تقسیم بندی معنایی تصویر با CRFهای کاملاً متصل و شبکه های پیچیده عمیق

چکیده

شبکه های عصبی مصنوعی عمیق (DCNNs) اخیرا وضعیت عملکرد هنری در وظایف بصری سطح بالا مانند طبقه بندی تصویر و تشخیص شی را نشان دادند .این کار روش های متفاوتی از DCNN ها و مدل های گرافیکی احتمالی برای رسیدگی به وظیفه طبقه بندی سطح پیکسل (همچنین "تقسیم بندی تصویر معنایی" نامیده می شود) را به ارمغان می آورد .ما نشان می دهیم که پاسخ ها در لایه نهایی DCNNs برای تقسیم بندی دقیق شیئ به اندازه کافی متمرکز نیستند. علت آن ویژگی های بسیار تغییرناپذیری ست که DCNNها را برای وظایف سطح بالا مناسب می سازد.ما با ترکیب پاسخ ها در لایه DCNN نهایی با یک فیلد تصادفی محرمانه کاملا متصل(CRF) بر این ویژگی محلی سازی نامرغوب شبکه های عمیق غلبه می کنیم. از لحاظ کیفیت، سیستم "DeepLab" ما قادر به محاسبه تقسیم مرزها به سطح دقت فراتر از روش های قبلی است . از لحاظ کیفیت ،روش ما جدیدترین حالت هنر را در PASCAL VOC-2012 وظیفه تقسیم بندی تصویر معنایی معین می کند، رسیدن به ۷۱,۶٪ دقت IOU در مجموعه آزمون. ما نشان می دهیم چگونه این نتایج را می توان به طور موثری به دست آورد: اهداف دقیق شبکه وکاربرد جدید از الگوریتم "سوراخ" از جامعه موجک محاسبه تراکم پاسخهای شبکه عصبی با ۸ فریم در ثانیه بر روی GPU مدرن را اجازه می دهد.

۱. معرفی

شبکه های عصبی مصنوعی عمیق (DCNNs) روش انتخابی برای شناخت سند از LeCun و همکاران بوده است (۱۹۹۸)، اما اخیرا تبدیل به جریان اصلی پژوهش بصری سطح بالا شده است. در طول دو سال گذشته، DCNNها عملکرد سیستم های بینایی کامپیوتر را به افزایش ارتفاع در یک مجموعه گسترده از مشکلات سطح بالا تحت تاثیر قرار داده اند، از جمله طبقه بندی تصویر(Krizhevsky و همکاران، ۲۰۱۳؛ Sermanet و همکاران، ۲۰۱۳؛ Simonyan و Zisserman، 2014؛ Szegedy و همکاران، ۲۰۱۴؛.Papandreou et al ، 2014)، تشخیص شی (Girshick و همکاران، ۲۰۱۴)، رده بندی دقیق دانه (Zhangو همکاران، ۲۰۱۴)، در میان دیگران. موضوع

مشترک در این آثار این است که DCNN ها در وضعیت پایان تا پایان آموزش دیده اند تا تحویل دهند به طرز قابل ملاحظه ای نتایج بهتر را نسبت به سیستم های تکیه بر بازنمودهای دقت مهندسی ارائه می دهد، مانند ویژگی های SIFT یا HOG. این موفقیت می تواند تا حدودی حمل بر تغییرناپذیریDCNNها به تحولات تصویر محلی باشد، که توانایی آنها در یادگیری انتزاع سلسله مراتبی از اطلاعات را پشتیبانی می کند(Zeiler & Fergus، 2014). در حالی که این تغییرات به وضوح برای وظایف بینایی سطح بالا مطلوب است،این می تواند مانع انجام وظایف در سطوح پایین مانند برآورد آستانه (Chen & Yuille، 2014؛ Tompson et al.،۲۰۱۴) و تقسیم بندی معنایی - جایی که ما می خواهیم دقیق تر محلی سازی کنیم، سریعتر از انتزاع جزئیات مکانی باشد.

