجربه خرید

این فناوری می‌تواند سبب بهبود تجربه خرید برای مشتریان نیز بشود. برای مثال، کیوسک‌های فروشگاه‌ها می‌توانند صف بندی انجام دهند، بر اساس سابقه خرید مشتریان به آنان کالاهایی را پیشنهاد دهند و آنان را به مسیر درست راهنمایی کنند. فناوری "پرداخت چهره‌ای" سبب می‌شود مشتریان مجبور به ایستادن در صف‌های طولانی صندوق و پرداخت هزینه با روش‌های کم سرعت نباشند.

۷ (بانکداری

بانکداری آنلاین بیومتریک، از مزایای دیگر بازشناسی چهره است. به جای استفاده از رمز عبور یک بار مصرف، کاربران می‌توانند با نگاه کردن به گوشی هوشمند یا رایانه خود، یک تراکنش را ثبت کنند.

با این روش، دیگر رمز عبوری برای هکرها وجود نخواهد داشت که به آن دسترسی یابند.

Languages

Python

54.9%

Jupyter Notebook

44.9%

Other

0.2%

وابستگی ها

پایتون 3.5

Numpy 1.15

تنسورفلو 1.8

LibROSA 0.6

FFmpeg 4.0

PyWorld

Database

پایگاه داده های حرکتی تعاملی(IEMOCAP) یک پایگاه داده چندوجهی و چندوجهی است که اخیرا در آزمایشگاه SAIL در USC جمع آوری شده است.این پایگاه داده شامل تقریبا ۱۲ ساعت داده صوتی تصویری، شامل ویدئو، گفتار، ضبط حرکت چهره، رونویسی متن است.پایگاه داده IEMOCAP توسط چندین حاشیه ساز در برچسب های قطعی، مانند خشم، شادی، غم، بی طرفی، و همچنین برچسب های ابعادی مانند ظرفیت، فعال سازی و تسلط نشان داده می شود.

اطلاعات دقیق ثبت حرکت، محیط تعاملی برای استخراج احساسات معتبر، و اندازه پایگاه داده، این پیکره را به یک افزودن ارزشمند به پایگاه داده های موجود در جامعه برای مطالعه و مدل سازی ارتباطات انسانی چند وجهی و بیانی تبدیل می کند.

نتیجه گیری

ما یک روش تبدیل گفتار احساسی بی نظیر مبتنی بر اتوکدرهای انتقال سبک پیشنهاد کردیم.

نیازی به داده های جفت شده، رونوشت ها یا هم ترازی زمانی نیست.

تا جایی که می دانیم، این اولین کار در زمینه تبدیل احساسات بی نظیر است.

با استفاده از انتقال سبک، کاره ای آینده شامل مدت زمان آوایی می شود.

تبدیل و طراحی یک مدل کلی برای بلندگوهای نامرئی

رویکردهای آموزشی بی نظیر

داده موازی به معنای جملاتی با محتوای زبانی یک سان است.

از آنجا که داده های موازی سخت هستند،

برای جمع آوری، روش های بی نظیری توسعه داده شده است.

ایده ها را از ترجمه تصویر به تصویر قرض بگیرید و خلق کنید.

مدل های GAN مناسب برای گفتار، مانندVC-VAW-GAN

VC-StarGAN ،VC-CycleGAN ،VC-StarGAN

روند دیگر مبتنی بر مدل های خودرگرسیو مانند

WaveNetاگرچه می تواند به طور مستقیم و بدون استخراج ویژگی ها، بار محاسباتی سنگین و حجم زیاد آموزش ببیند.

میزان داده های آموزشی مورد نیاز برای اکثر کاربران مقرون به صرفه نیست.

۲٫۳ .یادگیری نمایش درهم تنیده

کار ما از مطالعات اخیر در سبک تصویر الهام می گیرد.

انتقال: یک ایده اساسی پیدا کردن نمایش های درهم تنیده است که می توان به طور مستقل محتوا و سبک تصویر را مدل سازی کرد.

که یک شبکه عصبی کانولوشن(CNN) یک ایده آل است.

بازنمایی برای فاکتوربندی محتوای معنایی و سبک هنری آن ها روشی برای جداسازی و ترکیب مجدد محتوا معرفی کردند.

و سبک تصاویر طبیعی با تطبیق همبستگی ویژگی ها درلایه های کانولوشن مختلف برای ما، وظیفه یافتن بازنمایی های گسسته برای سیگنال گفتار است که بتواند احساسات را تقسیم کند.

از هویت گوینده و محتوای زبانی

روش ها

پژوهش در مورد بیان احساسات و ادراک انسان

دو نتیجه گیری عمده: نخست، درک احساسات انسان است.

فرآیند چند لایه.

مدل سه لایه و یادگیری اتصالات توسط یک سیستم استنتاج فازی. برخی از محققان دریافتند که اضافه کردن لایه های میانی

بر این اساس می توان دقت تشخیص احساسات را بهبود بخشید.

استفاده از پرسپترون های چندلایه را پیشنهاد می کنیم.

استخراج اطلاعات مرتبط با احساسات در سیگنال های گفتاری

دوم، فرآیند تولید احساسات گفتار انسان

جهت مخالف ادراک احساسات را دنبال می کند.

یعنی فرآیند رمزگذاری گوینده، عملیات معکوس فرآیند رمزگشایی شنونده است.

تولید و ادراک گفتاری احساسی، روش بازنمایی یکسانی دارند، یعنی کدکننده و رمزگشا عبارتند از:

عملیات معکوس با ساختارهای آینه ای دسته بندی های احساسی مختلف: هدف ما یادگیری یک نقشه است.

مدل

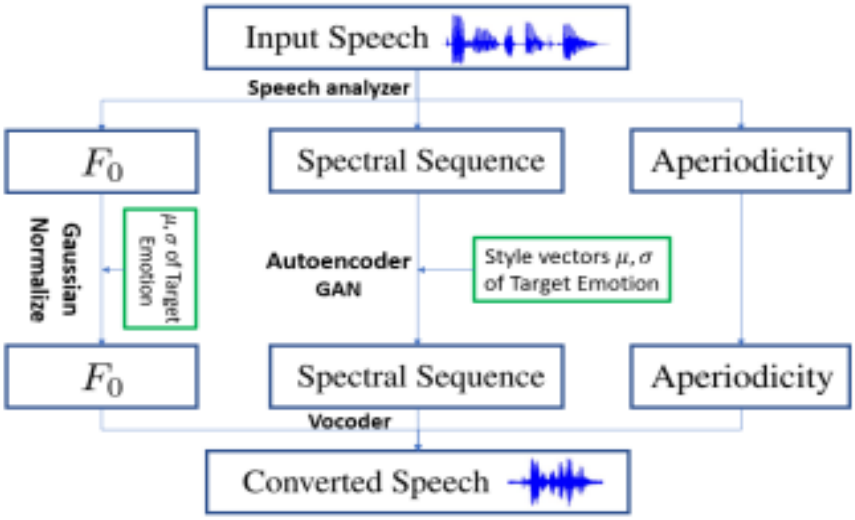
مدل مولد گفتار را به صورت جزئی نشان می دهد.

