



دانشکده مهندسی برق و کامپیوتر

داده افزایی برای هیجانات (احساسات) در گفتار

استاد راهنما: سید جلال ذهبی

متین فاضل

بهمن ۱۴۰۲

فهرست مطالب

۱	مقدمه	۱
۱	۱.۱ داده افزایی برای هیجانات (احساسات) در گفتار	۱
۱	۲.۱ شرح مسئله	۱
۱	۳.۱ مشارکت ها	۱
۳	۲ مفاهیم و اصطلاحات علمی	۳
۳	۱.۲ شبکه های مولد تخصصی	۳
۳	۱.۱.۲ معماری	۳
۴	۲.۱.۲ تابع هزینه	۴
۶	۳.۱.۲ انواع GAN برای تشخیص هیجانات در گفتار	۶
۶	۴.۱.۲ شبکه های تخصصی حلقوی پایدار	۶
۷	۵.۱.۲ تقسیم بندی با رویکرد یادگیری ماشین یا یادگیری عمیق	۷
۸	۲.۲ سیگنال	۸
۸	۳.۲ سری فوریه	۸
۸	۱.۳.۲ تعریف	۸
۸	Inception Score	۴.۲
۹	۵.۲ فاصله اولیه فریشه	۹
۹	۱.۵.۲ تعریف	۹
۹	۶.۲ شبکه عصبی	۹
۱۰	۷.۲ شبکه عصبی پیچشی	۱۰
۱۳	۸.۲ U-Net	۱۳
۱۵	۳ کار مربوطه	۱۵
۱۵	۱.۳ استخراج ویژگی ها	۱۵
۱۷	۱.۱.۳ Waveplots	۱۷
۱۷	۲.۱.۳ طیف نگاری	۱۷
۱۸	۲.۳ مجموعه داده	۱۸
۱۹	۱.۲.۳ SAVEE	۱۹
۱۹	۲.۲.۳ ESD	۱۹
۱۹	۳.۲.۳ Tess	۱۹
۲۰	۴.۲.۳ CREMA-D	۲۰

۲۰ روش پیشنهادی	۵.۲.۳	
۲۱ انتقال احساسات مبتنی بر شبکه‌های مولد متخاصمی	۳.۳	
۲۱ اجزاء شبکه CycleGan	۴.۳	
۲۲ استفاده از مجموعه داده بزرگ بدون برچسب	۵.۳	
۲۴			۴ نتایج
۲۴ معماری سیستم	۱.۴	

فصل ۱

مقدمه

۱.۱ داده افزایی برای هیجانات (احساسات) در گفتار

تشخیص عواطف داده های صوتی SER به تشخیص خودکار احساسات و عواطف انسانی اشاره دارد. به عنوان زمینه تحقیقاتی مهم، تشخیص عواطف داده های صوتی به سرعت در حال رشد است، همچنین دارای پتانسیل بهبود تعامل انسان و رایانه مبتنی بر صدا، مانند سیستم داخل خودرو برای درک وضعیت عاطفی رانندگان به هنگام ایجاد تغییرات ناگهانی است.

۲.۱ شرح مسئله

کمبود داده از عمده چالش های پیچیده در تشخیص هیجانات در گفتار است که در این سه مورد شرح داده شده است:

- مشکل اول فقدان دیتاست اصوات صوتی طبیعی است. دیتاست های کمی برای تحقیقات در این زمینه به اشتراک گذاشته شده است. مخصوصاً، بسیاری از دیتاست های صوتی که در شرایط واقعی تولید شده اند به علت یکسری محدودیت های قانونی در دسترس عمومی قرار نگرفته اند.
- مشکل دیگر برچسب گذاری داده های صوتی است. از آنجایی که احساسات ابراز شده متفاوت هستند، دسته بندی آنها بسیار مهم است. با این حال به دلیل عدم قطعیت بالا دسته بندی و تحلیل آنها کاری به شدت زمان بر است.
- در نهایت، داده های صوتی در اکثر پایگاه های داده به صورت نامتعادلی بر روی احساسات توزیع میشوند. به طور کلی، تعداد جملات با احساسات خنثی بیشترین تعداد را در بدنه گفتاری دارد. با این حال، برای ارزیابی دقت طبقه بندی، یک پایگاه داده متعادل برای تجزیه و تحلیل نیاز است. علاوه بر این، اگر یک جمله با احساسات مختلف ضبط شود، قضاوت انسان در مورد احساس درک شده میتواند صرفاً بر اساس محتوای عاطفی جمله بدون تاثیر محتوای واژگانی آن باشد.

۳.۱ مشارکت ها

شبکه های مولد متخاصم (GANs)، یک روش کارآمد برای تولید دیتا هستند. با استفاده از یک بازی متخاصم بین یک تشخیص دهنده و یک مولد، شبکه های مولد متخاصم برای تولید نمونه هایی که از داده

های واقعی قابل تشخیص نیستند، آموزش میینند. علاوه بر این دارای مشخصه های زیر هستند:

- شبکه های مولد متخاصم میتوانند توزیع های احتمالی با ابعاد بالا را در مسائل پیچیده دنیای واقعی بیاموزند.
- این شبکه ها را میتوان با داده ها از دست رفته آموزش داد، که برای یادگیری نیمه نظارتی مناسب است.
- شبکه های مولد متخاصم دارای خروجی های چندوجهی هستند، به این معنی که میتوانند چندین پاسخ صحیح مختلف تولید کنند و تنوع نمونه های تولید شده را افزایش دهند.

هدف این پروژه ارزیابی عملکرد SER زمانی است که داده های آموزشی واقعی در فضای ویژگی با داده های مصنوعی تولید شده توسط شبکه های مولد متخاصم افزایش میابد. به طور خاص، نیاز به طراحی یک مدل مبتنی بر GAN است که بردارهای ویژگی مصنوعی از گفته های احساسی مختلف را تولید کند، به طوری که عملکرد یک شبکه عصبی دسته بند با دریافت بردارهای ویژگی واقعی و مصنوعی به عنوان داده های آموزشی بهبود داده شود.

فصل ۲

مفاهیم و اصطلاحات علمی

۱.۲ شبکه های مولد تخصصی

مدل های مولد به هر مدلی اطلاق می شود که مجموعه ای از نمونه های آموزشی گرفته شده از یک توزیع را یاد می گیرد که تخمینی از آن توزیع را نشان دهد. این شبکه ها بر اساس تولید داده به دو دسته صریح^۱ و ضمنی^۲ تقسیم می شوند. مدل های صریح تابع چگالی توزیع را مستقیماً محاسبه میکنند در حالی که مدل های ضمنی بر تولید نمونه هایی از توزیع ارائه شده توسط مدل تمرکز میکنند.

۱.۱.۲ معماری

ک شبکه مولد متخصصی، یک مدل مولد بر بر اساس نظریه بازی است که دو عامل را روبه روی هم قرار میدهد:

- یک تشخیص دهنده D ^۳
- یک مولد G ^۴

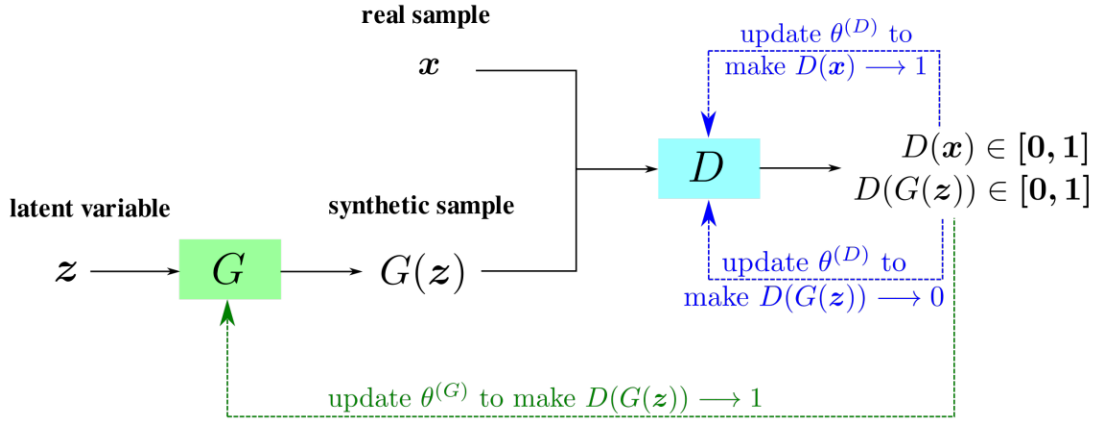
در شکل ۱.۱ طرح کلی یک شبکه مولد متخصصی رسم شده است، که با نام Vanilla GAN شناخته می شود. مولد به عنوان ورودی متغیر پنهان z را دریافت میکند که از یک مجموعه توزیع نویز $p_z(z)$ تهیه میشود و داده ساخته شده $G(z; \theta^G)$ را به عنوان خروجی تولید می کند. هدف مولد این است که یکسری داده تولید کند که از داده های واقعی غیر قابل تشخیص باشند. تشخیص دهنده از یک طرف داده های واقعی و از طرفی دیگر نمونه های مصنوعی $G(z)$ را به عنوان ورودی دریافت میکند. خروجی این مدل $D(x)$ یا $D(G(z))$ که احتمال واقعی بودن نمونه دریافتی را نشان می دهد. مولد تلاش می کند تا مقدار $D(G(z))$ را به عدد یک نزدیک کند تا تشخیص کننده را متقاعد کند که نمونه تولید شده شبیه به نمونه واقعی است، در حالی که تشخیص دهنده در تلاش است که مقدار $D(G(z))$ را صفر نزدیک کند و مقدار $D(x)$ را به یک نزدیک کند.

¹Implicit

²Explicit

³Discriminator

⁴Generator



شکل ۱.۲: ساختار یک شبکه مولد متخاصمی

۲.۱.۲ تابع هزینه

در نظر داریم که به عنوان ورودی شبکه داده‌های $(s_i, y_i)_{i=1}^N$ را دریافت می‌کنیم که نیمی از آن داده‌های واقعی x و نیمی دیگر از آن داده‌های تولید شده $G(z)$ است. هر نمونه آموزشی s_i متناظر با یک برچسب y_i است. همه داده‌های واقعی دارای برچسب یک و تمامی داده‌های واقعی حاوی برچسب صفر می‌باشند. با توجه به اینکه هدف تشخیص دهنده یک دسته‌بند دودویی است، تابع هزینه آن به صورت یک binary cross-entropy تعیین می‌شود.

