Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 10, Number 4, Winter 2022, pp. 88-98
Received: 03/03/2021, Revised: 11/09/2021, Accepted: 01/11/2022

Improving Video Semantic Segmentation using Deep Neural Networks and Optical Flow

Mohammad Mehdi Najafi[1] and Mohammad Fakhredanesh[2*]
1- Faculty of Electrical and Computer Engineering, Malek-e-Ashtar University of Technology, Tehran, Iran.
2*- Faculty of Electrical and Computer Engineering, Malek-e-Ashtar University of Technology, Tehran, Iran.
[1]m.najafi@mut.ac.ir, and[2*]fakhredanesh@mut.ac.ir

Nowadays, video semantic segmentation is used in many applications such as automatic driving, navigation systems, virtual reality systems, etc. In recent years, significant progress has been observed in semantic segmentation of images. Since the consecutive frames of a video must be processed with high speed, low latency, and in real time, using semantic image segmentation methods on individual video frames is not efficient. Therefore, semantic segmentation of video frames in real time and with appropriate accuracy is a challenging topic. In order to encounter the mentioned challenge, a video semantic segmentation framework has been introduced. In this method, the previous frames semantic segmentation has been used to increase speed and accuracy. For this manner we use the optical flow (change of continuous frames) and a GRU deep neural network called ConvGRU. One of the GRU input is estimation of current frames semantic segmentation (resulting from a pre-trained convolutional neural network), and the other one is warping of previous frames semantic segmentation along the optical flow. The proposed method has competitive results on accuracy and speed. This method achieves good performances on two challenging video semantic segmentation datasets, particularly 83.1% mIoU on Cityscapes and 79.8% mIoU on CamVid dataset. Meanwhile, in the proposed method, the semantic segmentation speed using a Tesla P4 GPU on the Cityscapes and Camvid datasets has reached 34 and 36.3 fps, respectively.

Keyword- Video Semantic Segmentation, Deep Neural Network, Optical Flow.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۰، شماره ۴، زمستان ۱۴۰۰ ، صفحه ۸۸-۹۸
دریافت: ۱۳۹۹/۱۲/۱۳، بازنگری: ۱۴۰۰/۰۸/۱۸، پذیرش: ۱۴۰۰/۱۰/۲۱

بهبود تقطیع معنایی ویدئو با استفاده از شبکه‌های عصبی عمیق و جریان نوری

محمد مهدی نجفی[۱]، و محمد فخردانش[۲]
۱- مجتمع مهندسی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر ، تهران، ایران.
*۲- مجتمع مهندسی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر، تهران، ایران.
[1]m.najafi@mut.ac.ir, and[2*]fakhredanesh@mut.ac.ir

* نشانی نویسنده مسئول: محمدفخردانش، تهران، دانشگاه صنعتی مالک اشتر، مجتمع مهندسی برق و کامپیوتر. کدپستی: ۱۷۷۴-۱۵۸۷۵.

چکیده- امروزه از تقطیع معنایی ویدئو در کاربردهای بسیاری از قبیل خودروهای بدون سرنشین، سیستم‌های ناوبری، سیستم‌های واقعیت مجازی و ... استفاده می‌شود. در سال‌های اخیر پیشرفت چشم‌گیری در تقطیع معنایی تصاویر مشاهده شده است. اما از آن‌جا که فریم‌های پشت سر هم یک ویدئو باید با سرعت بالا و تاخیر کم و به صورت بلادرنگ پردازش شوند استفاده از تقطیع معنایی تصویر روی تک تک فریم‌های ویدئو با مشکل مواجه می‌شود؛ بنابراین تقطیع معنایی فریم‌های یک ویدئو به صورت بلادرنگ و با دقت مناسب موضوعی چالش برانگیز است. به منظور مقابله با چالش ذکر شده، در این مقاله یک چارچوب تقطیع معنایی ویدئو معرفی شده است که با در نظر گرفتن تغییرات فریم‌های پشت سر هم (با استفاده از جریان نوری) و بهره گیری از شبکه عمیق بازگشتی GRU، از اطلاعات تقطیع معنایی فریم‌های قبلی به منظور افزایش سرعت و دقت استفاده شده است. یک ورودی شبکه GRU تخمینی از تقطیع معنایی فریم فعلی (حاصل از یک شبکه عمیق کانولوشنال از پیش آموزش دیده)، و ورودی دیگر آن لغزش یافته تقطیع معنایی فریم قبلی در راستای جریان نوری دو فریم قبلی و فعلی می‌باشد. روش پیشنهادی دارای دقت و سرعت قابل رقابت با شناخته شده‌ترین و بهترین روش‌ها می‌باشد. دقت تقطیع معنایی بر اساس معیار ارزیابی mIoU روی مجموعه داده‌های Cityscapes و Camvid به ترتیب برابر با ۸۳.۱ و ۷۹.۸ می‌باشد. این در حالیست که در روش پیشنهادی سرعت تقطیع معنایی با استفاده از یک GPU تسلا مدل P4 روی مجموعه داده‌های Cityscapes و Camvid به ترتیب به ۳۴ و ۳۶.۳ فریم بر ثانیه رسیده است.

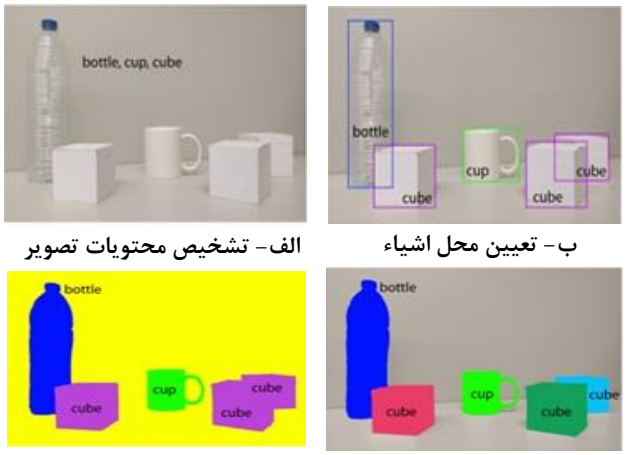
واژه‌های کلیدی: تقطیع معنایی ویدئو، شبکه عصبی عمیق، جریان نوری.

۱- مقدمه

انجام پردازش‌های مختلف روی دنباله فریم‌های ویدئو از اهمیت زیادی در بینایی ماشین برخوردار است. تجزیه و تحلیل یک تصویر یا فریم ویدئویی می‌تواند در چند مرحله انجام شود. تشخیص محتویات، تعیین محل اشیاء، تقطیع معنایی و تقطیع نمونه مواردی هستند که مربوط به تجزیه و تحلیل یک تصویر یا فریم می‌باشند. (شکل ۱)

تقطیع معنایی یک پردازش پایه‌ای است که در آن به هر پیکسل[۱] تصویر ورودی یک برچسب تخصیص داده می‌شود. تقطیع معنایی ویدئو، به معنای تعیین یک برچسب برای هر پیکسل فریم‌های ویدئو و به دنبال آن تقسیم‌بندی اشیاء موجود در هر فریم بر اساس مفهوم می‌باشد [۲، ۳، ۴]. نکته اصلی که پردازش فریم‌های ویدئو را از تصویر جدا می‌سازد این است که در ویدیو یک ارتباط زمانی (همراه با گذر زمان) بین فریم‌ها وجود دارد. بنابراین الگوریتمی مطلوب است که ارتباط بین

فریم‌ها و حرکت احتمالی اشیاء در فریم‌های متوالی را به عنوان یک ویژگی کلیدی در پردازش ویدئو در نظر بگیرد.



الف- تشخیص محتویات تصویر ب- تعیین محل اشیاء

ج- تقطیع معنایی د- تقطیع نمونه

شکل ۱: تجزیه و تحلیل تصویر [۱]

از آنجا که در یک ویدئو با حجم عظیمی از داده‌ها مواجه هستیم پردازش بلادرنگ با سرعت بالا و تاخیر زمانی کم و دقت قابل قبول یک چالش قابل توجه در انواع پردازش ویدئو از جمله تقطیع معنایی می‌باشد. از این رو در این مقاله ضمن بررسی ابعاد مختلف تقطیع معنایی ویدئو، مهمترین تحقیقات انجام شده سال‌های اخیر در این زمینه مرور شده و روشی برای رویارویی با چالش اشاره شده پیشنهاد می‌شود.