فرض کنیم یک کد نهفته مشترک و مرتبط با احساسات داشته باشیم.

دارای یک رمزگشا قطعی و معکوس آن است.

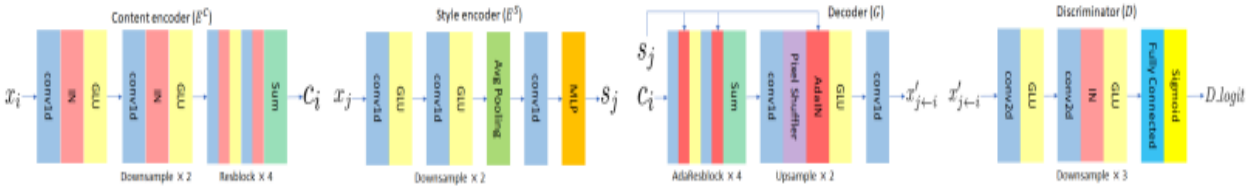
برای تبدیل احساسات، ما فقط کد منبع را استخراج و دوباره کامپایل کنید

شکل : مروری بر سیستم تبدیل احساسات غیر موازی



برای توالی طیفی، ما از نمایش کم - بعدی در دامنه mel-cepstrum برای کاهش پیچیدگی استفاده می کنیم.

این کدکننده ها و کدگشاها با سی ان ان دروازه ای پیاده سازی شده اند. علاوه بر این، یک ماژول GAN توسط بهینه سازی قوی برای تولید چارچوب های طیفی واقع گرایانه اضافه و آموزش داده می شود.



ساختار شبکه ای کدکننده محتوا، کدکننده سبک، کدکننده و تفکیک کننده.GAN

با اطلاعات زبانی، بخشی تغییرناپذیر در محیط ما مدل از نظر احساسات کمی بهتر از VC-StarGAN است.

توانایی تبدیل (۴۸ % در مقابل ۴۴ %)و شباهت بلندگو یک دلیل این است که VC-StarGAN برای تبدیل صدا در میان بلندگوهای مختلف طراحی شده است، در حالی که ما

علاوه بر این، VC-StarGANکیفیت صدای ضعیفی در جهت ساد ۲نگ (۱٫۷۱)و ساد ۲ هاپ (۱٫۸۱)دارد.

تمام احساسات با هم آموزش داده می شوند، بنابراین نسبت به حوزه غم انگیز ناعادلانه است زیرا که نسبت سیگنال به نویز کم تر است و ممکن است نویز را تقویت کند وقتی

به احساسات پرانرژی تر تبدیل می شد.

تعریف انکودر

انکودر (Encoder) در واقع یک سنسور اتوماسیون است، سنسوری که در آن حرکات دورانی یا خطی را برای ما به‌صورت دیجیتالی رمزنگاری می‌کند تا بتوان حالات حرکت (موقعیت، مسیر، سرعت و شمارش) را برای دستگاه‌های کنترلی نظیر PLC قابل‌فهم نماید. در ادامه دستگاه کنترل‌کننده از این سیگنال برگشتی استفاده کرده و عکس‌العمل موردنظر سیستم را تعیین می‌کند به همین دلیل در برخی منابع به آن "رمزگذار" نیز گفته می‌شود.

به نرم افزار یا سخت افزاری که عمل فشرده سازی را انجام می‌دهند انکودر Encoder و به نرم افزار یا سخت افزاری که ویدئو فشرده شده را از فشرده سازی خارج نماید دیکودر Decoder گویند. این انکودر و دیکودر می‌تواند یک نرم افزار ساده و رایگان باشد و یا یک دستگاه مخصوص که در رکهای شبکه نصب شده و دارای قیمت بالا باشد. البته این فشرده سازی برای صدا هم انجام میپذیرد که در اکثر مواقع این امر در دستگاه انکدر به صورت هم زمان برای ویدئو و صدا صورت می‌گیرد. البته لازم به ذکر است که فشرده ساز صدا نیز به تنهایی وجود دارد.

کدبردار) به انگلیسی (Decoder :دستگاه، مدار، مبدل، نرم‌افزار، الگوریتم یا شخصی است که پیام/اطلاعات کدگذاری شده توسط کدگذار را به حالت اولش باز می‌گرداند به طوری که اطلاعات اصلی را می‌توان بازیابی کرد.

در مدار منطقی رمزگشا یا دیکدر مداری است که دارای n پایه ورودی و حداکثر ۲ nپایه خروجی است که بسته به ترکیب سیگنال‌های ورودی، در هر لحظه تنها یکی از ۲ nپایه فعال می‌شود.

دیکدرها دارای انواع دو به چهار، سه به هشت، چهار به شانزده و ... هستند. کاربرد اصلی دیکودر در مدارهای دیجیتال، دسترسی به خانه‌های حافظه است. مدار دیکدر می‌تواند شامل یک سیگنال فعال ساز (En)باشد. اگر سیگنال فعال‌ساز وجود نداشته باشد، مدار دیکودر غیرفعال خواهد شد و عمل نخواهد کرد. دیکدرها را می‌توان به یکدیگر متصل کرد تا یک دیکدر بزرگتر حاصل شود .

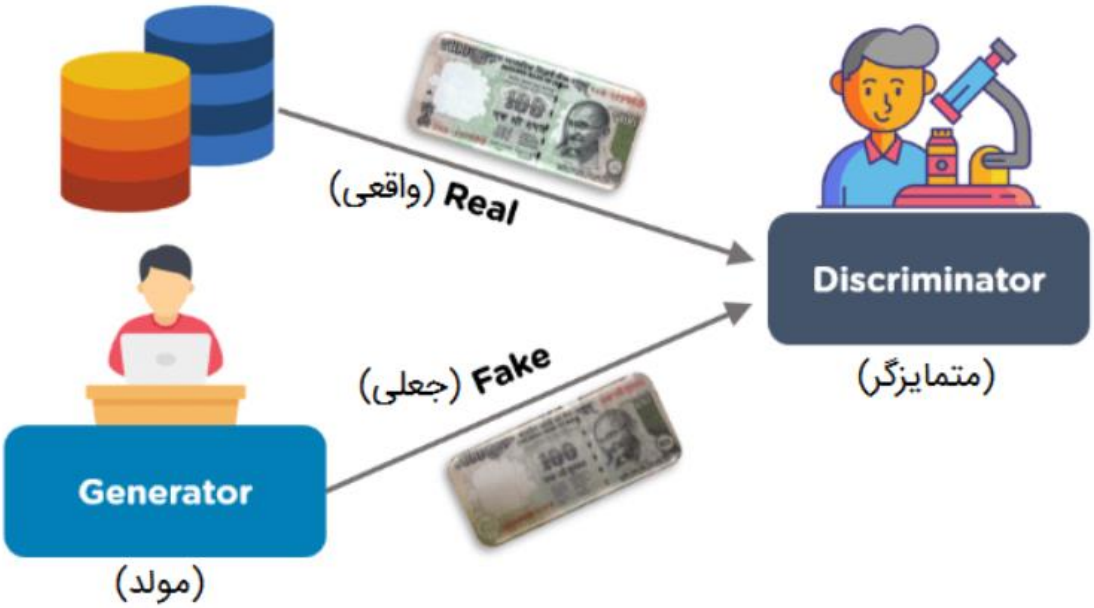
شبکه عصبی GAN چیست؟

شبکه عصبیGAN ، یکی دیگر از شبکه‌های عصبی مشهور در یادگیری ماشین است که عمر آن به ده سال نمی‌رسد. شبکه عصبی GAN در سال 2014 توسط Ian Goodfellow و همکارانش پیشنهاد شد (لینک مقاله). شبکه های عصبی GAN مدل‌های مولدی (Generative Models) هستند که داده‌های جدید شبیه داده‌های آموزشی تولید می‌کنند.