(۱.۲)

$$J^{(D)}(D, G) = H((s_i, y_i)_{i=1}^N, D) = -\frac{1}{N} \sum_{i=1}^N (y_i \log(D(s_i)) + (1 - y_i) \log(1 - D(s_i)))$$

اگر y_i را برابر با یک برای $s_i = x$ و برابر با صفر برای وقتی که $s_i = G(z)$ در نظر بگیریم، همچنین با جایگذاری میانگین‌ها به رابطه تابع هزینه تشخیص دهنده به عبارت زیر می‌رسیم:

$$J^{(D)}(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [1 - \log D(G(z))] \quad (۲.۲)$$

که p_{data} یک توزیع داده بر روی نمونه داده‌های واقعی x است. در یک بازی minimax که به نام بازی zero-sum نیز شناخته می‌شود، مجموع هزینه‌های تمامی بازیکنان همواره صفر است. که نشان می‌دهد تابع هزینه مولد مخالف $J^{(D)}$ است. در حالی که برای محاسبه تابع هزینه مولد در نزول گرادیانی عبارت دوم در معادله تابع هزینه اهمیت دارد. بنابراین تابع هزینه مولد در یک بازی minimax به صورت زیر تعریف شده است:

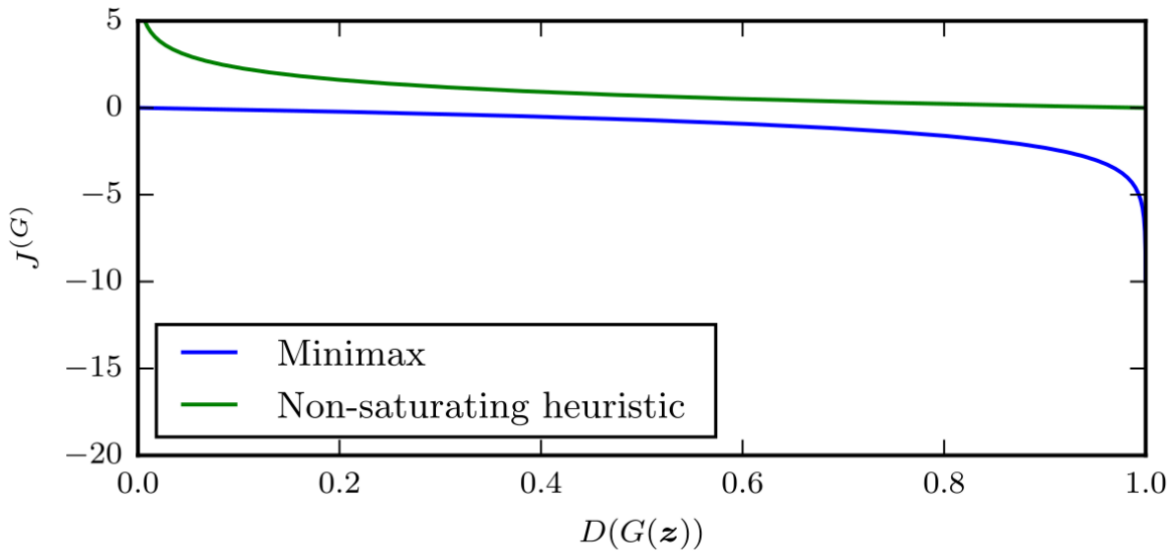
$$J^{(G)}(G) = \mathbb{E}_{z \sim p_z(z)} [1 - \log D(G(z))] \quad (۳.۲)$$

در حالت کلی این بازی در یک تابع ارزش خلاصه می شود که به صورت زیر تعریف می شود:

$$\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [1 - \log D(G(z))] \quad (۴.۲)$$

که مولد سعی در کمینه کردن آن و تشخیص دهنده سعی در بیشینه کردن آن دارد. در صورتی که تابع هزینه مولد در یک بازی minimax در واقعیت به خوبی عمل نمی کند، به دلیل اینکه وقتی مولد سعی در بیشینه کردن مقداری که تشخیص دهنده سعی در کمینه کردن آن را دارد، باعث می شود که تشخیص دهنده به آسانی تمامی داده های تولید شده توسط مولد را مصنوعی تشخیص دهد و آنها را رد کند. در نتیجه، نزول گرادیانی در مولد دچار اختلال می شود. مولد به جای آنکه مقدار تشخیص درست تشخیص دهنده را کمینه کند، مولد در یک بازی non-saturating heuristic سعی در بیشینه کردن اشتباه تشخیص دهنده دارد. و تابع هزینه آن در این بازی به صورت زیر تعریف می شود:

$$-\mathbb{E}_{x \sim p_{data}(x)} [\log D(G(x))] \quad (۵.۲)$$



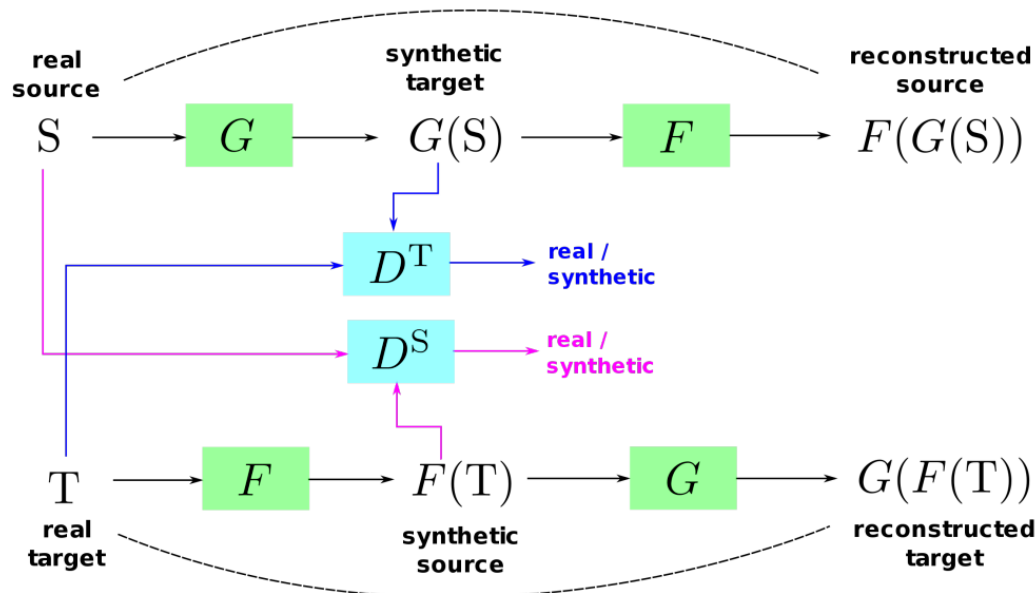
شکل ۲.۲: مقایسه مقدار تابع هزینه مولد در بازی های minimax و non-saturating heuristic

در شکل ۱.۲ اختلاف مقدار تابع هزینه در حالت minimax و non-saturating heuristic قابل مشاهده است. محور افقی احتمال قبول شدن یک نمونه داده مصنوعی به عنوان یک داده واقعی را نشان می دهد. هر چه مقدار این عدد بیشتر باشد، مولد میزان هزینه کمتری را میگیرد. قسمت سمت چپ تابع که میزان $D(G(z))$ نزدیک به صفر است، در ابتدای مرحله آموزش مدل اتفاق می افتد. در این زمان، تشخیص دهنده به راحتی می تواند تشخیص بدهد که داده نمونه متعلق به کدامین کلاس است. زیرا مولد در ابتدا شروع به ساخت داده های مصنوعی با توجه به توزیع تصادفی $p_z(z)$ با یکسری پارامتر رندوم می کند. واضح است که منحنی بازی minimax یک خط مستقیم است، که نشان می دهد مولد دارای گرادیان بسیار کمی است. با استفاده از نزول گرادیانی، مولد روند بهبود مدل را در مراحل ابتدایی متوقف کرده است. در نقطه مقابل، منحنی بازی non-saturating heuristic مقدار گرادیان خودش را در قسمت راست

از دست می‌دهد، و در این نقطه بهینه نمونه داده‌های تولید شده قادر به گمراه کردن تشخیص دهنده هستند. بنابراین، بازی non-saturating معمولاً در دنیای واقعی کاربرد بیشتری دارد و نسبتاً بازی minimax جنبه نظری دارد.

۳.۱.۲ انواع GAN برای تشخیص هیجانات در گفتار

در سال‌های گذشته شبکه‌های مولد تخصصی برای تشخیص هیجانات در گفتار استفاده شده است. برای مثال در یک آزمایش از یک DCGAN برای تحلیل گفتار احساسی به روش نیمه نظارت شده استفاده کردند. در این قسمت ما تمرکز بر استفاده از شبکه‌های مولد متخصصی برای تولید داده داریم، که در نهایت به ما کمک می‌کند که داده‌هایی تولید کنیم که شبیه به توزیع داده‌های واقعی در مسئله داده شده باشد. در این قسمت ما به تحلیل و بررسی سه نوع رایج شبکه‌های مولد متخصصی می‌پردازیم که شامل: adversarial autoencoder, conditional GAN, and CycleGAN است.