در استفاده از DCNN ها برای انجام وظایف برچسب گذاری تصویر دو مانع فنی وجود دارد:نمونه برداری پایین سیگنال و "غیر حساسیت" فضایی (غیر انسانی). اولین مشکل مربوط به کاهش وضوح سیگنال داده شده توسط ترکیبی مکرر از حداکثر ادغام و نمونه برداری پایین ("قدم زدن") انجام شده در هر لایه از DCNN های استاندارد است(Krizhevsky و همکاران، ۲۰۱۳؛ Simonyan & Zisserman، 2014؛Szegedy و همکاران، ۲۰۱۴). در عوض، همانطور که در Papandreou و همکاران (۲۰۱۴)، ما الگوریتم "اتلاف" (با سوراخ) را به کار بردیم که در اصل برای محاسبه کارآمد تلفات زیاد تبدیل موجک گسسته (مالات، ۱۹۹۹) رشدیافته است. این اجازه می دهد محاسبه تراکم کارآمد از پاسخ های DCNN در یک طرح به طور قابل توجهی ساده تر از راه حل های قبلی برای این مشکل (Giusti و همکاران، ۲۰۱۳؛ Sermanet و همکاران، ۲۰۱۳)باشد.

مشکل دوم مربوط به این واقعیت است که به دست آوردن تصمیمات شیء محوری از یک طبقه بندی انحراف به تحولات فضایی نیاز دارد، ذاتا دقت فضایی مدل DCNN محدود کننده است. ما توانایی مدل خود را برای ضبط جزئیات دقیق با استفاده از به کار بردن فیلد تصادفی شرطی به طور کامل متصل بالا ببریم (CRF). زمینه های تصادفی شرطی در بخش بندی معنایی به طور گسترده ای برای ترکیب امتیازات کلاس که توسط طبقه بندی های چند راهه محاسبه شده با اطلاعات کم سطح ضبط شده توسط تعاملات محلی پیکسلها و لبه ها (Rother et al.، 2004؛ Shotton et al.، 2009) یا سوپرپیکسل ها (Lucchi et al.، 2011) مورد استفاده قرار گرفته است. اگر چه کارهای مهارت پیشرفته برای مدل سازی وابستگی سلسله مراتبی (He et al.، 2004؛ Ladicky et al.، 2009؛ Lempitsky et al.، 2011)و / یا وابستگی های بالا از بخش های مختلف (Delong و همکاران،

۲۰۱۲؛ Gonfaus و همکاران، ۲۰۱۰؛ Kohli و همکاران،۲۰۰۹؛ چن و همکاران، ۲۰۱۳؛ وانگ و همکاران، ۲۰۱۵) پیشنهاد شده است، ما CRF جفت جفت به طور کامل متصل شده پیشنهاد شده توسط Kr¨ahenb¨uhl & Koltun (۲۰۱۱) برای محاسبه کارآمد و توانایی گرفتن جزئیات لبه های خوب را استفاده می کنیم. در حالی که همچنین برای وابستگی های طولانی مدت فراهم شده است. این مدل که در Kr¨ahenb¨uhl & Koltun نشان داده شده بود(۲۰۱۱) تا حد زیادی عملکرد یک طبقه بندی سطح پیکسل مبتنی بر افزایش(ترقی) در کار ما را بهبود داد. ما نشان می دهیم که این نتایج وقتی که با یک طبقه بندی سطح پیکسل مبتنی بر DCNN جفت می شود منجر به نتایج پیشرفته ای می شود.

سه مزیت اصلی سیستم "DeepLab" ما شامل (i) سرعت: با توجه به مزیت الگوریتم "atrous"، DCNN متراکم ما با سرعت ۸ فریم در ثانیه کار می کند، در حالی که استنتاج فیلد میانگین برای CRF کاملا متصل نیاز به ۰,۵ ثانیه دارد، (ii) دقت: ما نتایج پیشرفته ای در چالش تقسیم بندی معنایی PASCAL، که بهترین-دومین رویکرد از مستجابی و همکاران ۷,۲٪ بدست آمد. (۲۰۱۴) به دست آوردیم و (iii) سادگی: سیستم ما از یک آبشار از دو ماژول نسبتا معتبر ثابت ،DCNNها و CRFها تشکیل شده است.