شبکه های عصبی GAN می‌توانند تصاویری مانند چهره انسان تولید کنند که کاملا ساختگی هستند. چهره‌هایی که ممکن است در دنیای واقعی وجود نداشته باشند.

مولد و متمایزگر، برای بررسی، ضبط و تکرار تغییرات درون مجموعه داده با یکدیگر رقابت می‌کنند. می‌توان ازGAN ها برای تولید نمونه‌های جدیدی که به طرز قابل قبولی از مجموعه داده اصلی قابل تهیه هستند، استفاده کرد.

در شکل زیر، نمونه‌ای از GAN نشان داده شده است. یک پایگاه داده حاوی اسکناس‌های واقعی 100 روپیه‌ای وجود دارد. شبکه عصبی مولد، اسکناس‌های جعلی 100 روپیه‌ای را تولید می‌کند. شبکه متمایزگر، به شناسایی اسکناس‌های واقعی و جعلی کمک می‌کند.



مولد چیست؟

یک شبکه عصبی است که داده‌های جعلی تولید می‌کند تا متمایزگر توسط آن‌ها آموزش ببیند. مولد یاد می‌گیرد که داده‌های قابل قبول تولید کند. مثال ها/نمونه‌های تولید شده، برای متمایزگر، نمونه‌های منفی آموزشی به حساب می‌آیند. مولد، یک بردار نویز تصادفی با طول ثابت را به عنوان ورودی می‌گیرد و یک نمونه تولید می‌کند.

generator چیست؟

هدف اصلی مولد این است که متمایزگر را طوری فریب دهد که خروجی خود را با عنوان "واقعی" دسته بندی کند. قسمتی از GAN که مولد را آموزش می‌دهد شامل موارد زیر است:

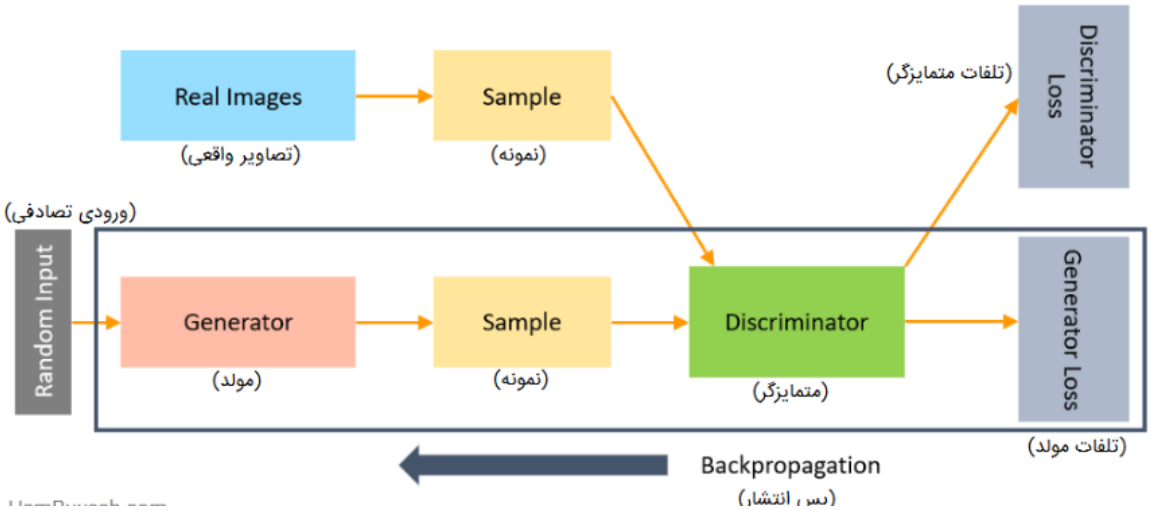
بردار ورودی نویزی

شبکه مولد، که ورودی تصادفی را به یک نمونه داده تبدیل می‌کند

شبکه متمایزگر، که داده‌های تولید شده را دسته بندی می‌کند

تلفات مولد، که مولد را به دلیل احمق پنداشتن متمایزگر، مجازات می‌کند!

از روش پس انتشار (backpropagation) برای تنظیم هر وزن در جهت مناسب با محاسبه تاثیر وزن بر خروجی استفاده می‌شود. همچنین از این روش برای به دست آوردن گرادیان استفاده می‌شود و این گرادیان‌ها می‌توانند به تغییر وزن‌های مولد کمک کنند.



رمزگذاران و رمزگشایان با D-CNNs 1پیاده سازی می شوند تا ثبت وابستگی های زمانی، در حالی که تفکیک کننده های GAN با D-CNN 2ها برای ثبت الگوهای طیفی اجرا می شوند.

همه شبکه ها از گیت استفاده می کنند.

واحدهای خطی(GLU) برای پی گیری اطلاعات ترتیبی.

شبکه عصبی کانولوشن (CNN) چه کار متفاوتی انجام می‌دهد؟

شبکه عصبی کانولوشن نوع خاصی از شبکه عصبی با چندین لایه است. داده‌هایی را که دارای آرایش شبکه‌ای هستند پردازش می‌کند و سپس ویژگی‌های مهم را استخراج می‌کند .

شبکه‌های عصبی کانولوشنال بر اساس یافته‌های علوم اعصاب است. آن‌ها از لایه‌های نورون مصنوعی به نام گره ساخته شده‌اند. این گره‌ها توابعی هستند که مجموع وزنی ورودی‌ها را محاسبه می‌کنند و یک نقشه فعال‌سازی را برمی‌گردانند. این قسمت تجمع شبکه عصبی است.

هر گره در یک لایه با مقادیر وزنی آن تعریف می‌شود. وقتی به لایه، برخی از داده‌ها را می‌دهید، مانند تصویر، مقادیر پیکسل را می‌گیرد و برخی از ویژگی‌های بصری را انتخاب می‌کند.

هنگامی‌که با داده‌های CNN کار می‌کنید، هر لایه نقشه فعال‌سازی را برمی‌گرداند. این نقشه‌ها به ویژگی‌های مهم مجموعه داده اشاره دارند. اگر به CNN تصویری داده باشید، به ویژگی‌های مبتنی بر مقادیر پیکسل مانند رنگ‌ها اشاره می‌کند و عملکرد فعال‌سازی را به شما می‌دهد.