شکل ۳.۲: ساختار CycleGAN

۴.۱.۲ شبکه‌های تخصصی حلقوی پایدار

شبکه‌های تخصصی حلقوی پایدار که معروف به CycleGAN است یک روش بسیار موفق برای حل مسائل ترجمه متن به متن یا به کارگیری مجموعه داده‌های جفت نشده است. نداشتن دو طرفه‌ای که توسط شبکه CycleGAN آموخته می‌شود، می‌تواند ویژگی‌های خاص یک مجموعه تصویر را ثبت کند و بفهمد که چگونه می‌توان این ویژگی‌ها را به مجموعه تصویر دیگر ترجمه کند. موفقیت چشم‌گیر این نوع شبکه در انگیزه‌ای برای استفاده در انتقال احساسات است. شکل ۱.۳ ساختار کلی یک CycleGAN را تصویر کشیده است. این شبکه حاوی دو تابع نگاشت G و F است. تابع G یاد می‌گیرد که داده‌های نمونه را

¹Cycle-consistent adversarial networks

از منبع S به دامنه هدف T تبدیل کند. همچنین، تابع نگاشت F نیز یک نگاشت معکوس نسبت به G است. هر دوی این توابع نگاشت G و F را می توان یکسری مولد برای هدف و منبع تولید داده در نظر گرفت. به علاوه، این شبکه دارای دو تشخیص دهنده D^T و D^S است. تشخیص دهنده D^T در مقابل مولد \square وظیفه تشخیص واقعی بودن داده های تولید شده توسط G را برای داده های T دارد. همچنین، D^S تشخیص هم وظیفه تشخیص داده واقعی S را از مجموعه داده تولید شده $F(T)$ دارد. همچنین در ادامه برای اطمینان حاصل کردن از اینکه تصویر ساخته شده قابلیت بازگردانی به داده نمونه اصلی را دارد، این شبکه سعی در ساخت نمونه داده های هدف و منبع را دارد به طوری که $F(G(S))$ باید شبیه به S و $G(F(T))$ باید شبیه به T باشد، که مربوط به قسمت پایداری این شبکه تخصصی است. میزان خطای CycleGAN تشکیل شده از یک خطای تخصصی و خطای حلقوی پایدار است. خطای تخصصی را می توانیم برای دو قسمت S و T در نظر بگیریم. خطای تخصصی به صورت زیر محاسبه می شود:

(۶.۲)

$$\mathcal{L}_{GAN}(G, D^T, S, T) = \mathbb{E}_{t \sim p_t(t)}[1 - \log D^T(t)] + \mathbb{E}_{s \sim p_s(s)}[1 - \log 1 - D^T(G(s))]$$

(۷.۲)

$$\mathcal{L}_{GAN}(F, D^S, T, S) = \mathbb{E}_{s \sim p_s(s)}[1 - \log D^S(s)] + \mathbb{E}_{t \sim p_t(t)}[1 - \log 1 - D^S(G(t))]$$

قابل ذکر است که میزان خطای تخصصی در قالب یک تابع ارزش بیان شده است. بنابراین حذف ما از این خطا $\min_F \max_{D^S} \mathcal{L}_{GAN}(F, D^S, T, S)$ و $\min_G \max_{D^T} \mathcal{L}_{GAN}(G, D^T, S, T)$ است.

با داشتن ظرفیت به اندازه بزرگ، شبکه می تواند همگی تصاویر ورودی را به هر ترتیب تصادفی از تصاویر در دامنه مقصد نگاشت کند. به گونه ای که هر یک از نگاشت های یاد گرفته شده می تواند یک توزیع خروجی را ایجاد کند که با توزیع مقصد همخوانی داشته باشد. بنابراین این شبکه به یک تابع خطای دیگر هم نیاز دارد، که در پایین ذکر شده است:

(۸.۲)

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{t \sim p_t(t)}[\|G(F(t)) - t\|_1] + \mathbb{E}_{s \sim p_s(s)}[1 - \log 1 - D^T(G(s))]$$

در نتیجه به صورت کلی تابع هزینه به این صورت نمایش داده میشود:

(۹.۲)

$$\mathcal{L}(G, F, D^T, D^S) = \mathcal{L}_{GAN}(G, D^T, S, T) + \mathcal{L}_{GAN}(F, D^S, T, S) + \lambda \mathcal{L}_{cyc}(G, F)$$

۵.۱.۲ تقسیم بندی با رویکرد یادگیری ماشین یا یادگیری عمیق

یادگیری ماشینی برای تقسیم خود کار قسمت های مختلف یک تصویر استفاده می شود. معماری های U-Net در حل مسائل تقسیم بندی کارآمد هستند. در ادامه با توجه به اینکه روش استفاده شده مبتنی بر یادگیری عمیق است ابتدا اصطلاحات و تئوری مربوط به آن را بررسی می نمایم.

۲.۲ سیگنال

سیگنال یک تغییر در یک کمیت معین در طول زمان است. برای صوت، مقداری که تغییر می کند فشار هوا است. ما می توانیم از فشار هوا در طول زمان نمونه برداری کنیم. سرعت نمونه برداری از داده ها می تواند متفاوت باشد، اما معمولاً ۱.۴۴ کیلوهرتز یا ۴۴۱۰۰ نمونه در ثانیه است. آنچه ما گرفته ایم یک شکل موج برای سیگنال است و می توان آن را با نرم افزار کامپیوتری تفسیر، اصلاح و تحلیل کرد.

۳.۲ سری فوریه

در ریاضیات، تبدیل فوریه^۱ یک تبدیل ریاضیاتی است که توابعی را که بر حسب زمان یا فضا هستند، به توابعی بر حسب فرکانس زمانی یا فضایی تجزیه می کند، مانند بیان یک آکورد موسیقی بر حسب حجم ها و فرکانس های نت های تشکیل دهنده آن. اصطلاح تبدیل فوریه هم به نمایش دامنه فرکانس و هم به عملیات ریاضی مربوط به آن که نمایش دامنه فرکانس را به تابعی از مکان یا زمان مرتبط می کند گفته می شود. تبدیل فوریه یک تابع از زمان، یک تابع مقدار مختلط از فرکانس است، که اندازه آن (قدر مطلق)، فرکانس موجود در تابع اصلی را نشان می دهد، و آرگومان آن اختلاف فاز سینوسی پایه در آن فرکانس است. تبدیل فوریه فقط محدود به توابع زمان نیست، اما به دامنه عملکرد اصلی، معمولاً دامنه زمان گفته می شود. معکوس تبدیل فوریه نیز وجود دارد که به صورت ریاضی تابع اصلی را از نمایش دامنه فرکانسی آن تولید می کند، که توسط قضیه عکس فوریه اثبات شده است.

۱.۳.۲ تعریف

تبدیل فوریه، نامیده شده به اسم ریاضیدان فرانسوی ژوزف فوریه، یک تبدیل انتگرالی است که هر تابع $f(t)$ را به یک تابع دیگر $F(\omega)$ منعکس می کند. در این صورت، به $F(\omega)$ تبدیل فوریه تابع $f(t)$ می گویند. حالت خاص تبدیل فوریه، سری فوریه نام دارد و آن زمانی کاربرد دارد که تابع $f(t)$ متناوب باشد، یعنی: $f(t+T) = f(t)$. چنانچه تابع متناوب نباشد یا به عبارتی، تناوب آن برابر بی نهایت باشد ($T \rightarrow \infty$)، از سری فوریه عبارت زیر به دست می آید:

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \quad (۱۰.۲)$$

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega \quad (۱۱.۲)$$

۴.۲ Inception Score

امتیاز ادراکی (Inception Score) یک معیار ارزیابی است که برای اندازه گیری کیفیت تصاویر تولید شده توسط شبکه های مولد و دسته بند نورو نی مصنوعی مانند شبکه های مولد متخاصمی استفاده می شود. این معیار توسط شبکه عصبی Inception که از مدل Inception v3 گرفته شده است، محاسبه می شود. Inception Score برای ارزیابی دو ویژگی اساسی در تصاویر تولید شده استفاده می شود:

¹Fourier transform

- تنوع (Diversity): معیار Inception Score سعی در اندازه گیری تنوع تصاویر تولید شده دارد. تصاویر با امتیاز بالاتر این دارایی را دارای تنوع بیشتری می داند که نشان از پوشش گسترده تر فضای تصاویر دارند.
- کیفیت (Quality): این معیار به دنبال اندازه گیری کیفیت تصاویر است. تصاویر با امتیاز بالاتر در این ویژگی، به عنوان تصاویر با کیفیت بالا تشخیص داده می شوند که نشان دهنده واقعیت و شباهت به داده های واقعی است.

۵.۲ فاصله اولیه فریشه

فاصله اولیه فریشه^۱ (FID) معیاری است که برای ارزیابی کیفیت تصاویر ایجاد شده توسط یک مدل مولد، مانند یک شبکه متخاصم مولد استفاده می شود. بر خلاف معیار قدیمی تر امتیاز اولیه (IS)، که فقط توزیع تصاویر تولید شده را ارزیابی می کند، فاصله اولیه فریشه توزیع تصاویر تولید شده را با توزیع مجموعه ای از تصاویر واقعی مقایسه می کند.

۱.۵.۲ تعریف

برای هر دو توزیع احتمالی u, v بر روی \mathbb{R}^n که دارای میانگین و انحراف از معیار متناهی هستند، فاصله اولیه فریشه به صورت زیر تعریف می شود:

$$d_F(u, v) := \sqrt{\inf_{\gamma \in \Gamma(u, v)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 d\gamma(x, y)}, \quad (12.2)$$

به صورتی که $\Gamma(u, v)$ شامل تمامی مقادیر بر روی $\mathbb{R}^n \times \mathbb{R}^n$ است، با مقادیر حاشیه ای u, v به ترتیب بر روی عوامل اول و دوم. برای دو توزیع گاوسی چند بعدی، $N(\mu_1, \Sigma_1)N(\mu_1, \Sigma_1)$ این رابطه به صورت زیر است:

$$d_F(N(\mu_1, \Sigma_1), N(\mu_1, \Sigma_1))^2 = \|\mu_1 - \mu_2\|^2 + Tr\left(\Sigma_1 + \Sigma_2 - 2\sqrt{\Sigma_1 \Sigma_2}\right) \quad (13.2)$$

۶.۲ شبکه عصبی

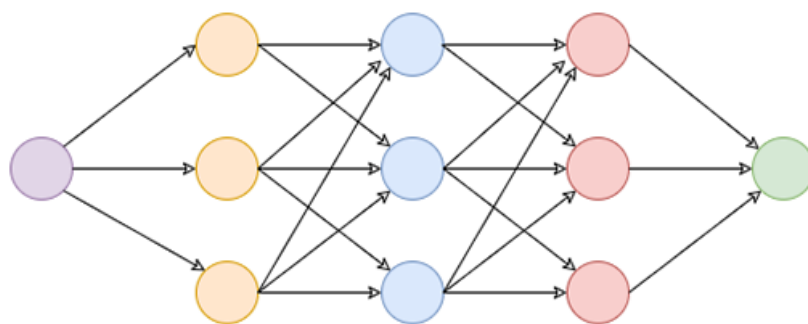
شبکه های عصبی^۲ از مغز ما الگو گرفته اند. نودهایی^۳ وجود دارند که لایه ها^۴ را در شبکه تشکیل می دهند و دقیقاً مانند نورون های مغز ما، نواحی مختلف را به هم متصل می کنند. به ورودی های نودها در یک لایه، وزنی اختصاص می یابد که تأثیری را که پارامتر بر نتیجه پیش بینی کلی دارد، تغییر می دهد. از آنجا که وزن ها به پیوندهای بین نودها اختصاص داده می شوند، ممکن است هر نود تحت تأثیر وزن های مختلف قرار گیرد. شبکه عصبی تمام داده های آموزش را در لایه ورودی می گیرد. سپس داده ها را از میان لایه های پنهان عبور داده، مقادیر را براساس وزن هر نود تغییر می دهد و در نهایت مقداری را در لایه خروجی برمی گرداند.