۲ کار مرتبط

سیستم ما به طور مستقیم بر روی نمایش پیکسل کار می کند، به طور مشابه به Long و همکاران. (۲۰۱۴). این با رویکردهای دو مرحله ای که در بخش بندی معناشناختی با DCNNs در حال حاضر رایج است در مقابل است: این تکنیک ها معمولا از یک آبشار تقسیم بندی تصویر پایین و بالا و طبقه بندی منطقه مبتنی بر DCNN استفاده می کنند، که باعث می شود سیستم متعهد به اشتباهات بالقوه از سیستم تقسیم بندی جلو عقب باشد. برای مثال، پیشنهادات محدوده جعبه و مناطق آسیب دیده (ماسک زده) که توسط (Arbel'aez et al.،۲۰۱۴؛ Uijlings و همکاران، ۲۰۱۳) در Girshick و همکاران استفاده می شود تحویل داده می شوند. (۲۰۱۴) و (Hariharan و همکاران، b۲۰۱۴) به عنوان ورودی به DCNN برای معرفی اطلاعات شکل در فرآیند طبقه بندی استفاده می شود. به طور مشابه، نویسندگان مستجابی و همکاران (۲۰۱۴) به نمایندگی سوپرپیکسل تکیه می کنند. یک پیشگام غیر DCNN مشهور به این کارها دومین روش ادغام منظم از (.Carreira et al، 2012) است که همچنین برچسب ها ی پیشنهادات مناطق ارائه شده توسط (Carreira & Sminchisescu، 2012) را اختصاص

می دهد. درک خطرات مرتکب به یک تقسیم بندی واحد، نویسندگان Cogswell و همکاران. (۲۰۱۴) ساخت (Yadollahpourو همکاران، ۲۰۱۳) برای کشف یک مجموعه متنوع از پیشنهادات تقسیم بندی بر اساس CRF، نیز توسط(Carreira & Sminchisescu، 2012) محاسبه شده. این پیشنهادات تقسیم بندی به ترتیب با توجه به یک DCNN به ویژه برای این کار مجدد آموزش دیده است. حتی اگر این روش به صراحت تلاش کند که این طبیعت تند مزاج الگوریتم تقسیم بندی جلو عقب را کنترل کند، هنوز هیچ بهره برداری صریحی از امتیازات DCNN در الگوریتم تقسیم بندی مبتنی برCRF وجود ندارد : DCNN تنها کاربرد post-hoc دارد در حالی که آن را مستقیما می توانید از نتایج آن در طول تقسیم بندی استفاده کنید.

حرکت به سوی آثار به رویکرد ما نزدیک تر است، چندین محقق دیگر در نظر گرفته اند که از ویژگی های DCNN محاسبه شده پیچیده برای نشانه گذاری تصویر متراکم استفاده کنند. در میان اولینFarabet و همکاران (۲۰۱۳) کسی که DCNN ها را در وضوح تصویری مختلف(چندگانه) اعمال می کند و سپس درخت تقسیم بندی برای صاف کردن نتایج پیش بینی را به کار می برند ؛ اخیرا، هری هاراران و همکاران. (a۲۰۱۴) الحاق نقشه های میان محوری میان محتوی محاسبه شده را در DCNN ها را برای طبقه بندی پیکسل پیشنهاد می کند، و دای و همکاران (۲۰۱۴) پیشنهاد می کند که نقشه های میان-میانجی را براساس پیشنهادات منطقه ای بسازد. با اینکه این کارها هنوز الگوریتم های تقسیم بندی را که از نتایج طبقه بندی DCNN جدا می شوند، استفاده می کنند ، ما که تقسیم بندی تنها در مرحله بعد استفاده می شود معتقدیم مفید است ، ازتعهد به تصمیمات زودرس اجتناب کنیم.

اخیرا، تکنیک های بدون تقسیم بندی از(.Long et al، 2014؛ Eigen & Fergus، 2014) به طور مستقیم DCNN ها را به کل تصویر در یک شکل پنجره کشویی اعمال می کند، آخرین لایه های DCNN کامل متصل شده توسط لایه های کانولوشن جایگزین می شود. برای مقابله با محلی سازی فضایی موضوعاتی که در ابتدای مقدمه ذکر شده است، Long et al.(۲۰۱۴) نمونه و پیوسته است این نمرات از نقشه های ویژگی های میانجی، در حالی که Eigen & Fergus (۲۰۱۴) نتیجه پیش بینی را اصلاح می کند از بزرگ به خوب با انتشار نتایج درشت به DCNN دیگر.