معمولاً با تصاویر، CNNدر ابتدا لبه‌های تصویر را پیدا می‌کند. سپس این تعریف جزئی از تصویر به لایه بعدی منتقل می‌کند. سپس آن لایه شروع به شناسایی مواردی مانند گوشه‌ها و گروه‌های رنگی می‌کند. سپس این تعریف تصویر به لایه بعدی منتقل می‌شود و چرخه تا پیش‌بینی ادامه می‌یابد.

هنگامی‌که بیشتر لایه‌ها تعریف می‌شوند، به این حداکثر تجمع می‌گویند. این فقط مرتبطترین ویژگی‌ها را از لایه موجود در نقشه فعال‌سازی برمی‌گرداند. این همان چیزی است که به هر لایه پی‌درپی منتقل می‌شود تا زمانی که لایه نهایی را به دست آورید.

انواع مختلفCNN

مدل :CNN 1D با این‌ها هسته CNN در یک‌جهت حرکت می‌کند CNN .های 1 Dمعمولاً روی داده‌های سری زمانی استفاده می‌شوند.

مدل 2 :CNN Dاین نوع هسته‌های CNN در دو جهت حرکت می‌کنند. این موارد را با برچسب‌گذاری و پردازش تصویر مشاهده خواهید کرد.

مدل 3 :CNN Dاین نوع CNN دارای هسته‌ای است که در سه جهت حرکت می‌کند. با استفاده از این نوعCNN ، محققان از آن‌ها در تصاویر سه‌بعدی مانند سی‌تی‌اسکن و MRI استفاده می‌کنند.

در بیشتر موارد، CNNهای دوبعدی را مشاهده خواهید کرد زیرا معمولاً با داده‌های تصویر مرتبط هستند. در اینجا برخی از برنامه‌هایی که ممکن است CNN مورد استفاده را مشاهده کنید، آورده شده است.

تشخیص تصاویر با پیش‌پردازش کم

تشخیص دست نوشته‌های مختلف

برنامه‌های دید رایانه‌ای

استفاده در بانکداری برای خواندن رقم چک

استفاده در سرویس‌های پستی برای خواندن کد پستی روی پاکت نامه

نمونه‌ای از CNN در پایتون

به‌عنوان نمونه‌ای از استفاده از CNN در مورد یک مشکل واقعی، قصد داریم برخی از اعداد دست‌نویس را با استفاده از مجموعه داده‌های MNIST شناسایی کنیم.

تشخیص گفتار اولین مرحله‌ی تبدیل گفتار به متن

تشخیص گفتار، توانایی دستگاه در شناخت کلمات و عبارات بیان‌شده و تبدیل اون‌ها به قالب قابل‌فهم توسط ماشینه. در سیستم‌های تشخیص گفتار، چند عامل اهمیت دارن:

گوینده: صدای گوینده‌ها متفاوته. هر مدلی یا باید برای یک گوینده خاص طراحی بشه یا طوری باشه که با صدای هر گوینده‌ای خودش رو تطبیق بده.

نحوه‌ی بیان واژه‌ها: نحوه‌ی صحبت گوینده هم در تشخیص گفتار نقش داره. بعضی از مدل‌ها می‌تونن گفته‌های پیوسته یا گفته‌های ناپیوسته رو با مکثی که در این بین وجود داره، تشخیص بدن.

واژه‌ها: اندازه‌ی واژه‌ها در تعیین پیچیدگی، عملکرد و دقت سیستم نقش مهمی داره.

مدل تشخیص گفتار پایه

برای تبدیل گفتار به متن، از مدل‌های DTW و HMM به‌همراه مدل‌های مختلف شبکه‌ی عصبی استفاده میشه. این مدل‌ها با طبقه‌بندی واج‌ها، تشخیص کلمات و تشخیص صدای گوینده به‌خوبی کار می‌کنن. نقش شبکه عصبی در تکامل هوش مصنوعی، بسیار مهمه. هر سیستم تشخیص گفتار، مراحل استانداردی مثل استخراج ویژگی، تولید مدل و دسته‌بندی الگو رو طی می‌کنه.

شبکه -های مولد متخاصمGenerative Adversarial Networks –

شبکه -های مولد متخاصم با استفاده از معماری شبکه ەhmای عصبی کانولوشنی قادرند تا از مجموعه ای از تصاویر (در اینجا، دیتاست) یاد بگیرد و تصاویری مشابه تصاویر واقعی اما کاملاً جدیدی که در دیتاست موجود نیست را تولید کند. این شبکه برای اولین بار توسط IanGoodfellow معرفی شد.

اجزای اصلی شبکه:GAN

دو جزء اساسی در GAN وجود دارد که سعی در بهبود شبکه برخلاف یکدیگر دارند:

مولد – تولید کننده که با خلق تصاویر بسیار نویزدار از دیتای ورودی که اغلب بصورت نویز گوسی به شبکه داده می-شود کار خود را شروع می -نماید. وظیفه ه-ای که مولد باید در ادامه انجام دهد این است که تصاویری، تا حد ممکن حقیقی تولید کند که به اندازه-ی کافی طبیعی جلوه کنند.

تمیز دهنده- تشخیص دهنده که وظیفه آن، تشخیص تصاویر حقیقی از تصاویر جعلی است بدین صورت که با نگاه کردن به تصاویر تولیدی از مولد باید تشخیص دهد که تصاویر به اندازه ی کافی طبیعی جلوه می- کنند یا خیر. این وظیفه را با مقایسه- ی بین تصاویر دیتاست و تصاویر تولید شده توسط مولد انجام می -دهد.

از این پس مولد تولید -کننده را به اختصار G و تمیز دهنده-تشخیص دهنده را نیز D می- نامیم.

مولد – تولید کننده که با خلق تصاویر بسیار نویزدار از دیتای ورودی که اغلب بصورت نویز گوسی به شبکه داده می-شود کار خود را شروع می -نماید. وظیفه ه-ای که مولد باید در ادامه انجام دهد این است که تصاویری، تا حد ممکن حقیقی تولید کند که به اندازه-ی کافی طبیعی جلوه کنند.

تمیز دهنده- تشخیص دهنده که وظیفه آن، تشخیص تصاویر حقیقی از تصاویر جعلی است بدین صورت که با نگاه کردن به تصاویر تولیدی از مولد باید تشخیص دهد که تصاویر به اندازه ی کافی طبیعی جلوه می- کنند یا خیر. این وظیفه را با مقایسه- ی بین تصاویر دیتاست و تصاویر تولید شده توسط مولد انجام می -دهد.

از این پس مولد تولید -کننده را به اختصار G و تمیز دهنده-تشخیص دهنده را نیز D می- نامیم.

بنابراین در GAN هدف این است که:

مقدار خروجی تمیز دهنده وقتی که تصویر از Pdata باشد، کمینه گردد و در غیرِ این صورت بیشینه باشد؛

G((z) D)بیشینه باشد و D -1 (G((z) کمینه باشد؛

مزیت استفاده از شبکه ی عصبی این است که به راحتی می توان مشتق ها را محاسبه نمود و از انتشار به عقب جهت آموزش شبکه و تصحیح وزن دهی و ... استفاده نمود.