¹Fréchet inception distance

²Neural Networks - NNs

³Nodes

⁴Layers



شکل ۴.۲: شبکه عصبی با چندین لایه پنهان. هر لایه چندین گره دارد.

تنظیم درست یک شبکه عصبی برای رسیدن به نتایج سازگار و قابل اعتماد ممکن است کمی زمان‌بر باشد. آزمایش و آموزش شبکه عصبی، یک فرآیند متعادل‌سازی برای تعیین مهم‌ترین ویژگی‌های مدل است.

۷.۲ شبکه عصبی پیچشی

شبکه عصبی پیچشی^۱ نوع خاصی از شبکه عصبی با چندین لایه است که داده‌هایی را که آرایش شبکه‌ای دارند، پردازش کرده و سپس ویژگی‌های مهم آن‌ها را استخراج می‌کند. یک مزیت بزرگ استفاده از CNN ها این است که نیازی به انجام پیش‌پردازش زیادی روی تصاویر نیست.

در بیشتر الگوریتم‌هایی که پردازش تصویر را انجام می‌دهند، فیلترها معمولاً توسط یک مهندس بر اساس روش‌های اکتشافی (heuristic) ایجاد می‌شوند. CNN ها می‌توانند مهم‌ترین ویژگی فیلترها را بیاموزند و چون به پارامترهای زیادی احتیاج نیست، صرفه‌جویی زیادی در وقت و عملیات آزمون و خطا صورت می‌گیرد.

هدف اصلی الگوریتم CNN این است که با حفظ ویژگی‌هایی که برای فهم آنچه داده‌ها نشان می‌دهند مهم هستند، داده‌ها را به فرم‌هایی که پردازش آن‌ها آسان‌تر است، درآورد. آن‌ها همچنین گزینه خوبی برای کار با مجموعه داده‌های عظیم هستند.

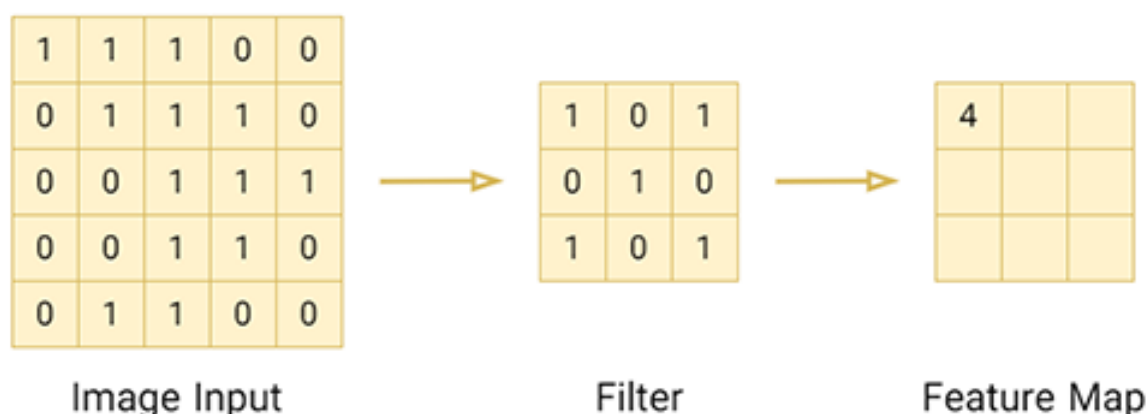
یک تفاوت بزرگ بین CNN و شبکه عصبی معمولی این است که CNN ها برای مدیریت ریاضیات پشت صحنه، از کانولوشن استفاده می‌کنند. حداقل در یک لایه از CNN ، به جای ضرب ماتریس از کانولوشن استفاده می‌شود. کانولوشن‌ها تا دو تابع را می‌گیرند و یک تابع را برمی‌گردانند.

CNN ها با اعمال فیلتر روی داده‌های ورودی شما کار می‌کنند. چیزی که آن‌ها را بسیار خاص می‌کند، این است که CNN ها می‌توانند فیلترها را هم‌زمان با فرایند آموزش، تنظیم کنند. به این ترتیب، حتی وقتی مجموعه داده‌های عظیمی مانند تصاویر داشته باشید، نتایج به‌خوبی و در لحظه دقیق‌تر می‌شوند. از آنجا که می‌توان فیلترها را برای آموزش بهتر CNN تازه‌سازی کرد، نیاز به فیلترهای دستی از بین می‌رود و این انعطاف‌پذیری بیشتری در تعداد و ارتباط فیلترهایی که بر روی مجموعه داده‌ها اعمال می‌شوند، به ما می‌دهد. با استفاده از این الگوریتم، می‌توانیم روی مسائل پیچیده‌تری مانند تشخیص چهره کار کنیم. کمبود داده یکی از مشکلاتی است که مانع استفاده از CNN می‌شود. با وجود اینکه می‌توان شبکه‌ها

¹Convolutional Neural Network - CNN

را با تعداد داده نسبتاً کمی، تقریباً ۱۰۰۰۰، آموزش داد، هرچه اطلاعات بیشتری در دسترس باشد، CNN بهتر تنظیم می‌شود. داده‌ها باید بدون نقص و دارای برجسب باشند تا CNN بتواند از آن‌ها استفاده کند و این چیزی است که باعث می‌شود کار کردن با آن‌ها زمان‌بر و نیازمند منابع سنگین محاسباتی باشد. شبکه‌های عصبی پیچشی براساس یافته‌های علوم اعصاب^۱ عمل می‌کنند. آن‌ها از لایه‌هایی از نورون‌های مصنوعی به نام نود^۲ ساخته شده‌اند. این نودها توابعی هستند که مجموع وزنی ورودی‌ها را محاسبه می‌کنند و یک نگاشت فعال‌سازی^۳ را برمی‌گردانند. این بخش پیچشی^۴ شبکه عصبی^۵ است.

Convolution



شکل ۵.۲: بخش پیچشی شبکه عصبی

هر نود در یک لایه توسط مقادیر وزنی آن تعریف می‌شود. وقتی به یک لایه داده‌هایی را می‌دهید، برای مثال یک تصویر، مقادیر پیکسل را می‌گیرد و برخی از ویژگی‌های بصری را جدا می‌کند. هنگامی که داده‌ها را به CNN می‌دهید، هر لایه نگاشت‌های فعال‌سازی را برمی‌گرداند. این نگاشت‌ها ویژگی‌های مهم مجموعه داده را شناسایی می‌کنند. اگر به CNN تصویری را بدهید، ویژگی‌های مبتنی بر مقادیر پیکسل مانند رنگ‌ها را شناسایی می‌کند و تابع فعال‌سازی را به شما ارائه می‌دهد. معمولاً در تصاویر، CNN در ابتدا لایه‌های تصویر را پیدا می‌کند. سپس این تعریف جزئی از تصویر به لایه بعدی منتقل می‌شود و آن لایه شروع به شناسایی مواردی مانند گوشه‌ها و گروه‌های رنگی می‌کند. سپس این تعریف جدید از تصویر به لایه بعدی منتقل می‌شود و چرخه تا پیش‌بینی ادامه پیدا می‌کند. همان‌طور که در تصویر زیر مشخص است، با افزایش تعداد لایه‌ها حداکثر تجمع (max-pooling) باید انجام شود. حداکثر تجمع فقط مرتبط‌ترین ویژگی‌ها از لایه موجود در نقشه فعال‌سازی را برمی‌گرداند و به لایه‌های بعدی منتقل می‌کند تا زمانی که به لایه آخر برسید. آخرین لایه CNN لایه طبقه‌بندی است.

¹neuroscience

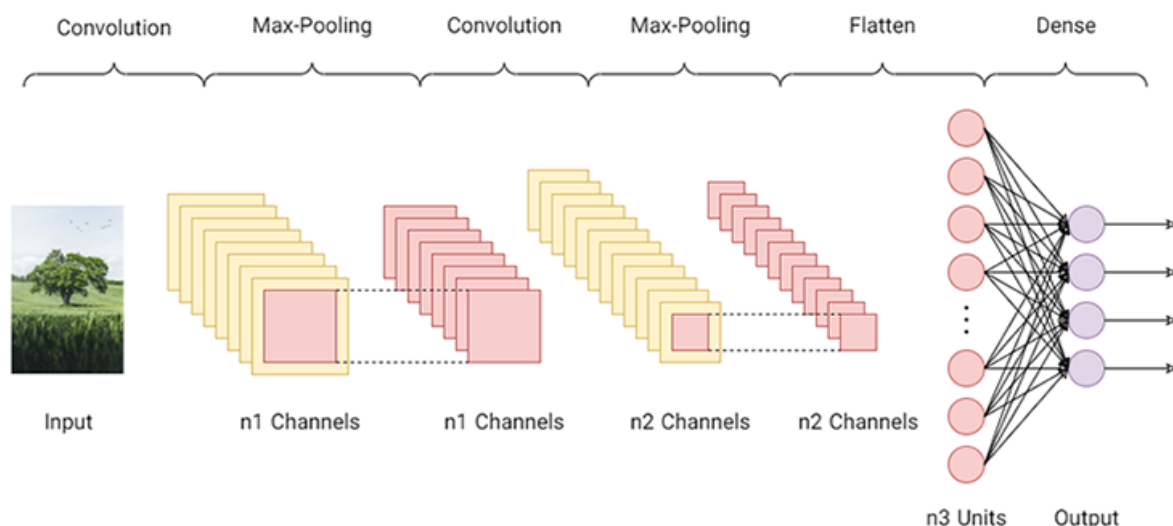
²node

³activation map

⁴Convolutional

⁵Neural Network

که مقدار پیش‌بینی شده را براساس نگاشت فعال‌سازی تعیین می‌کند. اگر یک نمونه دست خط را به CNN بدهید، لایه‌ی طبقه‌بندی حروف موجود در تصویر را به شما می‌گوید. این همان چیزی است که وسایل نقلیه خودران برای تعیین اینکه یک شیء اتومبیل، شخص و یا یک مانع است، استفاده می‌کنند.