تفاوت اصلی بین مدل ما و دیگر مدل های پیشرفته کشور ترکیبی ازCRF های سطح پیکسل و اصطلاحات Unary مبتنی بر DCNN است. تمرکز بر نزدیکترین آثار در این جهت،Cogswell et al (۲۰۱۴) استفاده از CRF ها به عنوان مکانیسم پیشنهادی برای سیستم بازاریابی مبتنی بر DCNN است،در حالی که فاراب و همکاران (۲۰۱۳) سوپرپیکسل ها را به عنوان گره ها برای یک CRF زوج محلی پردازش می کنند و از برش های نمودار برای استنتاج گسسته استفاده می کنند ؛ به همین ترتیب نتایج آنها می تواند توسط اشتباهات در محاسبات superpixel محدود شده باشد، درحالی که وابستگی های سوپرپیکسل دوربرد نادیده گرفته می شود. رویکرد ما به جای رفتار کردن هر پیکسل به عنوان گره CRF ، از وابستگی های طولانی مدت و استفاده می کنه ازنتایج CRF برای بهینه سازی مستقیم عملکرد هزینه هدایت DCNN بهره برداری می کند. ما یادآوری می کنیم که زمینه متوسط برای جداسازی تصویر سنتی /وظایف شناسایی لبه از همه جا مطالعه شده، برای مثال (Geiger & Girosi، 1991؛ Geiger & Yuille، 1991؛ Kokkinos et al.،۲۰۰۸)، اما اخیرا Kr¨ahenb¨uhl & Koltun (2011) نشان داد که استنتاج می تواند برای CRF به طور کامل متصل شده بسیار موثر باشد و به ویژه در زمینه تقسیم بندی معنایی موثر است.

پس از اینکه اولین نسخه از نوشته ما به صورت عمومی منتشر شد، متوجه شدیم که دو گروه دیگر به طور مستقل و همزمان یک جهت بسیار مشابه دنبال می کنند، DCNNs ترکیبی و CRF های متصل به انسداد (.Bell et al، 2014؛ .Zheng et al، 2015). چندین تفاوت در جنبه های فنی مدل های مربوطه وجود دارد. بل و همکاران (۲۰۱۴) بر مشکل طبقه بندی مواد تمرکز دارند ، در حالی که Zheng و همکاران (۲۰۱۵) مراحل استنتاج میانگین میدانی CRF را به تبدیل این سیستم کامل به یک شبکه رو به جلو تربیت شدنی end-to-end باز می کنند.

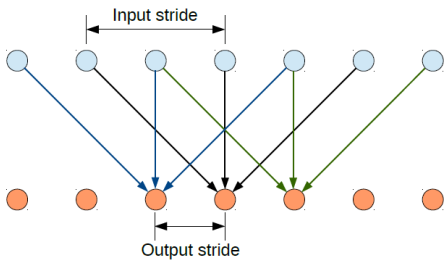
ما سیستم DeepLab پیشنهادیمان را با روشهای بسیار بهبود یافته و نتایج را در آخرین کارمان به روز کرده ایم (چن و همکاران، ۲۰۱۶). ما خواننده علاقه مند به مقاله را برای جزئیات ارجاع می دهیم.

۳ شبکه های عصبی مصنوعی برای برچسب زدن تصویر مصنوعی

در اینجا ما توضیح می دهیم که چگونه دوباره پیشنهاد کردیم و تصاویر دردسترس عمومی را شبکه تقسیم بندی ۱۶ لایه از(VGG-16) (Simonyan & Zisserman 2014) به یک استخراج کننده مصنوعی مفید و مؤثر برای سیستم بخش بندی تصویر معنادار مصنوعی ما به خوبی میزان کردیم.