در آن واحد تنها یک شبکه مورد آموزش قرار می گیرد، ابتدا یک شبکه) مولد G یا تمیز دهنده ی (D را آموزش می دهیم و بعد از آموزش آن شبکه، وزن شبکه ی آموزش دیده را ثابت نگه می داریم و شبکه ی دیگر را آموزش می دهیم) تمیز دهنده ی D یا مولد.(G

روند آموزش شبکه ی GAN ؛ Pg توزیع تصویر تولید شده توسط مولدG است که با رنگ سبز و توزیع تصویر دیتاست، با رنگ سیاه و همچنین تمیز دهنده D را با رنگ آبی نشان دادیم (a .تصاویر تولیدی توسط مولد به تصاویر دیتاست شباهت چندانی ندارد (b .وزن دهی تمیزدهنده به روز رسانی شده است در حالی که وزن دهی مولد ثابت مانده است (c وزن دهی مولد به روز رسانی شده درحالی که وزن دهی تمیزدهنده ثابت نگه داشته شده است Pg (d .و Pdata بسیار شبیه به هم شدند.

GANدو کاربرد عمده دارد:

تولید تصاویر جدید براساس دیتا های آموزش دیده موجود در دیتاست.

ترمیم تصویر؛ که ممکن است بخشی از تصویر حذف و یا مسدود شده باشد.

در مسئله ی ترمیم تصویر فرض بر این است که تصویری داریم و می خواهیم کمبود و نقایص موجود در تصویر را برطرف کنیم، این کار را با جایگزینی آن با تصویرِ زمینه انجام می دهیم. فرض کنید یک تصویر از یک تعطیلات دوست داشتنی از یک صحنه ی زیبا دارید امّا یک سری افرادی که نمی شناسید نیز در تصویر وجود دارند و باعث از بین رفتن منظره شده اند. برای برطرف کردن این ناهماهنگی در تصویر ممکن است از نرم افزار Photoshop استفاده کنیم. در اینجا دو انتخاب داریم؛ انتخاب اول این است که اگه مشابه تصویر را در دسترس داریم از آن تصویر برای بازسازی بخش مورد نیاز استفاده کنیم. که در اینصورت باید به کل تصویر نگاه کنیم و تصویر متناسب با مفهوم تصویر را برای جایگزینی انتخاب کنیم و یا به عنوان انتخاب دوم اگر مشابه تصویر در دسترس نباشد، تنها راه برای پر کردن قسمت مورد نظر این است که از پیکسل های همسایه برای پرکردن ناحیه¬ی مسدود شده استفاده کنیم و یا اگر بیش از حد دقت داشته باشیم، ممکن است از بخش های مشابه موجود در همان تصویر استفاده کنیم.

روش اول اصطلاحاً روش مبتنی بر درک و روش دوّم اصطلاحاً روش مبتنی بر محتوا نامیده می¬شوند.

در شبکه های مولد متخاصم ورودی شبکه ی مولد، نویز گوسی است و ورودی های شبکه¬ی تمیزدهنده تصاویر تولیدی توسط مولد و تصاویر موجود در دیتاست می باشند. شبکه ی مولد در ابتدا با در اختیار داشتن نویز، تصویری جعلی تولید می نماید. این تصویر به تمیزدهنده می رود. اگر تمیز دهنده تشخیص دهد تصویر تولیدی توسط مولد به میزان کافی حقیقی است، تصویر به خروجی شبکه داده می شود و کار

تمام است. اما اگر تمیزدهنده با مقایسه با تصاویر موجود در دیتاست، تصویر را جعلی تشخیص دهد فیدبک به مولد داده می شود تا وزن های خود را بروزرسانی نماید که در نتیجه مولد تصویری حقیقی تر را تولید می نماید و این پروسه تا جایی ادامه می یابد که شبکه ی تمیزدهنده متوجه جعلی بودن تصویر خروجی از مولد نشود.

انواع شبکه های:Generative Adversarial Networks

-1شبکه های مولد متخاصم وانیلا (شبکه ی اصلی معرفی شده توسط ایان گودفلو) (The Vanilla GAN)

۲ -شبکه های مولد متخاصم عمیق کانولوشنی (Deep Convolutional Generative Adversarial Networks)

استفاده از شبکه های عصبی کانولوشنی استفاده شده در یادگیری بدون ناظر هم در مولد و هم در تمیز دهنده.

برای مثال در شبکه ی توسعه داده شده توسط Nvidia برای تولید تصاویر که در شبکه ی تمیز دهنده از CNNبرای تشخیص تصویر چهره ی حقیقی از غیر حقیقی مورد استفاده قرار گرفته است و در شبکه ی مولد برای تولید تصویر صورت از یک سری دیتای اولیه (در اینجا نویز گوسی) و با استفاده از DeCNNتصویر حقیقی تر و با کیفیت بالاتری تولید گردیده است.

۳ -شبکه ی مولد متخاصم شرطی:(Conditional Generative Adversarial Networks)

می توان به شبکه امر کرد که چه نوع دیتایی تولید نماید. به عنوان مثال دیتاست اعداد ۰ تا ۹ را در نظر بگیرید که هر کدام از آنها در شبکه ی مولد متخاصم معمول شبکه قادر به تولید تصاویر رندوم از اعداد است. اما در این نوع شبکه ما می توانیم با تغذیه ی ورودی C یک شرط برای آن تعریف نماییم تا تنها مورد دلخواهمان را تولید کند.

شبکه عصبیSOM

شبکه‌های عصبی SOM یا Self-Organizing Map که با نام شبکه کوهونن (Kohonen Network) نیز شناخته می‌شوند، یک روش غیرنظارت‌شده (Unsupervised Learning) برای استخراج ویژگی و کاهش ابعاد است که با وجود سادگی، توانایی زیادی از خود نشان داده است.

آموزش شبکه عصبیSOM

در این شبکه، تعدادی نورون با موقعیت اولیه تصادفی انتخاب می‌شوند که این نورون‌ها در یک شبکه منظم به نام Lattice در کنار هم قرار گرفته‌اند. در طول آموزش، نورون‌های شبکه به مکان‌هایی با چگالی بیشتر داده حرکت می‌کنند و فرم نهایی Lattice حاصل می‌شود.

الگوریتم آموزش SOM به‌صورت زیر است:

۱ .همه داده‌ها تک تک وارد شبکه می‌شوند.

۲ .فاصله همه نورون‌ها از بردار ورودی محاسبه می‌شود.

۳ .نزدیک‌ترین نورون به بردار ورودی تعیین و به‌عنوان نورون برنده انتخاب می‌شود.

۴ .موقعیت نورون برنده با استفاده از رابطه زیر به‌روزرسانی می‌شود:

$$W_J^{t+1} = W_J^t + \eta \cdot (x - W_J^t)$$

5.موقعیت نورون‌های موجود در همسایگی نورون برنده با استفاده از رابطه زیر به‌روزرسانی می‌شود

$$W_N^{t+1} = W_N^t + \theta \cdot \eta \cdot (x - W_N^t)$$

به این ترتیب، با تکرار این 5 مرحله، شبکه در نهایت به موقعیت مناسبی برای هر نورون دست می‌یابد.