شکل ۶.۲: معماری و لایه‌های یک شبکه عصبی پیچشی

انواع شبکه عصبی پیچشی به شرح زیر است:

- CNN یک‌بعدی: در این حالت، کرنل CNN در یک جهت حرکت می‌کند. CNN های یک‌بعدی معمولاً روی داده‌های سری زمانی استفاده می‌شوند.
 - CNN دوبعدی: در این نوع از CNN، کرنل‌ها در دو جهت حرکت می‌کنند. CNN های دوبعدی در برچسب‌گذاری و پردازش تصویر کاربرد دارند.
 - CNN سه‌بعدی: این نوع CNN دارای کرنلی است که در سه جهت حرکت می‌کند. محققان از این نوع CNN در تصاویر سه‌بعدی مانند سی‌تی‌اسکن و MRI استفاده می‌کنند.
- از آنجایی که بیشتر مسائل با داده‌های تصویر مرتبط هستند، اغلب از CNN های دوبعدی استفاده می‌شود. در ادامه برخی از کاربردهایی که ممکن است از CNN ها استفاده شود، آورده شده است.
- تشخیص تصاویر با پیش‌پردازش کم
 - تشخیص دست‌خط‌های مختلف
 - کاربردهای بینایی کامپیوتر (Computer Vision)
 - استفاده در بانک‌داری بای خواندن ارقام در چک
 - استفاده در سرویس‌های پستی برای خواندن کدپستی روی پاکت نامه

۸.۲ U-Net

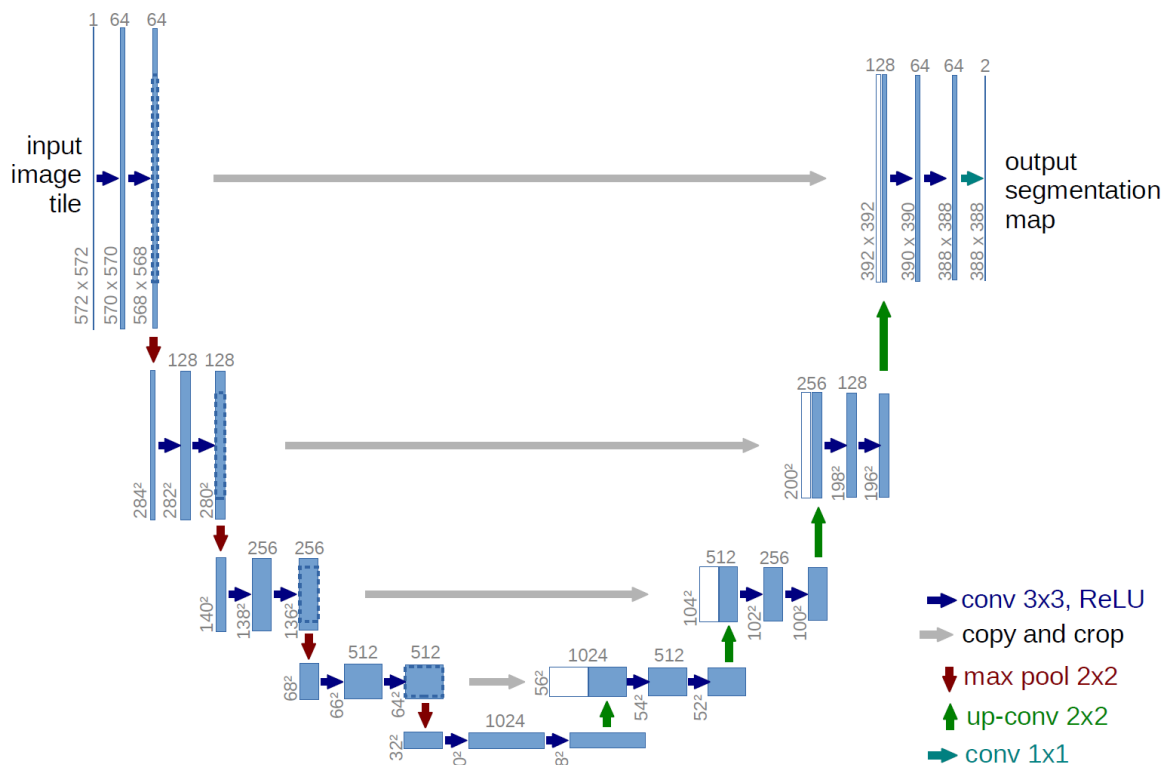
U-Net یک شبکه عصبی پیچشی است که برای بخش‌بندی تصاویر زیست پزشکی در دانشگاه فرایبورگ توسعه یافته است. این شبکه تماماً پیچشی است و معماری آن برای کار با تصاویر آموزشی کمتر و ارائه بخش‌بندی‌های دقیق‌تر توسعه یافته است.

معماری U-Net از معماری شبکه تماماً پیچشی نشأت می‌گیرد. ایده اصلی از ترکیب دو مسیر کاهنده و افزایشده تشکیل شده است که داده‌های مکانی مسیر کاهنده با مسیر افزایشده که تعداد ویژگی‌های آن بسیار زیاد است ترکیب می‌شود تا تصویر کیفیت بالا و با درک زمینه از مسیر کاهنده داشته باشیم و هر پیکسل یک به یک کلاس دسته‌بندی شود بدون اینکه هیچ گونه لایه تمام متصل استفاده شود.

معماری U-Net از یک مسیر کاهنده و یک مسیر افزایشده تشکیل شده است که به آن معماری U شکل را می‌دهد. در مسیر کاهنده تعدادی لایه پیچشی، تابع فعال‌ساز (ReLU) و لایه ادغام حداکثری (max-pooling) پیایی قرار دارد که ساختاری همانند شبکه عصبی باقی‌مانده تشکیل می‌دهند. در طول این مسیر ابعاد کاهش می‌یابد و با افزایش تعداد کانال‌ها ویژگی‌ها بیشتر می‌شوند.

در مسیر افزایشده تعدادی لایه پیچشی ترانهاد قرار دارد که ابعاد تصویر رو افزایش می‌دهند و با داده‌های مکانی مسیر کاهنده ترکیب می‌شوند و در نهایت تعدادی لایه پیچشی 1×1 قرار دارد که به تعداد کلاس‌های مورد نظر خروجی می‌دهد که این ساختار برای دسته‌بندی پیکسل‌ها در تصاویر مناسب است.

وجود مسیرهای انتقال گرادیان بین مسیر کاهنده و مسیر افزایشده باعث انتقال معنا از مسیر کاهنده به افزایشده می‌شوند و با انتقال گرادیان، فرایند تعلیم را پایدارتر می‌کنند.



شکل ۷.۲: معماری U-Net

از U-Net برای بخش‌بندی تصاویر زیست پزشکی استفاده‌های زیادی همانند بخش‌بندی تصاویر مغز

یا تصاویر کبد می‌شود. نسخه‌هایی از U-Net برای بازسازی و ترمیم تصاویر پزشکی نیز استفاده می‌شود. از دیگر کاربردهای U-Net می‌توان به بخش‌بندی تصاویر اشاره کرد که بخش‌بندی مثالی یا بخش‌بندی معنایی را می‌تواند انجام دهد. تفاوت این دو در شناخت و در تعداد است به صورتی که مثلاً بخش‌بندی مثالی چند شی مختلف حتی با هم‌پوشانی در تصویر شناسایی می‌کند در حالی که بخش‌بندی معنایی تنها آن شی را در تصویر شناسایی می‌کند. شبکه-یو تنها در بخش‌بندی معنایی کاربرد دارد و برای بخش‌بندی مثالی از R-CNN و مشتقات آن استفاده می‌شود.

رگرسیون پیکسلی از دیگر کاربردهای U-Net است که برای رنگ‌آمیزی تصاویر استفاده می‌شود. رفع نویز از دیگر کاربردهای U-Net می‌تواند باشد و ساختاری مشابه خودرمزگذار رفع‌کننده نویز دارد که عملکرد بهتری نیز به خاطر کانال‌های عبور گرادین می‌تواند داشته باشد.

از دیگر کاربردها می‌توان به فراتفکیک‌پذیری اشاره کرد و مشابه هنگامی که برای ترمیم تصاویر یا رفع نویز استفاده می‌شود، به تصویر جزئیات اضافه کند.

از U-Net در ساختار انتشار پایدار (Stable Diffusion) نیز استفاده شده و در طی فرایند انتشار با مرحله به مرحله رفع نویز با اعمال شرطی که از متن به آن وارد شده باعث ساخت تصاویر معنادار و توصیف شده توسط متن از نویز ورودی می‌شود.

فصل ۳

کار مربوطه

۱.۳ استخراج ویژگی‌ها

استخراج ویژگی‌ها بخش بسیار مهمی در تجزیه و تحلیل و یافتن روابط بین چیزهای مختلف است. همانطور که قبلاً می‌دانیم که داده‌های ارائه شده از صدا را نمی‌توان مستقیماً توسط مدل‌ها درک کرد، بنابراین ما باید آنها را به یک قالب قابل درک تبدیل کنیم که استخراج ویژگی برای آن استفاده می‌شود. سیگنال صوتی یک سیگنال سه بعدی است که در آن سه محور زمان، دامنه‌ی نوسان و فرکانس را نشان می‌دهد. هر فایل موسیقی اساساً از دو چیز مهم تشکیل شده است:

- نرخ نمونه^۱

- داده نمونه^۲

اکنون با کمک نرخ نمونه و داده‌های نمونه می‌توان چندین تغییر شکل روی آن انجام داد تا ویژگی‌های ارزشمندی را از آن استخراج کرد که در قسمت زیر آمده است:

۱. نرخ عبور از صفر^۳: نرخ‌ی است که در آن یک سیگنال از مثبت به صفر به منفی یا از منفی به صفر به مثبت تغییر می‌کند.

$$zcr = \frac{1}{T-1} \sum_{t=1}^{T-1} 1_{\mathbb{R} < 0}(s_t s_{t-1}) \quad (1.3)$$

۲. انرژی:

$$\mathbb{E}_s = \langle f(t), f(t) \rangle = \int_{-\infty}^{\infty} |f(t)|^2 dt \quad (2.3)$$

¹Sample Rate

²Sample Data

³Zero-crossing rate

۳. آنتروپی: معیاری عددی برای اندازه گرفتن اطلاعات، یا تصادفی بودن یک متغیر تصادفی است. به بیان دقیق‌تر، آنتروپی یک متغیر تصادفی، متوسط اطلاعات آن است. با داشتن یک متغیر تصادفی گسسته X که مقادیری از الفبای X ، آنتروپی برای آن به صورت زیر تعریف می‌شود:

$$H(x) = - \sum_{x \in X} p(x) \log p(x) \quad (۳.۳)$$

۴. مرکز طیفی: نشان می‌دهد که مرکز جرم طیف در کجا قرار دارد. از نظر ادراکی، ارتباط قوی با تأثیر روشنایی صدا دارد. گاهی به آن مرکز جرم طیفی نیز می‌گویند.

$$C_i = \frac{\sum_{k=1}^{Wf_L} k X_i(k)}{\sum_{k=1}^{Wf_L} X_i(k)} \quad (۴.۳)$$

۵. گسترش طیفی: دومین لحظه مرکزی طیف است که برای محاسبه آن باید انحراف طیف را از مرکز طیفی مطابق معادله زیر گرفت:

$$S_i = \sqrt{\frac{\sum_{k=1}^{Wf_L} (k - C_i)^2 X_i(k)}{\sum_{k=1}^{Wf_L} X_i(k)}} \quad (۵.۳)$$

۶. آنتروپی طیفی: محاسبه توزیع توان طیفی همراه با قابلیت پیش‌بینی سیگنال سری زمانی است. این آنتروپی بر اساس شانون و آنتروپی اطلاعات در داده‌های اطلاعاتی است. آنتروپی طیفی سیگنال توسط:

$$SE(F) = - \frac{1}{\log N_u \sum_u (p_u(F) \log_e P_u(F))} \quad (۶.۳)$$

$$SSH(F) = - \sum_u (P_h(F) \log_e P_h(F)), \quad (۷.۳)$$

به صورتی که $P_u(F)$ نشان دهنده‌ی تابع چگالی طیفی توان، $P_h(F)$ نشان دهنده تخمین آنتروپی شانون $(SSH(F))$ ، و N_u کل فرکانس‌ها را نشان می‌دهد.