به طور کلی دو دسته روش برای تقطیع معنایی ویدئو وجود دارد. دسته اول روش‌های مبتنی بر تقطیع معنایی تصویر (تک فریم) است. این روش‌ها با هر فریم ویدئو مانند یک تصویر مجزا برخورد کرده و دنباله فریم‌های ویدئو را به صورت تک به تک و مجزا (بدون در نظر گرفتن ارتباط بین فریم‌ها) پردازش کرده و عمدتاً قادر به تقطیع معنایی بلادرنگ نمی‌باشند [۴، ۵، ۶]. دسته دیگر، روش‌هایی هستند که در آن‌ها از ارتباط بین فریم‌های پشت سر هم در ویدئو استفاده شده و توانایی استفاده مجدد از ویژگی‌ها و اطلاعات فریم‌های قبلی را دارا می‌باشند [۷، ۸، ۹].

در سال‌های اخیر شبکه‌های عصبی عمیق، عملکرد مناسبی در حوزه‌های مختلف پردازش تصویر و بینایی ماشین خود به نمایش گذاشته‌اند [۱۰، ۱۱]. شبکه‌های عصبی عمیق کانولوشنال[۲] (CNN) به طور قابل‌توجهی زمان پردازش رویکردهای سنتی را کاهش می‌دهند؛ چرا که روش‌های سنتی نیازمند زمان بیشتری برای پیش‌پردازش و استخراج ویژگی می‌باشند. از آن‌جا که در ویدیوها معمولاً تخمین مطمئنی از تقطیع معنایی یک فریم را می‌توان از تقطیع معنایی فریم‌های قبلی به دست آورد، استفاده از شبکه‌های عصبی عمیق بازگشتی در افزایش سرعت و دقت می‌تواند بسیار موثر واقع شود [۱۲].

استفاده از جریان نوری[۳] نیز در کاربردهای مختلف بینایی ماشین نظیر دسته‌بندی و تقطیع معنایی فریم‌های ویدئو مورد استفاده قرار می‌گیرد؛ چراکه حاوی اطلاعات مفیدی در رابطه با حرکت اشیاء در فریم‌های پشت سر هم بوده و می‌تواند به عنوان تخمینی از تغییر موقعیت نقاط در فریم‌های متوالی می‌باشد. کاربردهای متعددی برای جریان نوری در بینایی ماشین وجود دارد که برخی از این کاربردها در کنار استفاده از شبکه‌های عصبی عمیق بوده و به نوعی یک مرحله پیش پردازش می‌باشند [۱۳، ۱۴].

از آن‌جا که رسیدن به یک مصالحه بین دقت قابل قبول و سرعت مناسب (پردازش بلادرنگ) در تقطیع معنایی ویدئو یک چالش قابل توجه می‌باشد، در این مقاله چارچوبی ارائه شده است که به منظور غلبه بر چالش ذکر شده از یک شبکه عمیق از پیش آموزش دیده به منظور ارائه یک تقطیع معنایی اولیه از فریم‌ها و همچنین جریان نوری فریم‌های متوالی (جهت به دست آوردن میزان حرکت اشیاء در فریم‌های پشت سر هم) به عنوان ورودی‌های یک شبکه عمیق بازگشتی GRU[4] استفاده می‌شود. خروجی شبکه عمیق بازگشتی تقطیع معنایی فریم فعلی می‌باشد. (مطابق شکل ۴)

در ادامه این مقاله ابتدا در بخش دو پیشینه تحقیق مورد نقد و بررسی قرار گرفته است. سپس در بخش سه به تشریح روش پیشنهادی پرداخته شده و قسمت‌های مختلف آن مورد بررسی قرار گرفته است. در بخش چهار به بررسی ملاحظات پیاده‌سازی و آزمایش‌ها و تفسیر نتایج حاصل پرداخته شده و در پایان بخش پنج به بیان نتیجه‌گیری حاصل از این تحقیق اختصاص یافته است.

۲- پیشینه تحقیق

در این بخش یک مرور کلی روی تحقیقاتی که به منظور تقطیع معنایی تصویر و ویدئو صورت گرفته است ارائه می‌شود.

روش‌های مبتنی بر گراف

در پردازش تصویر از این روش‌ها برای نمایش ارتباط بین پیکسل‌های تصویر استفاده می‌شود؛ به این صورت که پیکسل‌های تصویر به صورت گره‌های گراف و ارتباطات بین پیکسل‌ها به صورت ارتباطات وزن دار نمایش داده می‌شوند. بنابراین از روش‌های مبتنی بر گراف به منظور کشف ازتباطات

بین پیکسل‌ها و یافتن اشیاء و ساختارهای موجود در تصویر یا فریم ویدئو استفاده می‌شود [۱۵، ۱۶، ۱۷].

روش‌های مبتنی بر گام تصادفی (Random Walk)

الگوریتم‌های Random Walk در سیستم‌های مبتنی بر گراف استفاده می‌شوند. ساختار این الگوریتم‌ها به این صورت است که نقاط مختلف اشیاء تصویر انتخاب شده و سپس احتمال رسیدن هر نقطه به نقاط دیگر محاسبه می‌شود. به منظور تقطیع معنایی، دسته مربوط به هر پیکسل با استفاده از احتمال حداکثر محاسبه می‌شود [۱۸، ۱۹].

روش‌های مبتنی بر ماشین بردار پشتیبان[۵] (SVM)

SVM یک دسته بندی کننده مناسب برای کاربردهای مختلف با تعداد دسته‌های محدود است. به منظور افزایش قدرت ماشین بردار پشتیبان در تفکیک داده‌های هم‌پوشان از هسته‌های مختلف با قدرت تفکیک کنندگی بیشتر می‌توان استفاده کرد. با توجه به قدرت ماشین بردار پشتیبان در دسته‌بندی پیکسل‌های تصویر و فریم‌های ویدئو، در روش‌های مختلفی از این مدل به منظور تقطیع معنایی استفاده شده است [۲۰، ۲۱].

روش‌های مبتنی بر میدان تصادفی مارکوف[۶] (MRF)

MRF یکی از زیرمجموعه‌های مدل‌های احتمالاتی می‌باشد که ارتباط بین متغیرهای تصادفی را با استفاده از یک گراف غیرمستقیم نمایش می‌دهد و توانایی نمایش و پشتیبانی از وابستگی‌ها را دارا می‌باشد. با توجه به ساختار و قابلیت MRF از این مدل در برخی تحقیقات در زمینه تقطیع معنایی استفاده شده است [۲۲، ۲۳].

روش‌های مبتنی بر میدان تصادفی شرطی[۷] (CRF)

CRF یکی دیگر از انواع مدل‌های احتمالاتی هستند که معمولاً در مسائل بازشناسی الگو استفاده می‌شود. این مدل توانایی تعیین برچسب یک نمونه بدون در نظر گرفتن همسایگی‌های آن را دارا می‌باشد. در مسائل بینایی ماشین معمولاً از CRF به منظور شناسایی اشیاء و قطعه‌بندی استفاده می‌شود [۱، ۲۴، ۲۵].

روش‌های مبتنی بر شبکه‌های عصبی

استفاده از شبکه‌های عصبی یکی از گسترده‌ترین راه‌های مدل‌سازی در یادگیری ماشین است. تا کنون تحقیقات مختلفی ارائه شده است که در آن‌ها از شبکه‌های عصبی به منظور تقطیع معنایی بهره گرفته شده است [۲۶، ۲۷، ۲۸].

روش‌های مبتنی بر شبکه‌های عصبی عمیق

در سال‌های اخیر شبکه‌های عصبی عمیق به عنوان رایج‌ترین ساختار برای تقطیع معنایی شناخته می‌شوند؛ بالاخص پس از موفقیت‌های اخیر شبکه‌های عمیق مانند شبکه‌های عصبی کانولوشنال (CNN)، شبکه‌های LSTM[8] [۲۹] و شبکه‌های GRU [۳۰].

روش ارائه شده توسط لانگ و همکاران [۳۱] یک روش پیش‌گام برای تقطیع معنایی تصویر است که در این روش لایه انتهایی شبکه عمیق CNN با لایه‌های کانولوشنال جایگزین شده است. در بسیاری از روش‌ها از ساختار مذکور الگو گرفته شده است. به عنوان مثال در برخی روش‌ها به منظور بهبود دقت از ترکیب شبکه عصبی کانولوشنال با CRF و یا ترکیب ویژگی‌های سطح پایین و سطح بالا استفاده شده است [۳۲] .