توجه داشته باشید که در مراحل 3 و 4 فاز رقابت (Competition) و در مرحله 5 فاز همکاری (Cooperation)وجود دارد و وجود این دو عامل، SOMرا خاص‌تر می‌کند.

پیاده سازی شبکه عصبی SOM در پایتون

حال برای پیاده‌سازی SOM وارد محیط برنامه‌نویسی شده و کتابخانه‌های مورد نیاز را فراخوانی می‌کنیم:

```
import numpy as np
import matplotlib.pyplot as plt
```

این دو کتابخانه برای کار با آرایه‌ها، تولید داده، آموزش مدل و رسم نمودار استفاده خواهند شد.

حال Random Seed و Plot Style را تنظیم می‌کنیم:

```
np.random.seed(0)
plt.style.use('ggplot')
```

برای پیاده‌سازی و آموزش مدل، به یک مجموعه داده ساده نیاز داریم. می‌توانیم یک مجموعه داده دوبعدی به شکل دایره توخالی (دونات) ایجاد کنیم.

تعداد داده‌ها، شعاع داخلی و شعاع خارجی دایره را تعیین می‌کنیم:

```
nD = 1000 # Data Size
r1 = 1 # Inner Radius
r2 = 2 # Outer Radius
```

حال یک آرایه برای ذخیره داده‌ها ایجاد می‌کنیم:

```
nX = 2
X = np.zeros((nD, nX)) # Placeholder for Data
```

عدد nX نشان‌دهنده تعداد ویژگی‌های هر داده است که به‌دلیل محدودیت در نمایش آن‌ها، از داده‌های دوبُعدی استفاده می‌کنیم.

حال یک حلقه While ایجاد می‌کنیم و تا زمانی که تعداد داده‌های مورد نیاز تأمین نشده باشد، عملیات را ادامه می‌دهیم:

```
i = 0
while i<nD:
```

حال باید یک داده به‌صورت تصادفی درون مربع اول انتخاب کنیم:

```
i = 0
while i<nD:
x = np.random.uniform(-r2, +r2, nX)
```

حال باید شعاع داده از مرکز مختصات را حساب کنیم و در صورتی که داده دارای شعاعی بین

r

1

و

r

2

بود، آن را به‌عنوان داده جدید به مجموعه داده اضافه کنیم:

```
i = 0
while i<nD:
```

```
x = np.random.uniform(-r2, +r2, nX)
r = np.linalg.norm(x)
if r1 <= r <= r2:
X[i] = x
i += 1
```

به این ترتیب، داده‌های مورد نیاز ایجاد می‌شود.

حال برای مصورسازی می‌توانیم به‌شکل زیر عمل کنیم:

```
plt.scatter(X[:, 0], X[:, 1], s=12, label='Data')
plt.title('Created Dataset')
plt.xlabel('X1')
plt.ylabel('X2')
plt.legend()
plt.show()
```

In [1]:
```
import tensorflow as tf
```

Input Data: .wav -> Pitch contour (f0s), Harmonic spectral envelope (sps), Aperiodic spectral envelope (aps)

In [2]:
```
import numpy as np
import os
import time
import argparse
import librosa
```

In [3]:

```
from utils import *
from ops import *
```

In [4]:

```
import librosa.display
from IPython.display import Audio
# import matplotlib
import matplotlib.pyplot as plt
```

In [5]:

```
%matplotlib inline
# matplotlib.rcParams['figure.figsize'] = (16, 4)
```

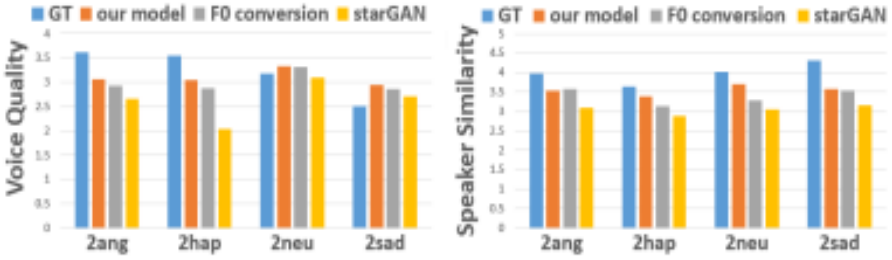
In [6]:

```
random_seed = 0
np.random.seed(random_seed)
```

Autoencoder: Style_Encoder, Content_Encoder, MLP, Decoder, Discriminator

In [7]:

```
def Style_Encoder(inputs, style_dim=16, reuse=False, scope='style_encoder'):

    inputs = tf.transpose(inputs, perm=[0, 2, 1], name='input_transpose')

    with tf.variable_scope(scope, reuse=reuse):

        h1 = conv1d_layer(inputs=inputs, filters=128, kernel_size=15, strides=1, name='h1_conv')
        h1_gates = conv1d_layer(inputs=inputs, filters=128, kernel_size=15, strides=1, name='h1_conv_gates')
        h1_glu = gated_linear_layer(inputs=h1, gates=h1_gates, name='h1_glu')

        # Downsample
        d1 = downsample1d_block_withoutIN(inputs=h1_glu, filters=256, kernel_size=5, strides=2, name_prefix='downsample1d_block1')
        d2 = downsample1d_block_withoutIN(inputs=d1, filters=512, kernel_size=5, strides=2, name_prefix='downsample1d_block2')

        d3 = downsample1d_block_withoutIN(inputs=d2, filters=512, kernel_size=3, strides=2, name_prefix='downsample1d_block3')
        d4 = downsample1d_block_withoutIN(inputs=d3, filters=512, kernel_size=3, strides=2, name_prefix='downsample1d_block4')

        # Global Average Pooling
        p1 = adaptive_avg_pooling(d4)
        style = conv1d_layer(inputs=p1, filters=style_dim, kernel_size=1, strides=1, name='SE_logit')

        return style


def Content_Encoder(inputs, reuse=False, scope='content_encoder'):
    # IN removes the original feature mean and variance that represent important style information
    inputs = tf.transpose(inputs, perm=[0, 2, 1], name='input_transpose')

    with tf.variable_scope(scope, reuse=reuse):

        h1 = conv1d_layer(inputs=inputs, filters=128, kernel_size=15, strides=1, name='h1_conv')
        h1_norm = instance_norm_layer(inputs=h1, name='h1_norm')
        h1_gates = conv1d_layer(inputs=inputs, filters=128, kernel_size=15, strides=1, name='h1_gates')
        h1_norm_gates = instance_norm_layer(inputs=h1_gates, name='h1_norm_gates')
        h1_glu = gated_linear_layer(inputs=h1_norm, gates=h1_norm_gates, name='h1_glu')
```

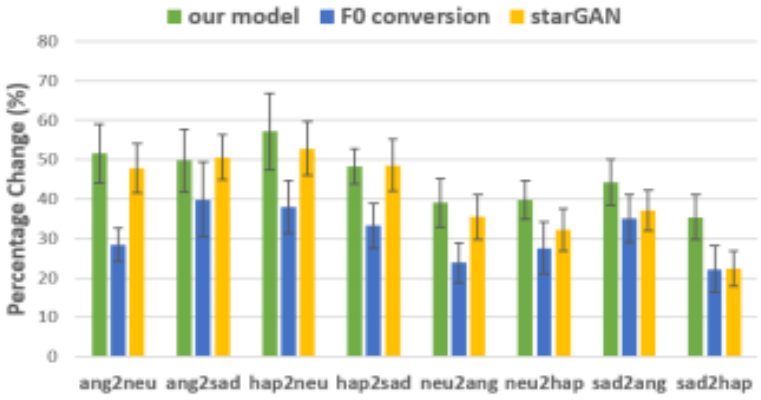


شکل MOS : برای کیفیت صدا و شباهت بلندگو

سمت چپ: کیفیت صدا

راست: تشابه گوینده، دوانگ به معنای هدف است.