۷. شار طیفی: شار طیفی تغییر طیفی بین دو فریم متوالی را اندازه‌گیری می‌کند و به عنوان اختلاف مجذور بین مقادیر نرمال شده طیف دو پنجره کوتاه مدت متوالی محاسبه می‌شود:

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_L} (EN_i(k) - EN_{i-1}(k))^2, \quad (۸.۳)$$

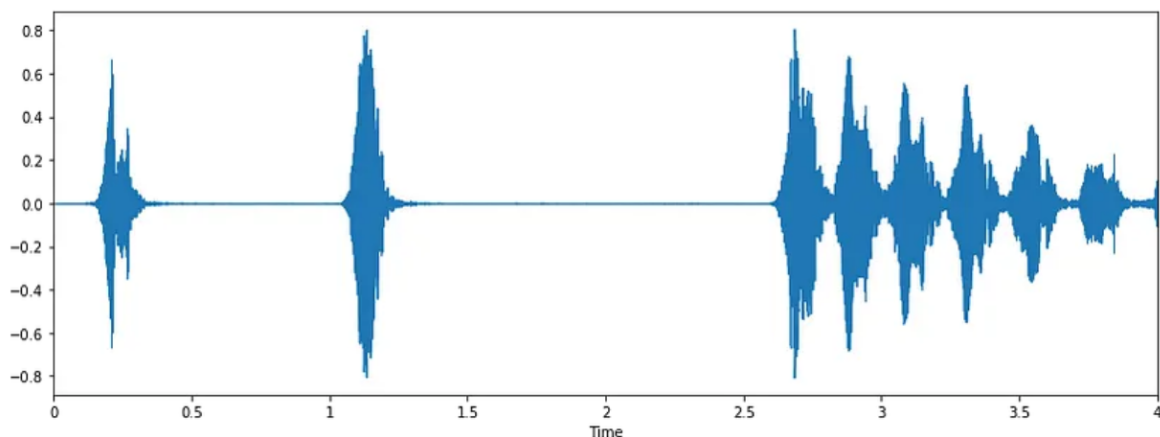
به صورتی که $EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_L} X_i(l)}$ نشان دهنده‌ی k امین ضریب نرمال شده در فریم i ام است.

۸. افت تدریجیه طیفی: ^۱ نقطه ای است که درصد معینی از کل انرژی طیفی زیر آن قرار دارد. به تمایز بین صداهای گفتاری صدادار و بدون صدا کمک می کند.

۹. ضرایب طیفی فرکانس های ملتویی: ^۲ یک روش مهم در پردازش سیگنال های صوتی است. این روش با استفاده از تقسیم سیگنال صوتی به قطعات کوتاه و استخراج ویژگی های فرکانسی و زمانی از آن، در تشخیص و شناسایی گفتار و الگوهای صوتی مؤثر استفاده می شود. MFCC اطلاعات حیاتی درباره ی طیف فرکانسی و ویژگی های زمانی سیگنال صوتی را ارائه می دهد و در سیستم های تشخیص گفتاری و شناسایی الگوهای صوتی به کار می رود.

۱.۱.۳ Waveplots

صدای خام را می توان به عنوان یک نمودار موج تجسم کرد که نمونه ای از آن در شکل زیر داده شده است. یک نمودار موج، پوشش دامنه سیگنال را در برابر زمان ترسیم می کند. تجسم اینکه یک سیگنال چگونه به نظر می رسد می تواند مفید باشد، اما معمولاً برای مدل های یادگیری ماشین در پیش بینی ها مفید نیست. برای اینکه یک سیگنال مفید باشد، لازم است ویژگی های کمتر آشکار استخراج شود. معمولاً ویژگی های استخراج شده را می توان به دو دسته تقسیم کرد: زمانی که به ویژگی های وابسته به زمان می پردازد و طیفی که با ویژگی های وابسته به فرکانس سروکار دارد.



شکل ۱.۳: طرحی از یک Waveplot

۲.۱.۳ طیف نگاری

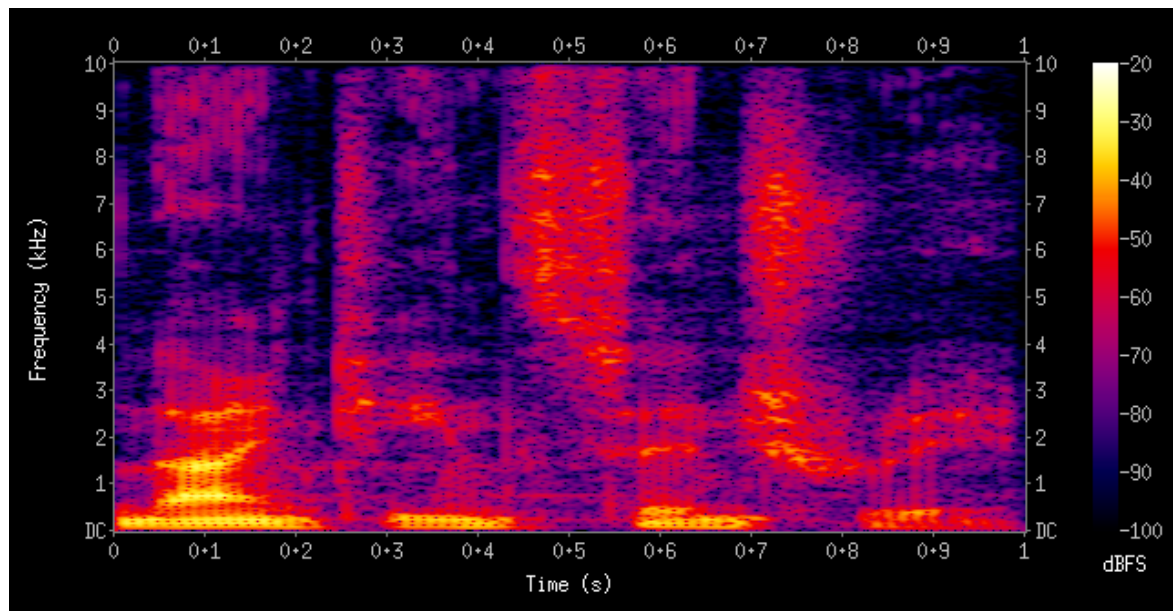
یک طرح طیف نگاره ^۳ روشی بصری برای نمایش قدرت سیگنال یا بلندی سیگنال در طول زمان در فرکانس های مختلف موجود در یک شکل موج خاص است. این به ما کمک میکند که در طول زمان در فرکانس های مختلف مشاهده کنیم چگونه سطوح انرژی تغییر می کند. طیف نگاره ها نمودارهای دو بعدی هستند که بعد سوم با رنگ ها نشان داده می شود. زمان از چپ (قدیمی ترین) به راست (جوان ترین)

^۱Spectral Roll off

^۲MFCC

^۳Spectrogram

در امتداد محور افقی می‌گذرد. محور عمودی نشان‌دهنده فرکانس است که می‌توان آن را به صورت زیر و بم یا تن نیز در نظر گرفت، با کمترین فرکانس در قسمت پایین و بیشترین فرکانس در قسمت بالایی قرار گرفته است. دامنه (یا انرژی یا "بلندی") یک فرکانس خاص در یک زمان خاص با بعد سوم، رنگ، با آبی تیره مربوط به دامنه‌های کم و رنگ‌های روشن‌تر تا قرمز مربوط به دامنه‌های به تدریج قوی‌تر (یا بلندتر) نشان داده می‌شود.



شکل ۲.۳: طرحی از یک طرح نگاره

۲.۳ مجموعه داده

ابتدا برای اینکه یک مدل تشخیص دهنده‌ای را ایجاد کنیم نیاز است که مجموعه داده صوتی احساسی را استفاده کنیم، اگر چه برخی از این پایگاه‌های داده به دلیل حفظ حریم شخصی افراد در دسترس عموم توسعه دهنده‌گان این حوزه قرار نگرفته است، برخی از این پایگاه‌های داده مورد استفاده در مدل مولد و مدل تشخیص احساسات در این پروژه به شرح زیر است:

- **Ravdess**: این پایگاه داده شامل ۱۴۴۰ فایل است که هر گوینده دارای ۶۰ فایل صوتی است. این دیتاست شامل ۲۴ گوینده (۱۲ مرد و ۱۲ زن) است، که به زبان انگلیسی با گویش آمریکای شمالی صحبت کرده‌اند. احساسات مورد استفاده شامل موارد زیر است:

- آرام (Calm)
- شاد (Happy)
- غمگین (Sad)
- خشمگین (Angry)

□ ترسناک (Fearful)

□ حیرت زده (Surprise)

□ منزجر (Disgust)

هر یک از موارد بالا در سه درجه عادی، خنثی، و شدید جمع آوری شده است.