کندل و همکاران در [۳۳] مدل SegNet را معرفی کرده‌اند که یک معماری مبتنی بر رمزگذار-رمزگشا[۹] می‌باشد. ژائو و همکاران در [۳۴] PSPNet را ارائه کردند که از تجمع مکانی هرم برای ترکیب اطلاعات سراسری و محل استفاده می‌کند. همه این متدها صرفاً بر مبنای تقطیع معنایی تصویر بوده و برای ویدئو هر فریم را به صورت مجزا پردازش کرده و ارتباط بین فریم‌ها را در نظر نمی‌گیرند.

تقطیع معنایی ویدئو با استفاده از شبکه‌های عصبی عمیق

روش‌های مختلفی برای تقطیع معنایی ویدئو ارائه شده است که در هریک از آن‌ها تلاش گسترده‌ای برای ایجاد یک تعادل بین دقت و سرعت صورت پذیرفته است [۳۵، ۳۶]. در برخی روش‌ها یک شبکه عصبی عمیق به طور مکرر روی فریم‌های مختلف عمل کرده و ویژگی‌های به دست آمده در لایه‌های بالاتر را به اشتراک می‌گذارند [۳۷]. گرچه این روش ها روی هر فریم مجزا به دقت مناسبی دست می‌یابند اما محاسبات زیاد و عدم استفاده از وابستگی بین فریم های پشت سر هم در یک ویدئو ایراد این روش‌ها می‌باشد.

دسته دیگری از روش‌ها از وابستگی بین فریم‌های پشت سر هم به منظور استفاده مجدد از ویژگی‌های استخراج شده و اطلاعات پردازش فریم‌های قبلی و بهبود سرعت پردازش بهره می‌برند [۲، ۳، ۳۸].

یکی از چالش‌های موجود در این دسته از روش‌ها چگونگی انتشار صحیح اطلاعات در طول زمان می‌باشد. به همین منظور می‌توان از ویژگی‌های سطح بالا استخراج شده از لایه‌های عمیق و انتشار آن در فریم‌های بعدی استفاده کردد [۳۹، ۴۰].

با استفاده از شبکه‌های کانولوشنال سه‌بعدی می‌توان در یکی از ابعاد شبکه گذر زمان در ویدئو و تغییرات فریم‌ها در طول زمان

را به شبکه آموزش داد. اما این شبکه‌ها بسیار پیچیده بوده و حجم محاسبات در آنها بسیار زیاد است [۴۱، ۴۲، ۴۳].

یکی از راه حل‌های بهره‌گیری از ویژگی‌ها و اطلاعات فریم‌های قبلی استفاده از شبکه‌های عصبی بازگشتی است. شبکه‌های LSTM و GRU از معروف‌ترین شبکه‌های عمیق بازگشتی هستند که کارایی مناسبی برای کار روی داده‌های سلسله مراتبی نظیر فریم‌های ویدئو دارند [۱۲، ۴۴].

فیاض و همکاران در [۱۲] مدلی برای تقطیع معنایی ویدئو معرفی کرده‌اند که در آن یک LSTM روی ویژگی‌های تولید شده برای هر فریم (توسط شبکه CNN) عمل کرده و روی هریک از فریم‌های ویدئو عملیات تقطیع معنایی را انجام می‌دهد. در روش ارائه شده توسط چاندرا و همکاران ابتدا کانال‌های رنگی هر فریم و جریان نوری مربوط به آن به یک شبکه FCN[10] داده می‌شود و سپس خروجی شبکه FCN به یک شبکه LSTM داده شده و تقطیع معنایی انجام می‌پذیرد [۴۴].

به منظور استفاده از شبکه‌های LSTM و GRU برای تقطیع معنایی دنباله فریم‌های ویدئو، Conv-LSTM توسط هاندا و همکاران[۴۵] و Conv-GRU توسط بالاس و همکاران [۴۶] معرفی شده است. در این مدل‌های بازگشتی کانولوشنال، وضعیت‌ها و دروازه‌های LSTM و GRU، تنسورهای[11] سه بعدی هستند و بردارهای وزن نیز به صورت کانولوشن‌های دو بعدی می‌باشند. از مدلConv-GRU در روش پیشنهادی این مقاله استفاده می‌شود.

به منظور کاهش زمان پردازش و بعضاً افزایش دقت در روش‌های مبتنی بر شبکه‌های عمیق راه حل‌های مختلفی وجود دارد. یک راه استفاده از مفهوم فریم کلیدی و انتخاب تعدادی از فریم‌ها به عنوان فریم کلیدی و انتشار و استفاده از ویژگی‌ها در فریم‌های بین آن فریم کلیدی و فریم کلیدی بعدی می‌باشد.

محاسنی و همکاران در [۳۶] روشی ارائه کرده‌اند که بر اساس فرایند تصمیم‌گیری مارکوف برای انتخاب فریم‌های کلیدی عمل می‌کند و ویژگی‌های فریم‌های کلیدی از طریق درون‌یابی به دیگر فریم‌ها انتشار می‌یابد.

در تحقیقات اخیر روش‌هایی پیشنهاد شده است که با استفاده از روش‌هایی مانند جریان نوری نتیجه تقطیع معنایی به فریم‌های دیگر تعمیم داده می‌شود. به عنوان مثال توسط گاد و همکاران روشی ارائه شده است که از ترکیب ویژگی‌های انتقال یافته از فریم‌های قبل و ویژگی‌های فریم فعلی برای تقطیع معنایی استفاده می‌شود [۴۷]. همچنین ژو و همکاران در [۲] مدلی تحت عنوان DFF[12] ارائه کرده‌اند که در آن ویژگی‌های سطح بالای فریم‌های کلیدی با جریان نوری یادگیری شده توسط FlowNet [۱۳] جمع‌آوری شده و نتایج قبلی بهبود داده شده است. هرچند تخمین جریان نوری با استفاده از Flownet موجب افزایش محاسبات و کاهش سرعت می‌شود، اما استفاده از فریم کلیدی در این تحقیق موجب بهبود سرعت می‌شود. روش FlowNet2 [۱۴] نیز توسط دنتون و همکاران بر پایه FlowNet توسعه داده شده است. البته در برخی تحقیقات، تخمین جریان نوری با استفاده از روش‌های یادگیری عمیق انجام نشده است؛ بلکه به صورت سنتی و با محاسبه ماتریس اختلاف نقاط دو فریم پشت سر هم صورت پذیرفته است. بنابراین بهره گیری از اطلاعات فریم‌های قبلی در تقطیع معنایی با استفاده از جریان نوری بسیار مورد توجه می‌باشد و در روش ارائه شده نیز از این ایده استفاده می‌شود. البته در روش پیشنهادی این مقاله جریان نوری فریم‌های پشت سر هم با استفاده از یک شبکه از قبل آموزش دیده شده محاسبه می‌شود که این امر منجر به استحکام مدل در تغییرات ناگهانی و بهبود کارایی می‌شود.

۳- روش پیشنهادی

همانطور که اشاره شد به طور کلی احتمال این‌که فریم‌های پشت سر هم در یک ویدئو ارتباط و مشابهت محتوایی قابل توجهی داشته باشند زیاد است (شکل ۲).

شکل۲: وابستگی و مشابهت اجزاء فریم‌های پشت سر هم در یک ویدئو

در روش ارائه شده تلاش بر این است که در تقطیع معنایی هر فریم ویدئو تا حد امکان از اطلاعات پردازش شده در فریم‌های قبلی استفاده شده و دقت و سرعت تقطیع معنایی به حد قابل قبولی برسد. برای این منظور، ابتدا جریان نوری حاصل از فریم فعلی و فریم قبلی محاسبه شده و خروجی تقطیع معنایی فریم قبلی در راستای این جریان نوری لغزش داده می‌شود. علت این امر آن است که با لغزاندن تقطیع معنایی فریم قبلی در راستای جریان نوری می‌توان به یک تخمین از تقطیع معنایی فریم فعلی (به خصوص در اشیاء تعییر نیافته نسبت به فریم قبل) دست یافت. از طرفی به منظور دستیابی به یک تخمین از تقطیع معنایی فریم فعلی (به خصوص در نواحی و محتویات

تغییر یافته نسبت به فریم قبل) از یک شبکه کانولوشنال ساده و از پیش آموزش دیده استفاده می‌شود. مقدار حاصل از لغزش یافته تقطیع معنایی فریم قبلی و همچنین مقدار حاصل از شبکه کانولوشنال اجرا شده روی فریم فعلی، به یک شبکه عصبی عمیق بازگشتی GRU داده شده و خروجی نهایی به دست می‌آید. (طبق شکل ۴) بنابراین در روش پیشنهادی از اطلاعات نهفته در پردازش و تقطیع معنایی فریم‌های قبلی استفاده حداکثری به عمل می‌آید. اجزاء الگوریتم پیشنهادی مطابق شکل زیر می باشد:

گام ۱- ورود فریم t
گام ۲- جریان نوری فریم قبلی (f_{t-1})و فریم فعلی (f_t) $\leftarrow O_t$
گام ۳- لغزش خروجی تقطیع معنایی فریم قبل (S_{t-1})با تابع لغزش $\phi \leftarrow m_t$
گام ۴- تخمین تقطیع معنایی فریم فعلی با شبکه CNN$\leftarrow C_t$
گام ۵- تولید خروجی تقطیع معنایی با شبکه GRU$\leftarrow S_t$
گام ۶- محاسبه خطا و انتشار به عقب
گام ۷- آموزش پارامترهای شبکه GRU و تنظیم دقیق (fine-tune) شبکه CNN

شکل ۳: الگوریتم روش پیشنهادی

مطابق الگوریتم، در روش پیشنهادی پس از ورود یک فریم (f_t) و محاسبه جریان نوری حاصل از فریم قبلی و فریم فعلی (طبق گام‌های ۱ و۲ الگوریتم) ابتدا خروجی تقطیع معنایی فریم قبلی (S_{t-1}) در امتداد جریان نوری حاصل از فریم قبلی (f_{t-1}) و فریم فعلی (f_t) لغزش داده می‌شود (با استفاده از عمل لغزش و مطابق گام ۳ الگوریتم). هدف از این کار دستیابی به یک تخمین از تقطیع معنایی فریم فعلی (S_t) مخصوصاً در بخش‌های تغییر نیافته نسبت به فریم قبل می‌باشد. این عمل با استفاده از رابطه زیر انجام می‌شود.

$$m_t=\phi(S_{t-1},O_t) \quad (1)$$

m_t به عنوان یک حالت مخفی به GRU وارد می‌شود. ورودی دیگر GRU، C_t است که حاصل اعمال CNN روی فریم فعلی (f_t) بوده و تخمینی از تقطیع معنایی فریم فعلی (طبق گام ۴) مخصوصا در بخش‌های تغییر یافته نسبت به فریم قبل می‌باشد.
خروجی GRU (گام ۵ الگوریتم)، تقطیع معنایی فریم t با توجه به ترکیب ویژگی‌های فریم فعلی (ویژگی‌های مکانی) و لغزش یافته خروجی فریم قبلی (وابستگی زمانی فریم‌ها) می‌باشد. کلیات روش ارائه شده در شکل ۴ نمایش داده شده است.
ساختار پیشنهادی حاوی اجزاء و قسمت‌های مختلفی می‌باشد که در ادامه به معرفی قسمت‌های مختلف آن همراه با جزئیات هر قسمت پرداخته می‌شود.

۳-۱- ماژول محاسبه جریان نوری

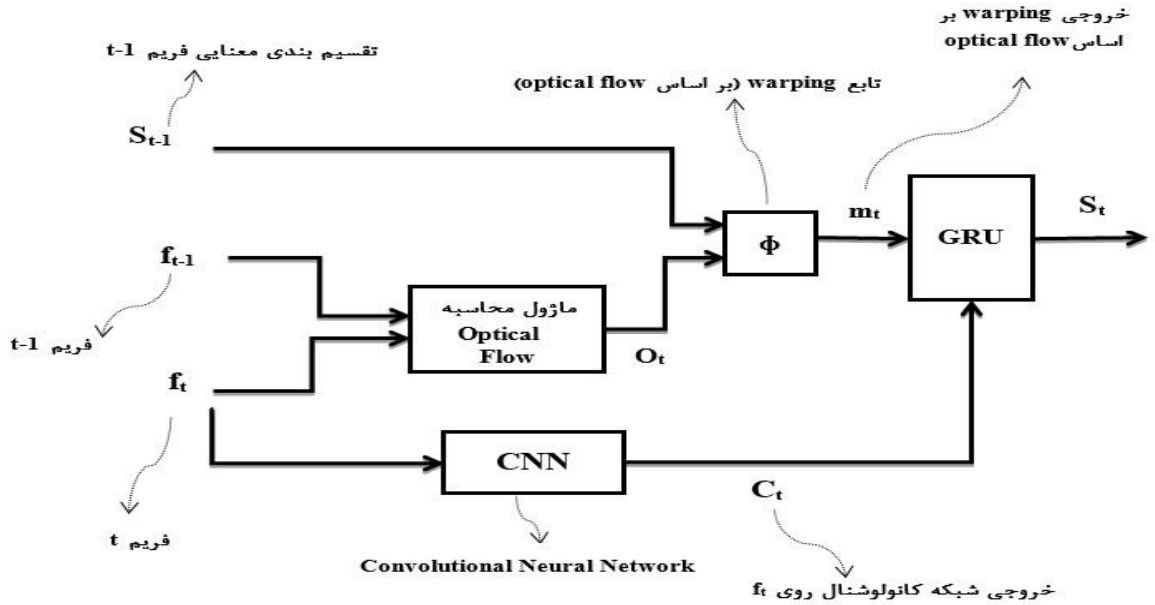
محاسبه جریان نوری بین فریم فعلی (f_t) و فریم قبلی (f_{t-1}) توسط این ماژول انجام می‌شود. برای این ماژول انتخاب‌های مختلفی وجود دارد که در اینجا به دو مورد اشاره می‌شود:

- محاسبه جریان نوری به روش کلاسیک
- محاسبه و تخمین جریان نوری از طریق یک شبکه عصبی کانولوشنال کوچک

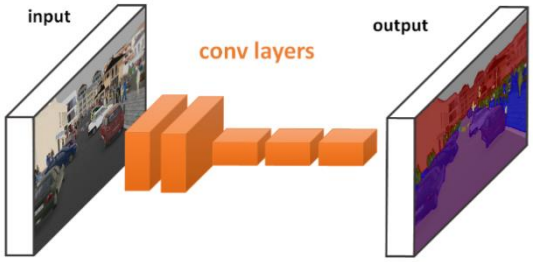
برای تخمین جریان نوری از طریق یک شبکه عصبی کانولوشنال تاکنون تحقیقات مختلفی انجام شده و ساختارها و شبکه‌های متعددی پیشنهاد و ارائه شده است. در این تحقیق برای تخمین جریان نوری هر دو فریم پشت سر هم از ساختارهای FlowNet و FlowNet2 [۴۸] استفاده می‌شود.
m_t به عنوان یک حالت مخفی به GRU وارد می‌شود. ورودی دیگر GRU، C_t است که حاصل اعمال CNN روی فریم فعلی (f_t) بوده و تخمینی از تقطیع معنایی فریم فعلی (طبق گام ۴) مخصوصا در بخش‌های تغییر یافته نسبت به فریم قبل می‌باشد.
خروجی GRU (گام ۵ الگوریتم)، تقطیع معنایی فریم t با توجه به ترکیب ویژگی‌های فریم فعلی (ویژگی‌های مکانی) و لغزش یافته خروجی فریم قبلی (وابستگی زمانی فریم‌ها) می‌باشد. کلیات روش ارائه شده در شکل ۴ نمایش داده شده است.
ساختار پیشنهادی حاوی اجزاء و قسمت‌های مختلفی می‌باشد که در ادامه به معرفی قسمت‌های مختلف آن همراه با جزئیات هر قسمت پرداخته می‌شود.
علت استفاده از این تابع و انجام عمل لغزش نقاط S_{t-1} آن است که معمولا موجودیت ها و اجزاء فریم‌های متوالی مقدار اندکی حرکت و تغییر دارند؛ مگر اینکه تغییر ناگهانی در صحنه رخ داده و دو فریم متوالی بسیار متفاوت باشند. با این استدلال می‌توان امید داشت که با این عمل می‌توان تقریبی از تقطیع معنایی فریم فعلی به دست آورد. بنابراین m_t حاصل از اعمال تابع ϕ (لغزش نقاط در راستای جریان نوری) روی S_{t-1} می‌باشد. m_t یکی از ورودی‌های شبکه GRU می‌باشد.

۳-۳- شبکه عصبی کانولوشنال برای تقطیع معنایی

این شبکه کانولوشنال تخمینی از تقطیع معنایی اجزاء موجود در یک فریم (فارغ از ارتباط با فریم‌های قبلی) می‌باشد. در واقع خروجی این CNN از قبل آموزش دیده تا حد زیادی برچسب صحیح هر نقطه فریم فعلی(f_t) را مشخص می‌کند. خروجی این شبکه به عنوان یکی دیگر از ورودی‌های GRU می‌باشد.