احساسات عصبانی کننده است و با گفتار خشمگین اصلی مقایسه می شود.

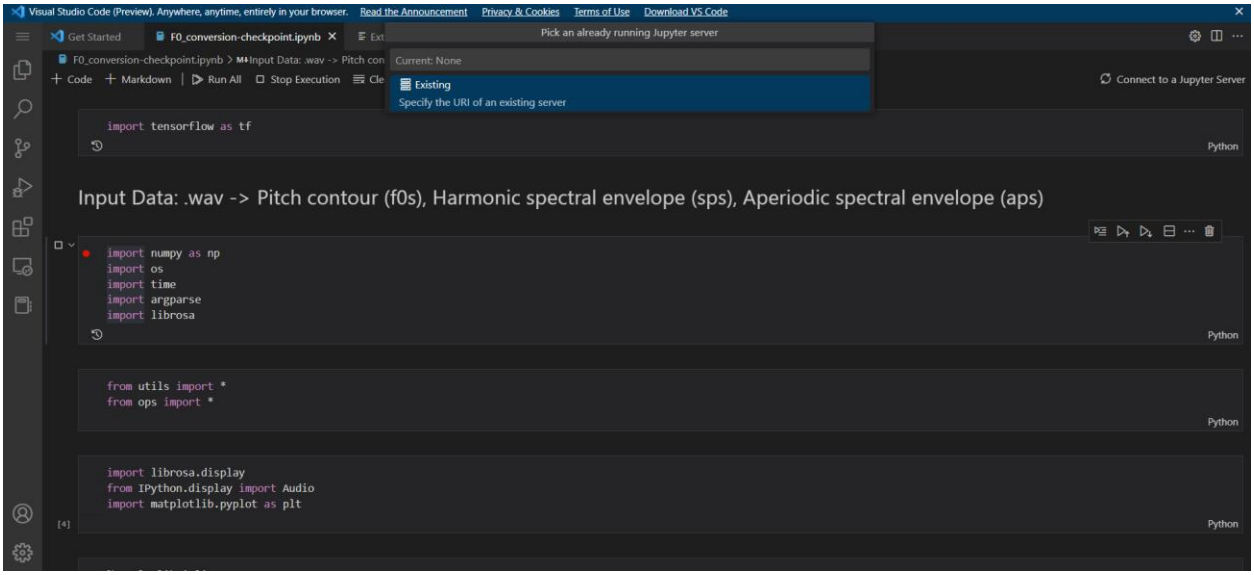


شکل ۶: مقایسه توانایی تبدیل احساسات مدل ما و سیستم های پایه:

F تبدیل(۱)

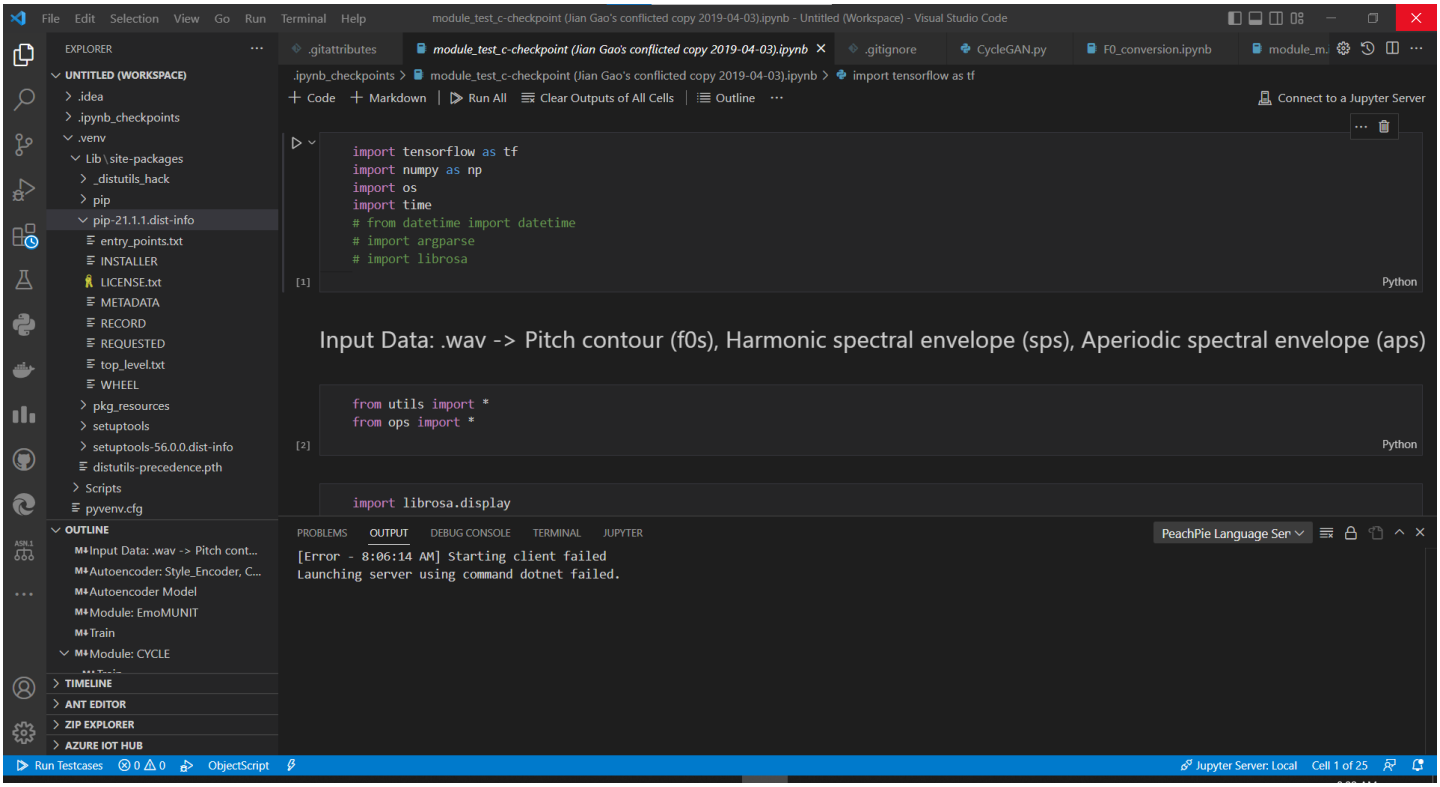
VCstarGAN. (۲)

ang2neu در حال تبدیل شدن از عصبانی به خنثی است.