۱.۲.۳ SAVEE

پایگاه داده SAVEE از چهار مرد بومی زبان انگلیسی (مشخص شده به عنوان DC، JE، JK، KL)، دانشجویان کارشناسی ارشد و محققان دانشگاه ساری در سنین ۲۷ تا ۳۱ سال ثبت شد. عاطفه از نظر روانشناختی در دسته بندی های مجزا توصیف شده است: خشم، انزجار، ترس، شادی، اندوه و تعجب. یک دسته خنثی نیز اضافه شده است تا ضبط ۷ دسته احساسات را ارائه دهد. محتوای متن شامل ۱۵ جمله TIMIT در هر احساس بود: ۳ جمله رایج، ۲ جمله خاص هیجان و ۱۰ جمله کلی که برای هر احساس متفاوت و از نظر آوایی متعادل بودند. ۳ جمله رایج و $6 \times 2 = 12$ جمله خاص هیجانی به عنوان خنثی ثبت شد تا ۳۰ جمله خنثی ارائه شود. این منجر به ۱۲۰ بیان برای هر گوینده شد.

۲.۲.۳ ESD

پایگاه داده ESD توسط دانشگاه ملی سنگاپور (NUS) و دانشگاه فناوری و طراحی سنگاپور (SUTD) در دسترس است. پایگاه داده ESD شامل ۳۵۰ گفتار موازی است که توسط ۱۰ انگلیسی بومی و ۱۰ فرد چینی (Mandarian) صحبت می شود و ۵ کلاس احساسی (خنثی، شادی، خشم، غم و حیرت) را پوشش می دهد. بیش از ۲۹ ساعت داده گفتار در محیط آکوستیک کنترل شده ثبت شده است. بنابراین، برای مطالعات تبدیل صدای عاطفی چند گوینده و چند زبانه مناسب است. کاربردهای این مجموعه داده به صورت زیر است:

- تبدیل صدای احساسی (تک زبانه و چند زبانه، وابسته به گوینده و مستقل از گوینده)
- تبدیل صدا (تک زبانه و چند زبانه)
- متن به گفتار احساسی
- بیان متن به گفتار

۳.۲.۳ Tess

در این مجموعه از ۲۰۰ کلمه هدف در عبارت حامل "کلمه را بگویید" توسط دو بازیگر زن (۲۶ و ۶۴ ساله) بیان شده است و ضبط هایی از مجموعه انجام شده است که هر یک از هفت احساس (خشم، انزجار، ترس، شادی، غافلگیری دلیزیر، غمگینی و خنثی) را به تصویر می کشد. در مجموع ۲۸۰۰ نقطه داده (فایل صوتی) وجود دارد. مجموعه داده به گونه ای سازماندهی شده است که هر یک از دو بازیگر زن و احساسات آنها در پوشه مخصوص به خود قرار دارند. در آن، تمام ۲۰۰ کلمه هدف فایل صوتی را می توان یافت. فرمت فایل صوتی فرمت WAV می باشد.

Parameter	Mandarin						English					
	Neu	Ang	Sad	Hap	Sur	All	Neu	Ang	Sad	Hap	Sur	All
# speakers	10	10	10	10	10	10	10	10	10	10	10	10
# utterances per speaker	350	350	350	350	350	1,750	350	350	350	350	350	1,750
# unique utterances	350	350	350	350	350	350	350	350	350	350	350	350
# characters/words per speaker	4,005	4,005	4,005	4,005	4,005	20,025	2,203	2,203	2,203	2,203	2,203	11,015
# unique characters/words	939	939	939	939	939	939	997	997	997	997	997	997
Avg. utterance duration [s]	3.23	2.68	4.04	2.84	3.32	3.22	2.61	2.80	2.98	2.70	2.73	2.76
Avg. character/word duration [s]	0.28	0.23	0.35	0.25	0.29	0.28	0.41	0.44	0.47	0.43	0.43	0.44
Total duration [s]	11,305	9,380	14,140	9,940	11,620	56,385	9,135	9,800	10,430	9,450	9,555	48,370

Emotion abbreviations are used as follows: *Neu* stands for neutral, *Ang* stands for anger, *Sad* stands for sadness, *Hap* stands for happiness and *Sur* stands for surprise. The number of characters is reported for Mandarin, and the number of words is reported for English.

شکل ۳.۳: جزئیات مجموعه داده ESD

۴.۲.۳ CREMA-D

CREMA-D یک مجموعه داده عاطفی بازیگر چندوجهی از ۷۴۴۲ کلیپ اصلی از ۹۱ بازیگر است. این کلیپ ها از ۴۸ بازیگر مرد و ۴۳ بازیگر زن بین ۲۰ تا ۷۴ سال بود که از نژادها و قومیت های مختلف (آمریکای آفریقایی، آسیایی، قفقازی، اسپانیایی تبار و سایر قومیتها) بودند. بازیگران از مجموعه ای از ۱۲ جمله صحبت کردند. جملات با استفاده از یکی از شش احساس مختلف (خشم، انزجار، ترس، خوشحالی، خنثی، و غمگین) و چهار سطح هیجانی مختلف (کم، متوسط، زیاد و نامشخص) ارائه شدند. شرکت کنندگان بر اساس ارائه ترکیبی سمعی و بصری، ویدئو به تنهایی و صوت به تنهایی، احساسات و احساسات را ارزیابی کردند. با توجه به تعداد زیاد رتبه بندی های مورد نیاز، این تلاش به صورت جمعی انجام شد و در مجموع ۲۴۴۳ شرکت کننده هر کدام ۹۰ کلیپ منحصر به فرد، ۳۰ کلیپ صوتی، ۳۰ تصویری و ۳۰ کلیپ صوتی و تصویری را رتبه بندی کردند. ۹۵ درصد از کلیپ ها بیش از ۷ امتیاز دارند.

۵.۲.۳ روش پیشنهادی

انتقال سبک^۱ یک موضوع مهم در پردازش تصویر است. هدف آن ترکیب یک تصویر بدیع با ترکیب محتوای یک تصویر با سبک تصویر دیگر است. اگر احساسات را به عنوان یک "سبک" عاطفی گفتار در نظر بگیریم، انتقال احساسات می تواند به عنوان روشی برای ایجاد نمونه های جدیدی از گفتار عاطفی مورد استفاده قرار گیرد.

به طور کلی، بین انتقال سبک هدایت شده با نمونه، که در آن سبک هدف از یک مثال منفرد حاصل می شود، و انتقال سبک مجموعه، که در آن سبک هدف با یک مثال تعریف می شود، تفاوت وجود دارد. مجموعه ای از تصاویر به عنوان مثال، انتقال سبک هدایت شده با نمونه می آموزد که سبک *The Starry Night* را تقلید کند، یک قطعه واحد از ون گوگ، در حالی که انتقال سبک مجموعه یاد می گیرد که به سبک کل مجموعه ون گوگ تصاویر تولید کند. از نظر انتقال عاطفه، انتقال سبک مجموعه ارجحیت دارد، زیرا می تواند ویژگی های عاطفی را که تمام گفته های مجموعه با آنها مشترک است به تصویر بکشد و از تأثیرات ویژگی های غیر عاطفی که برخی از گفته های منفرد حاوی آن هستند جدا شود. انتقال سبک مجموعه را می توان به عنوان فعالیتی برای ترجمه تصویر به تصویر در نظر گرفت که می

¹Style transfer

تواند برای کارهای دیگر مانند تغییر شکل شی، تصویر به برچسب های معنایی و غیره نیز اعمال شود. شبکه pix2pix یک شبکه مولد متخاصمی است که برای ترجمه تصویر به تصویر مورد استفاده قرار میگیرد، اما به داده های آموزشی جفتی نیاز دارد که به دست آوردن آنها دشوار و پرهزینه است. از آنجایی که مشکل ترجمه تصویر به تصویر بدون نظارت ذاتاً نامناسب است و نیاز به محدودیت های اضافی دارد. از تابع cycle-consistent به عنوان محدودیت در CycleGAN ها استفاده میشود و به نتایج قانع کننده ای در انتقال سبک تصویر حتی با داده های آموزشی جفت نشده دست میابد. به دلیل عدم وجود داده های آموزشی موازی در بیشتر پایگاه های گفتاری احساسی، CycleGAN مناسب ترین روش انتقال سبک تصویر است که می تواند با احساسات سازگار شود.

۳.۳ انتقال احساسات مبتنی بر شبکه های مولد متخاصمی

از آنجایی که یک مدل CycleGAN نگاشت یک به یک بین یک منبع و یک دامنه هدف را می آموزد، برای مجموعه داده گفتاری برچسب گذاری شده با کلاس های احساسی، ما باید بین هر جفت از آن ها نقشه برداری ایجاد کنیم، نگاشت $C(N, 2)$ که بسیار گران است. در روش ما، داده های برچسب گذاری شده هر نوع احساس به عنوان دامنه هدف استفاده می شود در حالی که دامنه منبع یک مجموعه داده بزرگ خارجی بدون برچسب است. مفهوم انتقال احساسات در روش ما یک CycleGAN منبع واقعی و هدف واقعی خود را به ترتیب به یک هدف مصنوعی و یک منبع مصنوعی نگاشت می کند. بنابراین، می توانیم یک مجموعه داده هدف مصنوعی تولید کنیم که به اندازه منبع واقعی باشد و احساساتی مشابه مجموعه داده هدف واقعی داشته باشد. هدف مصنوعی بیشتر برای تقویت داده ها برای تشخیص احساسات استفاده می شود، در حالی که منبع مصنوعی در حوزه مجموعه داده بدون برچسب است و نادیده گرفته می شود. به جای آموزش N CycleGAN به طور جداگانه، ما CycleGAN ها را در یک چارچوب کامل قرار می دهیم تا نمونه های تولید شده از هر احساس هدف را با یکدیگر مرتبط کنیم، این چارچوب در بخش های بعدی توضیح داده خواهد شد.