شکل ۴: نمای کلی روش پیشنهادی



شکل ۵: شبکه عمیق کانولوشنال

۴-۳- شبکه عمیق بازگشتی GRU

به منظور اتصال خروجی‌های تقطیع معنایی فریم‌های مختلف، از یک نسخه کانولوشنال تغییر یافته GRU استفاده می‌شود. این GRU اطلاعات تقطیع معنایی لغزش یافته از S_{t-1} و همچنین خروجی تخمین تقطیع معنایی فریم فعلی (C_t) را برای تقطیع معنایی فریم جدید مورد استفاده قرار می‌دهد.

در اینجا باتوجه به ورودی‌های شبکه GRU نوع و روابط موجود در آن تعیین می‌شود. از آن‌جا که ورودی‌های (m_t و C_t) به صورت بردار نبوده و تنسور هستند شبکه GRU استفاده شده در این تحقیق از نوع ConvGRU [۴۶] می‌باشد. خروجی GRU با یک لایه کانولوشنال و لایه غیرخطی softmax پردازش شده و خروجی تقطیع معنایی به دست می‌آید.

به منظور به دست آوردن خروجی تقطیع معنایی هر فریم ویدئو، GRU از دو ورودی c_t و m_t و دو دروازه به روزرسانی (z_t) و فراموشی (r_t) استفاده می‌کند. وضعیت‌ها و دروازه‌ها تنسورهای سه بعدی هستند و قابلیت تشخیص الگوهای مکانی-زمانی نهفته در دنباله فریم‌های ویدئو را دارا می‌باشند. اجزاء مختلف GRU به صورت زیر می‌باشند.

$$m_t = \phi(S_{t-1}, O_t) \quad (۲)$$

$$r_t = \sigma(c_t * w_{cr} + m_t * w_{mr} + b_r) \quad (۳)$$

$$z_t = \sigma(c_t * w_{cz} + m_t * w_{mz} + b_z) \quad (۴)$$

$$\tilde{S}_t = tanh(c_t * w_{c\tilde{S}} + (r_t \odot m_t) * w_{\tilde{S}m} + b_{\tilde{S}}) \quad (۵)$$

$$S_t = softmax((1 - z_t) \odot m_t + z_t \odot \tilde{S}_t) \quad (۶)$$

در روابط فوق $\odot$ نشان دهنده ضرب المانی[۱۳] و $*$ بیان‌گر عملیات کانولوشنال است، σ تابع سیگموئید و w‌ها وزن‌ها و b بیانگر بایاس می‌باشد، S_t یک ترکیب وزن‌دار از m_t و حافظه $\tilde{S}_t$ است، دروازه به روزرسانی z_t نشان دهنده این است که چه مقدار از حافظه در وضعیت جدید مشارکت دارد، و دروازه بازنشانی r_t تاثیر m_t روی $\tilde{S}_t$ را کنترل می‌کند؛ یعنی این‌که چه مقدار از وضعیت قبلی درون حافظه باقی مانده است. اگر r_t نزدیک به صفر باشد، GRU وضعیت قبلی S_{t-1} و متعاقباً m_t را فراموش می‌کند.

۴- آزمایش‌های تجربی

به منظور ارزیابی روش پیشنهادی، آزمایش‌های مختلفی انجام شده است. در این آزمایش‌ها از مجموعه داده‌های مختص تقطیع معنایی ویدئو که حاوی دنباله تصاویر شهری هستند استفاده شده است. ارزیابی روش‌ها نیز با استفاده از معیارهای ارزیابی مختص تقطیع معنایی صورت گرفته است.

۴-۱- مجموعه داده‌ها

در این تحقیق، از دو مجموعه داده Cityscapes [۴۹] و Camvid [۵۰] که حاوی ویدئوها و فریم‌های تصاویر شهری می‌باشند استفاده می‌شود.

مجموعه داده Cityscapes حاوی تصاویری از ۵۰ شهر مختلف می‌باشد. این مجموعه داده حاوی ۲۹۷۵ دنباله تصاویر آموزش، ۵۰۰ دنباله تصاویر اعتبارسنجی و ۱۵۲۵ دنباله تصاویر تست می‌باشد. در این مجموعه داده هر دنباله حاوی ۳۰ فریم با اندازه ۲۰۴۸ × ۱۰۲۴ نقطه است اما می‌توان در مرحله آموزش یک مدل با ساختار شبکه عصبی عمیق با استفاده از یک مرحله پیش پردازش اندازه تصاویر را ۵۱۲ × ۵۱۲ پیکسل در نظر گرفت. مدت زمان هر دنباله ۱.۸ ثانیه . به این ترتیب فرکانس نمونه برداری در این مجموعه داده حدود ۱۶.۶ فریم بر ثانیه بوده و آستانه زمانی پردازش بلادرنگ در این مجموعه داده حدود ۱۶.۶ فریم بر ثانیه است. مجموعه داده Cityscapes دارای ۱۹ دسته مختلف می‌باشد.

Camvid نیز یکی از مجموعه داده‌های حاوی دنباله فریم‌های ویدئویی است که به منظور تقطیع معنایی مورد استفاده قرار می‌گیرد. این مجموعه داده از ۷۰۰ دنباله تصویر و ۴۰۰۰۰ فریم تشکیل شده است که ۳۶۷ دنباله تصاویر آموزش، ۱۰۰ دنباله تصاویر اعتبارسنجی و ۲۳۳ دنباله تصاویر تست می‌باشند. در این مجموعه داده اندازه هر فریم ۴۸۰×۶۴۰ پیکسل می‌باشد. تعداد دسته‌ها در Camvid برابر با ۱۱ دسته می‌باشد.

مشخصات کلی مجموعه داده‌هایCityscapes و Camvid در جداول زیر بیان شده است.

جدول ۱: مشخصات مجموعه داده Cityscapes

مجموعه داده Citysapes	
اندازه تصاویر	۱۰۲۴×۲۰۴۸
تعداد تصاویر آموزش	۲۹۷۵ دنباله
تعداد تصاویر اعتبارسنجی	۵۰۰ دنباله
تعداد تصاویر آزمون	۱۵۲۵ دنباله
تعداد دسته‌ها	۱۹ دسته

جدول ۲: مشخصات مجموعه داده Camvid

مجموعه داده Camvid	
اندازه تصاویر	۶۴۰×۴۸۰
تعداد کل تصاویر	۷۰۰ دنباله تصویر
تعداد تصاویر آموزش	۳۶۷ دنباله
تعداد تصاویر اعتبارسنجی	۱۰۰ دنباله
تعداد تصاویر آزمون	۲۳۳ دنباله
تعداد دسته‌ها	۱۱ دسته

۴-۲- معیار ارزیابی

معیار ارزیابی استفاده شده در آزمایش‌ها معیار mIoU[14] می‌باشد. این معیار ارزیابی که به آن شاخص جاکارد نیز گفته می شود، در واقع روشی برای تعیین میزان همپوشانی صحیح یا هدف و خروجی تقطیع معنایی یک تصویر یا فریم ویدئویی می‌باشد. فرمول معیار ارزیابی mIoU به صورت زیر است:

$$mIoU = \frac{tp}{tp + fp + fn} \qquad (۵-۱)$$

در رابطه فوق tp عبارتست از نرخ مثبت صحیح، fp نرخ مثبت کاذب بوده و fn نیز عبارتست از نرخ منفی کاذب.

۴-۳- جزئیات پیاده سازی و تنظیم پارامترها

از آنجا که با در نظر گرفتن محدودیت زمان و حافظه مورد نیاز، آموزش شبکه‌های تقطیع معنایی از پایه و ابتدا امکان پذیر نیست به جز شبکه بازگشتی GRU، برای بخش‌های مختلف هر ساختار از مدل‌هایی با وزن‌های از پیش آموزش دیده استفاده می‌شود.

برای شبکه کانولوشنال مدل‌های از قبل آموزش دیده شده روی مجموعه داده Cityscapes مورد استفاده واقع می‌شود.