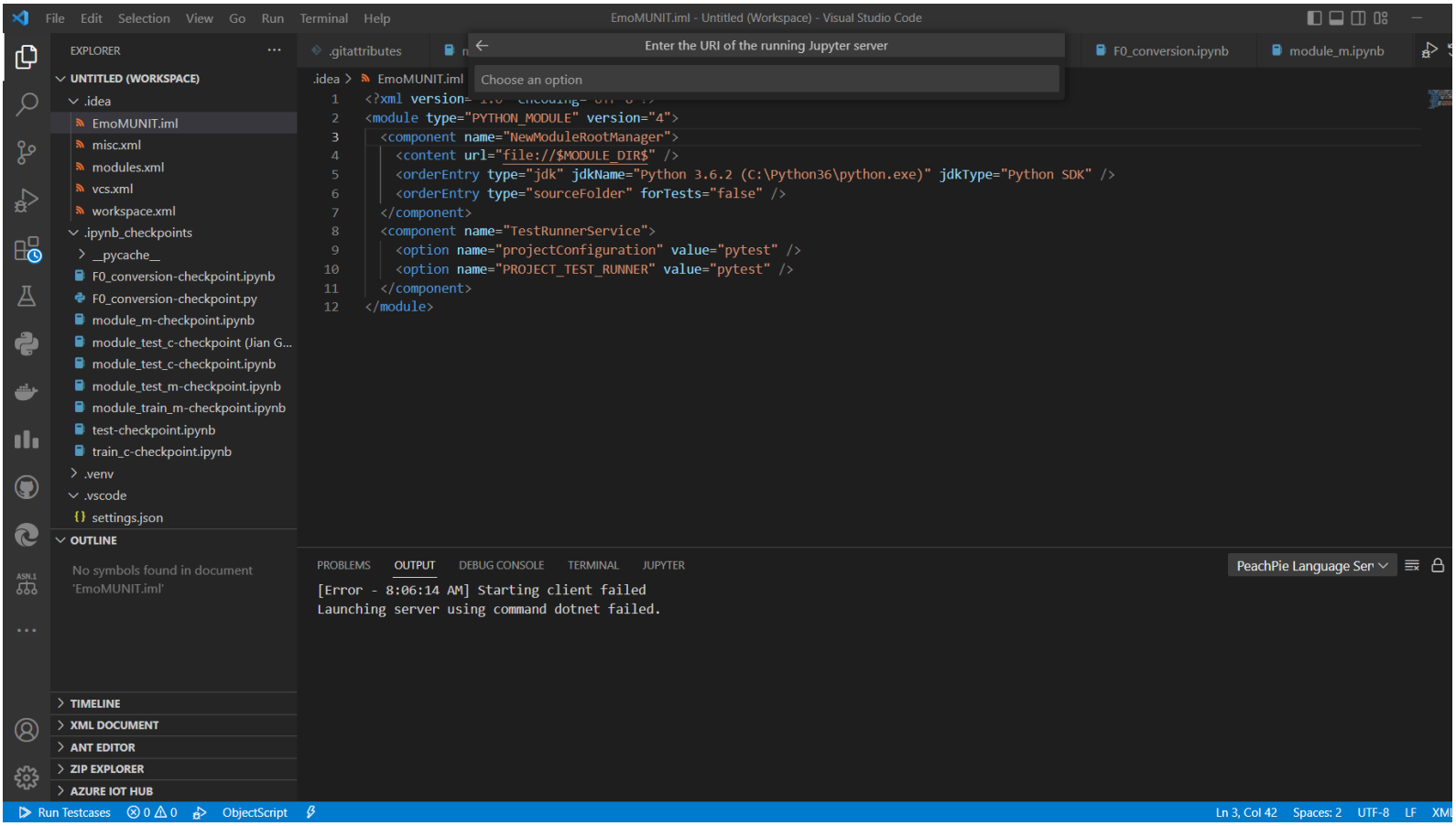


Module: F0

```python
class F0(object):
    def __init__(self, sess, folder='S01/', source='ang', target='neu'):

        self.train_A_dir = './../../../Database/Emotion/' + folder + source + '_' + target + '/' + source
        self.train_B_dir = './../../../Database/Emotion/' + folder + source + '_' + target + '/' + target
        self.validation_A_dir = './../../../Database/Emotion/' + folder + source + '_' + target + '/' + 'val_' + source
        self.validation_B_dir = './../../../Database/Emotion/' + folder + source + '_' + target + '/' + 'val_' + target

        self.audio_len = 128    # = n_frames, time_length
        self.audio_ch = 24      # = num_mcep, num_features

        self.dataset_name = source + '_' + target
        self.model_name = 'C'
        self.gan_type = 'lsgan'
        self.log_dir = "logs/" # + datetime.now().strftime("%Y%m%d-%H%M%S")
        self.sample_dir = 'samples'
        self.checkpoint_dir = 'checkpoint'
        self.A2B_dir = 'F0_results/' + source + '2' + target
        self.B2A_dir = 'F0_results/' + target + '2' + source

        self.sess = sess

        self.sampling_rate = 16000
        self.frame_period = 5.0
        self.num_mcep = 24
```



و موفق به ران شدیم


File Edit Selection View Go Run Terminal Help
EmoMUNIT.iml - Untitled (Workspace) - Visual Studio Code
EXPLORER
UNTITLED (WORKSPACE)
.idea
EmoMUNIT.iml
misc.xml
modules.xml
vcs.xml
workspace.xml
.ipynb_checkpoints
__pycache__
F0_conversion-checkpoint.ipynb
F0_conversion-checkpoint.py
module_m-checkpoint.ipynb
module_test_c-checkpoint (Jian G...
module_test_c-checkpoint.ipynb
module_test_m-checkpoint.ipynb
module_train_m-checkpoint.ipynb
test-checkpoint.ipynb
train_c-checkpoint.ipynb
.venv
.vscode
settings.json
OUTLINE
No symbols found in document 'EmoMUNIT.iml'
TIMELINE
XML DOCUMENT
ANT EDITOR
ZIP EXPLORER
AZURE IOT HUB
.gitattributes
Enter the URI of the running Jupyter server
Choose an option
F0_conversion.ipynb
module_m.ipynb
.idea > EmoMUNIT.iml
<?xml version="1.0" encoding="UTF-8"?>
<module type="PYTHON_MODULE" version="4">
<component name="NewModuleRootManager">
<content url="file://$MODULE_DIR$" />
<orderEntry type="jdk" jdkName="Python 3.6.2 (C:\Python36\python.exe)" jdkType="Python SDK" />
<orderEntry type="sourceFolder" forTests="false" />
</component>
<component name="TestRunnerService">
<option name="projectConfiguration" value="pytest" />
<option name="PROJECT_TEST_RUNNER" value="pytest" />
</component>
</module>
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
PeachPie Language Ser
[Error - 8:06:14 AM] Starting client failed
Launching server using command dotnet failed.
Run Testcases
0 0
ObjectScript
Ln 3, Col 42 Spaces: 2 UTF-8 LF XML