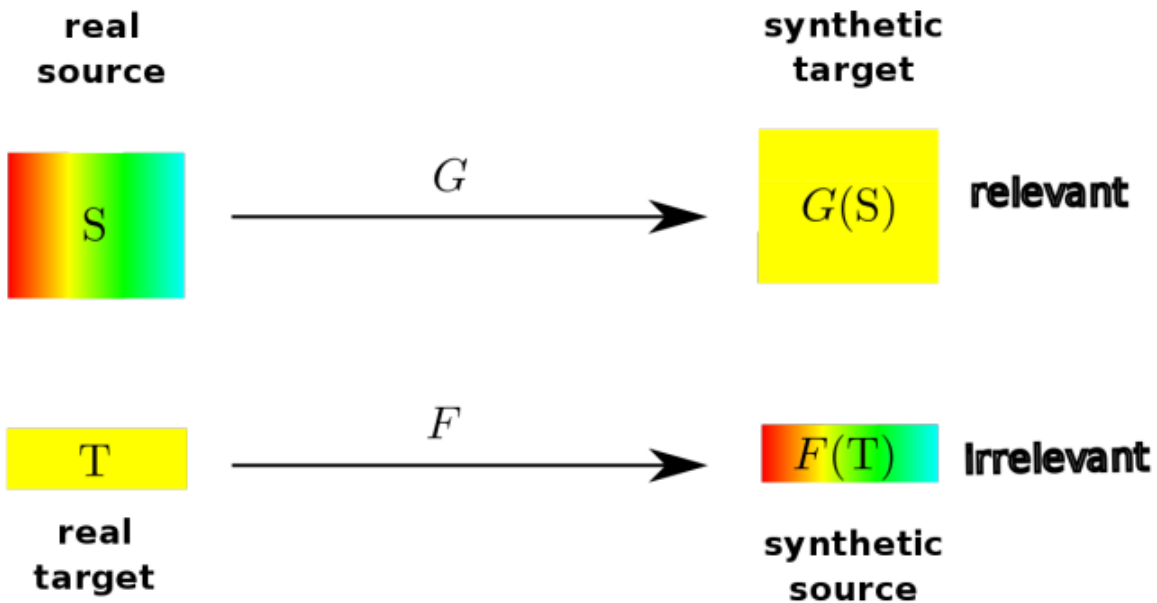
۴.۳ اجزاء شبکه CycleGan

با توجه به مجموعه داده برچسب دار X با کلاس های احساسی، نمونه های مصنوعی برای هر احساس i با استفاده از یک CycleGAN تولید می کنیم. همانطور که در شکل ۵.۱ مشاهده میکنید، این شبکه CycleGan یک نگاشت متقابل بین دامنه S و دامنه T_i به صورت یک به یک ایجاد میکند، که در آن مجموعه S یک مجموعه داده بدون برچسب خروجی است، و T_i مجموعه داده های احساسی کلاس i است. دو تابع نگاشت G_i و F_i به ترتیب برای ترجمه از منبع به مقصد و از هدف به منبع استفاده می شود. تشخیص دهنده متخاصم D_i^T ، مولد G_i را ترقیب به ساخت داده های مصنوعی متمایز از داده های واقعی میکند. سیستم این شبکه به ازای هر کلاس i به صورت زیر است:

(۹.۳)

$$\mathcal{L}_i^{GAN}(G_i, F_i, D_i^T, D_i^S, S, T_i) = \mathcal{L}_i^{GAN}(G_i, D_i^T, S, T_i) + \mathcal{L}_i^{GAN}(F_i, D_i^S, S, T_i)$$

به صورتی که مولدهای T_i و F_i سعی می کنند آن را به حداقل برسانند در حالی که تمایز کنندگان D_i^T و D_i^S سعی می کنند آن را به حداکثر برسانند.

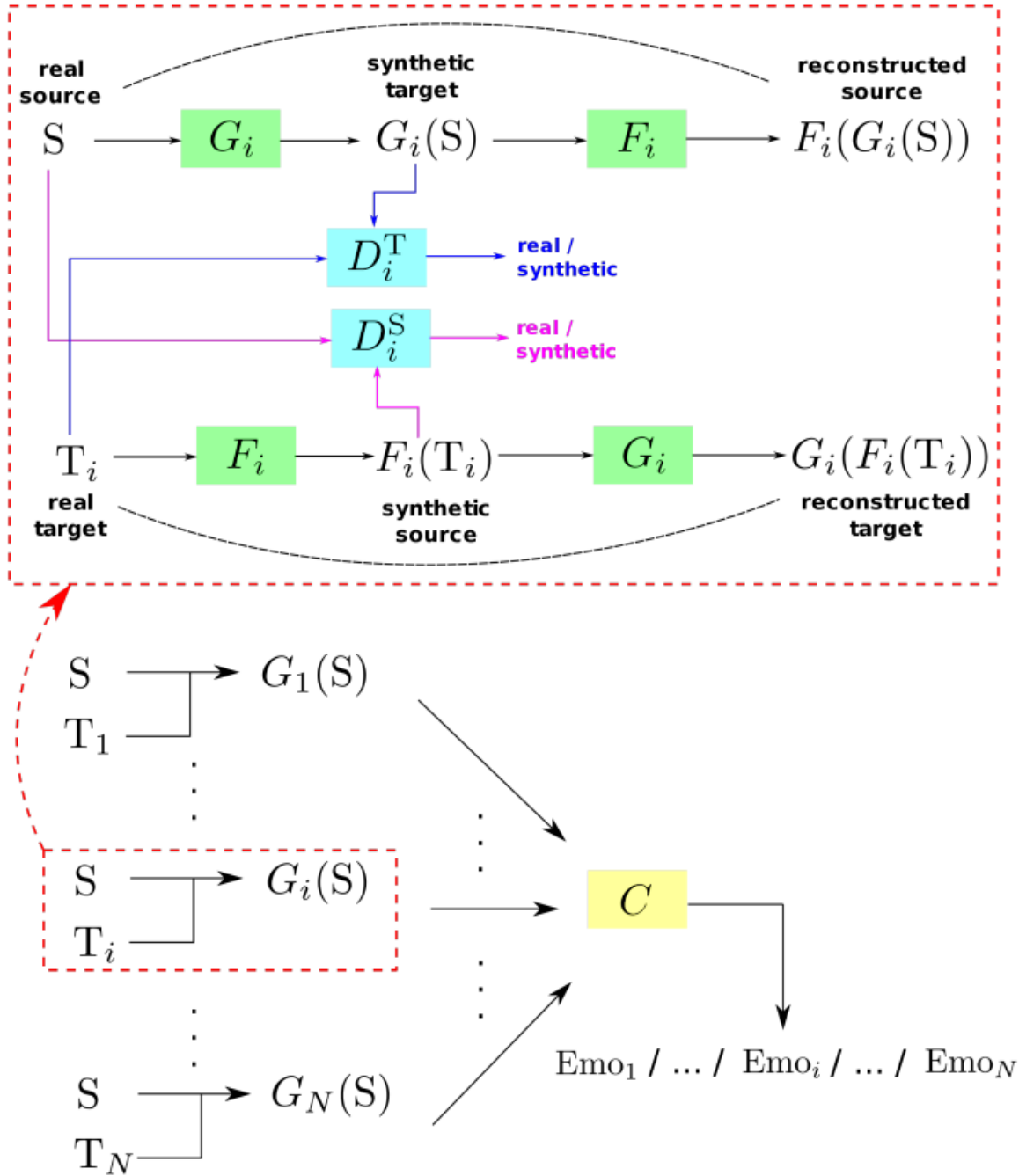


شکل ۴.۳: مفهوم انتقال احساسات برای جلوگیری از لزوم ساختن نقشه برداری برای هر جفت کلاس احساس.

۵.۳ استفاده از مجموعه داده بزرگ بدون برچسب

تولید داده مصنوعی فرآیندی است که داده های منبع را به یک احساس هدف نگاشت می کند. در حالی که احساسات هدف یکی از کلاس های برچسب گذاری شده است، نمونه های منبع به داده های برچسب گذاری شده محدود نمی شوند. به دلایل زیر از یک مجموعه داده خارجی بدون برچسب بزرگ به عنوان منبع استفاده می کنیم:

- داده های بدون برچسب فراوان هستند و به راحتی به دست می آیند
- از آنجایی که بدنه خارجی محتوای متفاوتی با داده های برچسب گذاری شده دارد، نمونه های تولید شده حاوی اطلاعات بالقوه مفید جدید هستند.
- هنگامی که همان داده های بدون برچسب به هر یک از احساسات هدف منتقل می شود، مجموعه داده مصنوعی از نمونه های متعادلی تشکیل می شود که فقط در احساسات متفاوت هستند. در نتیجه، طبقه بندی می تواند بیشتر بر روی احساسات متمرکز شود و مستقل از محتوا باشد.



شکل ۵.۳: معماری این سیستم

فصل ۴

نتایج

۱.۴ معماری سیستم

از آنجایی که پنج احساس وجود دارد که باید طبقه بندی شوند، مدل ما از پنج مولد، پنج متمایز کننده و یک دسته‌بند تشکیل شده است. همه آنها توسط شبکه های عصبی پیچشی اجرا می شوند. با توجه به دشواری مولدها برای یادگیری توزیع ابعاد بالا، ما هر جفت مولد G_i و F_i را بر اساس اختلاف بین S و $F_i(G_i(S))$ و همچنین اختلاف بین T_i و $G_i(F_i(T_i))$ از قبل آموزش می دهیم.

با توجه به اینکه تصاویر طیف نگاره های تولید به صورت $640 \times 480 \times 3$ هستند، از معماری شبکه U-net برای تولید عکس های ساختگی استفاده میکنیم، جزئیات دقیق این ساختار به همراه جزئیات معماری شبکه تشخیص دهنده در جدول ۱.۴ آمده است.

برای پیاده سازی این پروژه از ابزار PyTorch استفاده شده، همچنین جزئیات ابرپارامترهای آموزشی در جدول ۲.۴ به صورت کامل آمده است.

جدول ۱.۴: معماری GAN برگرفته شده از ساختار U-net

معماری مولد		
اندازه خروجی	عملیات	لایه
(۲۵۶،۱۶۰،۱۲۰)	کانولوشن + فعال ساز نرمال ساز	۳ لایه Encoder
(۲۵۶،۱۶۰،۱۲۰)	کانولوشن + فعال ساز نرمال ساز	۴ لایه Residual
(۳،۶۴۰،۴۸۰)	کانولوشن + فعال ساز نرمال ساز	۳ لایه Decoder
معماری تشخیص دهنده		
اندازه خروجی	عملیات	لایه
(۵۱۲،۸۰،۶۰)	کانولوشن + فعال ساز نرمال ساز	Feature Map
(۳،۶۴۰،۴۸۰)	کانولوشن + فعال ساز نرمال ساز	۴ لایه تقلیلی
(۱،۸۰،۶۰)	کانولوشن + فعال ساز نرمال ساز	لایه نهایی

جدول ۲.۴: ابرپارامترهای آموزش

پارامتر	مقدار
نرخ یادگیری	2×10^{-4}
الگوریتم بهینه سازی	Adam
تابع خطا	Binary Cross-Entropy
ابعاد ورودی	(۳،۶۴۰،۴۸۰)
اندازه هر دسته داده	۱