آزمایش‌ها با استفاده از پردازنده گرافیکی تسلا مدل P4 و در بستر تنسورفلو[15] انجام شده است. برای شبکه CNN از پیش آموزش دیده از سه مدل PSP [۳۴]، FCN-8 [۳۱] و Dilation [۵۱] استفاده شده است. همچنین برای ساختار از پیش آموزش دیده تخمین جریان نوری نیز از سه ساختار FlowNet، FlowNet2 و LiteFlowNet [۵۲] استفاده شده است. در این مقاله به منظور آموزش شبکه عمیق بازگشتی از روش کاهش گرادیان تصادفی[16] و پس انتشار خطا استفاده شده است. مقدار ممنتوم[17] برابر با ۰.۹، مقدار کاهش وزن[18] برابر با ۰.۰۰۰۵، نرخ یادگیری برابر با ۰.۰۱ و اندازه دسته‌ها برای مجموعه داده‌های Cityscapes و Camvid به ترتیب برابر با ۴۸ و ۲۴ در نظر گرفته شده است.

۴-۳- نتایج حاصل

در آزمایش‌های این بخش دقت و سرعت روش ارائه شده با دو دسته از روش‌ها مورد مقایسه قرار می‌گیرد. دسته اول روش‌های تقطیع معنایی فریم به فریم است که بدون توجه به ارتباط موجود بین فریم‌های ویدئو عمل می‌کنند و دسته دوم نیز روش‌هایی هستند که در آن‌ها وابستگی بین فریم‌ها در نظر گرفته شده و در تقطیع معنایی یک فریم، از اطلاعات فریم‌های قبلی استفاده می‌شود. هریک از ساختارهای مختلف روش پیشنهادی نیز مطابق جدول زیر می‌باشد.

جدول ۳: ساختارهای مختلف روش پیشنهادی

ساختار روش پیشنهادی	کانولوشنال	تخمین جریان نوری
Our Method (PSP+FlowNet2)	PSP	FlowNet2
Our Method (PSP+FlowNet)	PSP	FlowNet
Our Method (PSP+LiteFlowNet)	PSP	LiteFlowNet
Our Method (Dilation+FlowNet2)	Dilation	FlowNet2
Our Method (Dilation+FlowNet)	Dilation	FlowNet
Our Method (Dilation+LiteFlowNet)	Dilation	LiteFlowNet
Our Method (FCN-8s+FlowNet)	FCN-8s	FlowNet

۴-۴- ارزیابی نتایج آزمایش‌ها

در جدول‌های۴ و ۵ دقت و سرعت روش‌ها روی مجموعه داده‌های Cityscapes و Camvid آورده شده است.

جدول ۴: نتایج روی مجموعه داده Cityscapes

روش	دقت (mIoU)	سرعت (FPS[19])
تقطیع معنایی فریم به فریم		
HRNet-OCR [۷]	۸۵.۱	نامشخص
EfficientPS [۸]	۸۴.۲	۲.۲
Panoptic-DeepLab [۹]	۸۴.۲	۳.۱
PSPNet [۳۴]	۷۷	۰.۷۸
FCN-8s [۳۱]	۶۵.۳	۲
تقطیع معنایی جریان فریم‌ها		
Our Method (PSP+FlowNet2)	**۸۳.۱**	**۲۱.۱**
SFNet [۵۳]	۸۰.۴	۲۵.۷
Our Method (PSP+ FlowNet)	**۷۷.۴**	**۲۷.۴**
HyperSeg [۵۴]	۷۵.۸	۳۶.۹
STDC [۵]	۷۶.۸	۳۲.۴
Our Method (Dilation+LiteFlowNet)	**۷۰.۳**	**۳۳.۵**
GRFP (Dilation10+FlowNet2) [۵۵]	۶۹.۵	۱۹
BiseNet [۵۶]	۶۹	۳۱.۸
Our Method (FCN-8s + LiteFlowNet)	**۶۸.۱**	**۳۴**
Clockwork [۳۹]	۶۷.۷	نامشخص
EDANet [۵۷]	۶۷.۳	۳۰

جدول ۵: نتایج دقت روی مجموعه داده Camvid

تقطیع معنایی فریم به فریم		
DeepLabV3Plus + SDCNetAug [۵۸]	۸۱.۷	۵
ETC-Mobile [۵۹]	۷۶.۳	۷.۱
VideoGCRF [۴۴]	۷۵.۲	۶.۵
تقطیع معنایی جریان فریم‌ها		
DDRNet-23 [۶۰]	۸۰.۶	۳۲
Our Method (PSP+FlowNet2)	**۷۹.۸**	**۲۲**
HyperSeg [۵۴]	۷۹.۱	۱۸.۷
BiSeNet V2-Large [۶۱]	۷۸.۵	۳۲.۷
BiSeNet V2 [۶۱]	۷۶.۷	۳۳.۵
Our Method (Dilation+FlowNet2)	**۷۳.۳**	**۳۱.۵**
DeepLabv2-CRF [۶۲]	۷۰.۴	۳۰.۴
Netwarp[۴۷]	۷۰.۳	۳۱
Our Method (Dilation+FlowNet)	**۶۹.۹**	**۳۶.۳**
GRFP [۵۵]	۶۹.۵	۳۲.۵

با توجه به نتایج جدول‌های ۴ و ۵ موارد زیر قابل استنباط است:

۱- به طور کلی روش‌های مبتنی بر پردازش جریان فریم‌ها دارای کفیت (ترکیب دقت و سرعت) بهتری نسبت به روش‌های تقطیع معنایی فریم به فریم می‌باشند؛ زیرا در این روش‌ها از اطلاعات تقطیع معنایی فریم‌های قبلی نیز استفاده می‌شود.

۲- هرچند برخی روش‌هایی که تقطیع معنایی را به صورت فریم به فریم انجام می‌دهند از دقت بالایی برخوردارند اما سرعت پردازش در این روش‌ها مناسب نیست؛ چراکه در این روش‌ها پردازش بلادرنگ مطرح نبوده و به طور کلی این روش‌ها برای تقطیع معنایی تصویر طراحی شده‌اند و فریم‌های پشت سر هم را به عنوان تصاویر مجزا در نظر می‌گیرند.

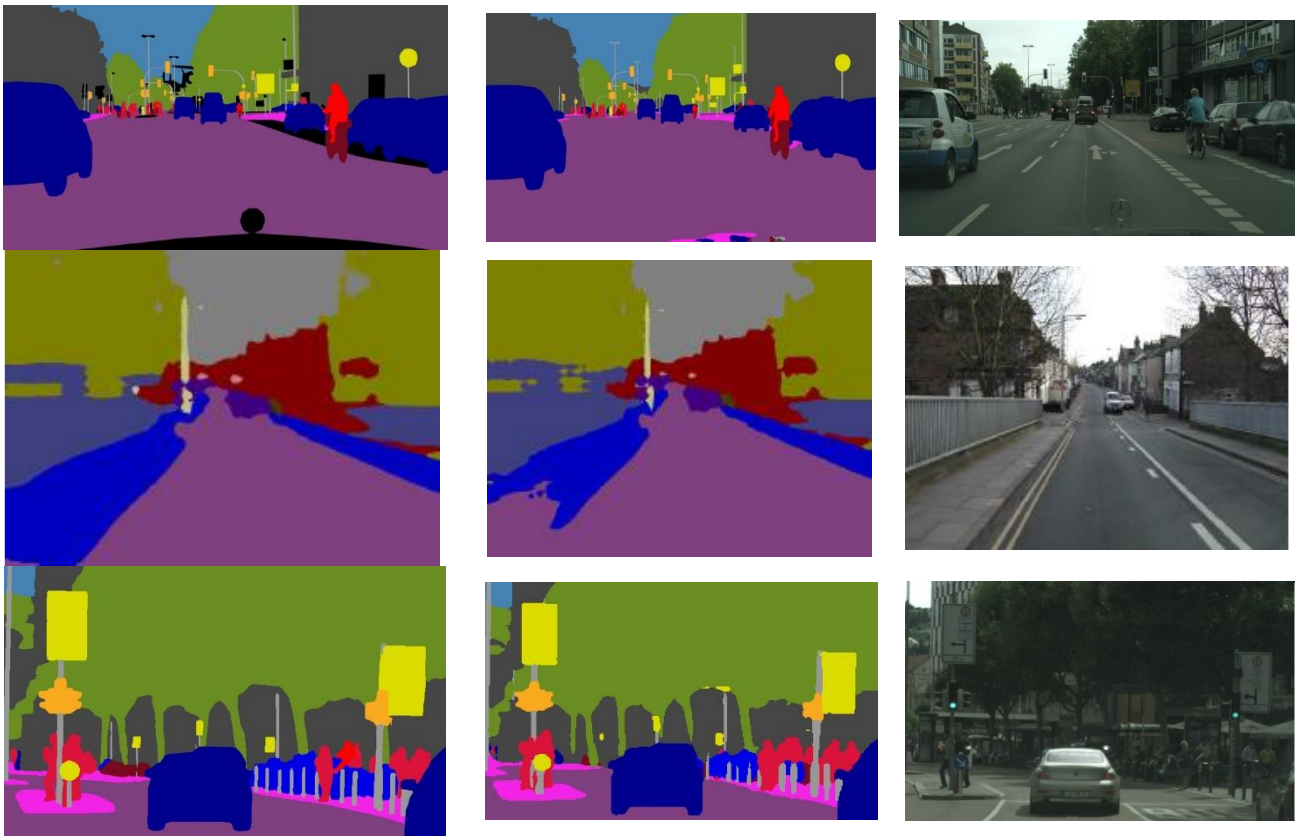
۳- با مقایسه ساختارهای مختلف روش پیشنهادی مشاهده می‌شود که هرچه شبکه کانولوشنال استفاده شده و شبکه تخمین جریان نوری بزرگ‌تر و دقیق‌تر باشند با افزایش دقت و کاهش سرعت مواجه می‌شویم. به عنوان مثال استفاده از شبکه PSP و NetFlow2، موجب افزایش دقت روی مجموعه داده Cityscapes تا ۸۳.۱٪ و کاهش سرعت به ۲۱.۱ فریم بر ثانیه می‌شود. استفاده از همین ساختار روی مجموعه داده Camvid موجب افزایش دقت تا ۷۹.۸٪ و کاهش سرعت تا ۲۲ فریم بر ثانیه می‌شود.

۴- از طرفی هرچه شبکه FCN و شبکه تخمین جریان نوری ساده‌تر و کوچک‌تر باشند، دقت کل در روش پیشنهادی کاهش یافته و سرعت تقطیع معنایی افزایش می‌یابد. به طوری که کمترین دقت و بیشترین سرعت روش پیشنهادی روی مجموعه داده Cityscapes مربوط به ترکیب FCN-8s (به عنوان شبکه کانولوشنال) و FlowNet (به عنوان شبکه تخمین جریان نوری) و همچنین کمترین میزان دقت روش پیشنهادی روی مجموعه داده Camvid مربوط به ترکیب Dilation و LiteFlowNet می‌باشد.

* با بررسی دقیق موارد ۳ و ۴ می‌توان نتیجه گرفت هم نوع ساختار کانولوشنال و هم نوع ساختار تخمین جریان نوری روی دقت روش تاثیرگذار می‌باشند. بنابراین با انتخاب نوع ساختار هریک می‌توان به دقت و سرعت مورد نظر دست یافت.

۵- با بررسی نتایج مجموعه داده CityScapes مشاهده می‌شود که تاثیر نوع شبکه کانولوشنال و دقیق‌تر بودن این شبکه (همانند شبکه PSP) تاثیر بیشتری روی دقت دارد؛ چرا که این مجموعه داده دارای تعداد دسته‌ها و انواع اشیاء بیشتری بوده و احتمال تکرار و وجود یک شیء در فریم‌های مختلف کمتر است. بنابراین انتخاب مناسب شبکه کانولوشنال تخمین تقطیع معنایی، نسبت به دقیق‌تر بودن نوع ساختار تخمین جریان نوری تاثیر بیشتری روی دقت تقطیع معنایی دنباله فریم‌ها دارد.

در شکل ۶ خروجی تقطیع معنایی روش ارائه شده روی چند فریم



شکل ۶: تصویر اصلی، خروجی روش SFNet، خروجی روش ارائه شده (PSP+FlowNet2)

از مجموعه داده Cityscapes نمایش داده شده و با خروجی حاصل از روش SFNet مقایسه شده است.

۵- نتیجه گیری

استفاده از اطلاعات موجود در تقطیع معنایی فریم‌های قبلی یک ویدئو موجب بهبود عملکرد روش‌های تقطیع معنایی ویدئو می‌شود. در مواجهه با دنباله فریم‌های یک ویدئو راه‌کارهای مختلفی جهت استفاده از دانش موجود در تقطیع معنایی فریم‌های قبلی وجود دارد که از جمله این راه‌کارها می‌توان به استفاده از جریان نوری بین فریم‌ها و همچنین استفاده از شبکه‌های عصبی عمیق بازگشتی اشاره کرد. در روش پیشنهادی این مقاله نشان داده شده است که ترکیب شبکه‌های عصبی عمیق بازگشتی و مدل‌های پیش پردازش فریم‌ها (نظیر محاسبه جریان نوری) از نظر کارایی (دقت و سرعت) منجر به دستیابی به نتایج مطلوبی می‌شود. در این روش، استفاده از مدل‌های از پیش آموزش دیده برای کارهای پیش پردازش همانند تخمین جریان نوری و یا تخمین اولیه تقطیع معنایی فریم‌ها منجر به افزایش قدرت مدل و بهبود سرعت و دقت تقطیع معنایی دنباله فریم‌های ویدئویی می‌شود.

در روش پیشنهادی انتخاب نوع شبکه کانولوشنال و همچنین ساختار تخمین جریان نوری، تعیین کننده کارایی مدل است و با انتخاب انواع مختلف مدل‌ها برای این دو منظور می‌توان یک مصالحه بین دقت و سرعت پردازش ایجاد کرد.

به منظور ادامه زنجیره تحقیق موارد مختلفی را می‌توان مدنظر قرار داد، از جمله:

۱- استفاده از مفهوم فریم کلیدی جهت افزایش سرعت و انتشار ویژگی‌های فریم‌های کلیدی در فریم‌های بعدی.

۲- ارائه یک مدل انتخاب ویژگی جهت استفاده در ورودی‌های شبکه GRU و روی خروجی شبکه کانولوشنال و همچنین روی خروجی لغزش یافته تقطیع معنایی فریم قبلی در راستای جریان نوری.

مراجع

[1] F.J.Chang, Y.Y.Lin, and K.-J. Hsu, “Multiple structured-instance learning for semantic segmentation with uncertain training data”, Proceedings of the IEEE Computer Vision and Pattern Recognition, pp. 360-367, 2014.

[2] X. Zhu, Y, Xiong, J, Dai, L, Yuan, and Y. Wei,“Deep feature flow for video recognition”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2349–2358, 2017.

[3] D. Lin Y. Li J. Shi, “Low-Latency Video Semantic Segmentation”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[4] P.Hu, F.Caba, O.Wang, Z.Lin, S.Sclaroff and F.Perazzi, “Temporally distributed networks for fast video semantic segmentation”, CVPR, pp. 8818–8827, 2020.

[5] M.Fan, Sh.Lai, J.Huang, X.Wei, Z.Chai, J.Luo and X.Wei, “Rethinking BiSeNet For Real-time Semantic Segmentation”, CVPR, 2021.

[6] H.Wang, W.Wang and J.Liu, “TEMPORAL MEMORY ATTENTION FOR VIDEO SEMANTIC SEGMENTATION”, CVPR, 2021.

[7] A.Tao, K.Sapra and B.Catanzaro, “Hierarchical Multi-Scale Attention for Semantic Segmentation”, CVPR, 2020.

[8] EfficientPS: R.Mohan and A.Valada, “Efficient Panoptic Segmentation”, International Journal of Computer Vision volume 129, p.1551–1579, 2021.

[9] B.Cheng, M.D.Collins, Y.Zhu, T.Liu, T.S.Huang and H.Adam, “Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation”, CVPR, 2020.

[10] M.Khalooei, M.Fakhredanesh, M.Sabokrou, “Dominant and rare events detection and localization in video using Generative Adversarial Network”,Journal of Soft Computing and Information Technology (JSCIT), Volume 8, Number 3, pp. 40-51, 2019.

[11] M.Fakhredanesh, S.Roostaei, “Action Change Detection in Video Based on HOG”, Journal of Electrical and Computer Engineering Innovations (JECEI), pp. 135-144, 2020.

[12] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy and R. Klette, “STFCN: spatio-temporal FCN for semantic video segmentation”, CoRR,2016.

[13] P. Fischer, A. Dosovitskiy, E. Ilg, P. Hausser, C. Hazırbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox,“Flownet: Learning optical flow with convolutional networks”, IEEE International Conference on Computer Vision (ICCV), 2015.

[14] E. L. Denton, S. Chintala, R. Fergus, et al., “Deep generative image models using a laplacian pyramid of adversarial networks”, in Proc. Neural Information Processing Systems(NIPS), pp 1486-1494, 2017.

[15] F.Galasso, M.Keuper, T.Brox and B. Schiele, "Spectral graph reduction for efficient image and streaming video segmentation", IEEE Conference on Computer Vision and Pattern Recognition, pp. 49-56, 2014.

[16] A.Khoreva, F.Galasso, M.Hein and B.Schiele, "Classifier based graph construction for video segmentation", Computer Vision and Pattern Recognition (CVPR) 2015 IEEE Conference, pp. 951-960, 2015.

[17] S. Hickson, S. Birchfield, I. Essa, and H. Christensen, "Efficient hierarchical graph-based segmentation of RGBD videos", IEEE Conference on Computer Vision and Pattern Recognition, pp. 344-351, 2014.

[18] S.Ardeshir, K.Malcolm and M.Shah, "Geo-semantic segmentation", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2792-2799, 2015.

[19] G.Bertasius, L.Torresani, S.X.Yu and J.Shi, "Convolutional Random Walk Networks for Semantic Image Segmentation" , arXiv:1605.07681, 2016.

[20] M.P.Kumar, H.Turki, D.Preston and D.Koller, "Parameter estimation and energy minimization for region-based semantic segmentation", IEEE transactions on pattern analysis and machine intelligence, vol. 37, pp. 1373-1386, 2015.

[21] M.Volpi and V.Ferrari, "Semantic segmentation of urban scenes by learning local class interactions", IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-9, 2015.

[22] A.Sharma, O.Tuzel and D.W.Jacobs, "Deep hierarchical parsing for semantic segmentation", IEEE Conference on Computer Vision and Pattern Recognition, pp. 530- 538, 2015.

[23] Z.Liu, X. Li, P. Luo, C.-C. Loy and X. Tang, "Semantic image segmentation via deep parsing network", IEEE International Conference on Computer Vision, pp. 1377- 1385, 2015.

[24] B. Liu, X. He, and S. Gould, "Multi-class semantic video segmentation with exemplar-based object reasoning", IEEE Winter Conference on Applications of Computer Vision, pp. 1014- 1021, 2015.

[25] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black, "Optical flow with semantic segmentation and localized layers", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[26] G. Csurka and F. Perronnin, "An efficient approach to semantic segmentation", International Journal of Computer Vision, vol. 95, pp. 198-212, 2011.

[27] C.-F. Tsai, K. McGarry, and J. Tait, "Image classification using hybrid neural networks", 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 431-432, 2003.

[28] T. Blaschke, C. Burnett, and A. Pekkarinen, "Image segmentation methods for object-based analysis and classification", Remote sensing image analysis: Including the spatial domain, ed: Springer, pp. 211-236, 2004.

[29] S.Hochreiter and J.Schmidhuber, “Long short-term memory”, Neural computation, pp. 1735–1780, 1997.

[30] K.Cho, B.Merrienboer, C.Gulc¸ F.Bougares, H.Schwenk and Y.Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation”, EMNLP, 2014.

[31] J.Long, E.Shelhamer, and T.Darrell, “Fully convolutional networks for semantic segmentation”, CVPR, pp. 3431– 3440, 2015.

[32] S.Zheng , “Conditional random fields as recurrent neural networks”, IEEE Int. Conf. Computer Vision, pp. 1529-1537, 2015.

[33] V.Badrinarayanan, A.Kendall and R.Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”, CoRR, 2015.

[34] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia “Pyramid scene parsing network”, CVPR, 2017.

[35] A.Kundu, V.Vineet and V.Koltun, “Feature space optimization for semantic video segmentation”, CVPR, 2016.

[36] B.Mahasseni, S.Todorovic, A.Fern, “Budget-Aware Deep Semantic Video Segmentation”, IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[37] X.Jin, X.Li, H.Xiao, X.Shen, Z.Lin, J.Yang, Y.Chen, J.Dong, L.Liu and Z.Jie, “Video scene parsing with predictive feature learning”, ICCV, 2017.

[38] S.Jain, X.Wang and J.Gonzalez, “Accel: A corrective fusion network for efficient semantic segmentation on video”, CVPR, 2019.

[39] E. Shelhamer, K. Rakelly, J. Hoffman, and T,“Darrell. Clockwork convnets for video semantic segmentation”, European Conference on Computer Vision (ECCV) Workshops, pp. 852-868 , 2016.

[40] J.Carreira, V.Patraucean, L.Mazare, A.Zisserman and S.Osindero, “Massively parallel video networks”, ECCV, 2018.

[41] Y.He, W.Chiu, M.Keuper and Mario Fritz, “Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling”, CVPR, 2017.

[42] G.Hinton, O.Vinyals and J.Dean, “Distilling the knowledge in a neural network”, arXiv:1503.02531, 2015.

[43] G.Huang, Z.Liu, L.V.Maaten and K.Weinberger, “Densely connected convolutional networks”, CVPR, 2017.

[44] S.Chandra, C.Couprie and I.Kokkinos, "Deep Spatio-Temporal Random Fields for Efficient Video Segmentation", IEEE Conference of Computer Vision and Pattern Recognition, pp. 8915–8924, 2018.

[45] A.Handa, V.Patraucean and R.Cipolla, "Spatio-temporal video autoencoder with differentiable memory", ICLR Workshop, 2016.

[46] N. Ballas, L. Yao, C. Pal, and A.Courville, "Delving deeper into convolutional networks for learning video representations", 2016.

[47] R. Gadde, V. Jampani, and P. V. Gehler,"Semantic video cnns through representation warping",IEEE International Conference on Computer Vision (ICCV), 2017.

[48]E.Ilg, N.Mayer, T.Saikia, M.Keuper, A.Dosovitskiy and T.Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks", CVPR, 2016.

[49] https://www.cityscapes-dataset.com, Accessed: Feb. 21, 2019.

[50] http://mi.eng.cam.ac.uk/research/projects/VideoRec/Camvid, Accessed: Ap. 3, 2019.

[51] Yu and F.Koltun, "Multi-scale context aggregation by dilated convolutions", ICLR, 2016.

[52] T.W.Hui, X.Tang and C.Ch.Loy, "LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[53] X.Li, A.You, Z.Zhu, H.Zhao, M.Yang, K.Yang, Sh.Tan andY.Tong, 'Semantic Flow for Fast and Accurate Scene Parsing", ECCV 2020, pp. 775-793, 2020.

[54] Y.Nirkin, L.Wolf and T.Hassner, "HyperSeg: Patch-wise Hypernetwork for Real-time

Semantic Segmentation", CVPR, 2021.

[55] D.Nilsson and C.Sminchisescu, "Semantic Video Segmentation by Gated Recurrent Flow Propagation", IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[56] Ch.Yu, J.Wang, Ch.Peng and Ch.Gao, "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation", ECCV 2018, pp. 334-349, 2018.

[57] M.D.Yang, J.Boubin, H.P.Tsai and H.Tseng, "Adaptive autonomous UAV scouting for rice lodging assessment using edge computing with deep learning EDANet", Computers and Electronics in Agriculture, 2020.

[58] Y.Zhu, K.Sapra, F.Reda; K.Shih, Sh.Newsam, A.Tao and Bryan Catanzaro, "Improving Semantic Segmentation via Video Propagation and Label Relaxation", IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[59] Y.Liu, Ch.Shen, Ch.Yu and J.Wang, "Efficient Semantic Video Segmentation with Per-Frame Inference", ECCV, pp.352-368, 2020.

[60] Y.Hong, H.Pan, W.Sun and Y.Jia, "Deep Dual-resolution Networks for Real-time and Accurate Semantic Segmentation of Road Scenes", CVPR, 2021.

[61] Ch.Yu, Ch.Gao, J.Wang, G.Yu, Ch.Shen and N.Sang, "BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation", International Journal of Computer Vision volume 129, p. 3051–3068, 2021.

[62] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs", ICLR, 2015.

پاورقی‌ها:

[1] pixel

[2] Convolutional Neural Network

[3] Optical Flow

[4] Gated Recurrent Unit

[5] Support Vector Machine

[6] Markov Random Field

[7] Conditional Random Field

[8] LSTM (Long-Short Term Memory)

[9] Encoder-Decoder

[10] Fully Convolutional Network

[11] Tensor

[12] DFF(Deep Feature Flow)

[13] منظور از ضرب المانی ضرب تک تک عناصر تنسورها در یکدیگر می‌باشد. در اینجا منظور از ضرب المانی hadamard می باشد.

[14] mean intersection-over-union

[15] Tensorflow

[16] SGD (Stochastic Gradient Descent)

[17] Momentum

[18] weight decay

[19] Frame per Second