۳,۱ کارآیی استخراج ویژگی اسلاید کردن مصنوعی پنجره با الگوریتم حفره

ارزیابی امتیاز فضایی مصنوعی در موفقیت استخراج کننده ویژگی CNN مصنوعی ما مفید است. مانند گام اول برای پیاده سازی این، ما لایه های کاملا به هم متصل شده از VGG-16 را به موارد پیچیده دیگر تبدیل کردیم این شبکه را در یک سبک پیچیده بر روی تصویر در وضوح اصلی خود واجرا کردیم . با این حال این به اندازه کافی نیست، زیرا این بازده به صورت خیلی پراکنده نمرات تشخیص را (با ۳۲ پیکسل در ثانیه) محاسبه می کند. برای محاسبه نمرات به میزان بیشتر مصنوعی در گام هدفمند ما از ۸ پیکسل، ما یک تغییر از روش را توسعه می دهیم که قبلا توسط Giusti و همکاران مورد استفاده قرار گرفته است. (۲۰۱۳)؛ Sermanet و همکاران (۲۰۱۳). بعد از آن نمونه زیر را پرش می کنیم این دو لایه آخر max-pooling در شبکه ازSimonyan & Zisserman (۲۰۱۴) و اصلاح می کنیم این فیلترهای پیچیده را در لایه هایی که آنها را با اضافه کردن صفر به افزایش طول آنها دنبال می کنیم (۲ سانتیمتر)



شکل ۱: الگوریتم سوراخ در ۱-D، زمانی که اندازه کرنل = ۳، ورودی stride = 2 وگام خروجی = ۱.

سه لایه کانولوشن آخری و ×4 در اولین لایه کامل متصل شده). ما می توانیم این را بصورت کارآمد با نگه داشتن فیلترهای دست نخورده و در عوض، نمونه پراکنده این نقشه های ویژگی را که بر روی آنها ، با استفاده از گام ورودی ۲ یا ۴ پیکسل به کار می روند به ترتیب انجام دهیم. این رویکرد، نشان داده شده در شکل ۱، به عنوان الگوریتم سوراخ شناخته شده است (الگوریتم atrous) و قبل از آن برای محاسبات کارآمد از تبدیل موجک ناپیوسته رشد یافته بود (Mallat، 1999). ما این در چارچوب کافه (جیا و همکاران، ۲۰۱۴) با اضافه کردن به

عملکرد im2col (این تبدیل می کند نقشه های ویژگی چند کاناله را به تکه های برداری) که گزینه ای برای نمونه پراکنده این نقشه ویژگی اساسی است پیاده سازی کرده ایم. این رویکرد به طور کلی قابل اجرا است و به ما امکان می دهد تا به طور کارآمد نقشه های ویژگی CNN متراکم را در هر نرخ زیرنمونه هدف بدون معرفی هیچ تقریبی محاسبه کنیم.

ما وزن مدل های شبکه VGG-16 پیشگیری شده تصویری را به منظور تطبیق آن با وظیفه طبقه بندی تصویر که به روش ساده، این روش را Long و همکاران دنبال کردند، استفاده می کنیم. (۲۰۱۴). ما جایگزین کردیم این دسته کننده Imagenet ۱۰۰۰ نوع را در آخرین لایه VGG-16 با یک ۲۱ بیتی.تابع اتلاف ما مجموع عناصر انتروپی متقابل برای هر موقعیت مکانی در نقشه خروجی CNN است(در مقایسه با تصویر اصلی ۸ برابر شده است). همه مواضع و برچسب ها در این تابع اتلاف کلی به یک اندازه دارای وزن هستند. اهداف ما برچسبهای حقیقی زمین هستند (زیرنمونه با ۸ عدد). ما تابع هدف را با توجه به وزن در تمام لایه های شبکه با روش SGD استاندارد از Krizhevsky و همکاران (۲۰۱۳) بهینه سازی کردیم.

در حین آزمایش، ما نیاز به نقشه نمره کلاس در رزولوشن تصویر اصلی داشتیم. همانطور که در شکل ۲ نشان داده شده است و در بخش ۴,۱ بیشتر توضیح داده شده است، نقشه نمرات کلاس (مربوط به احتمال ورود)کاملا صاف است، که به ما اجازه می دهد تا با استفاده از یک درون یابی دوسویه ساده رزولوشن آن را با یک عامل ۸ در هزینه محاسباتی ناچیز افزایش دهیم. توجه داشته باشید که روش لانگ و همکاران (۲۰۱۴) از الگوریتم سوراخ و تولید نمرات بسیار سنگین (زیر نمونه ای با یک عامل ۳۲) درخروجی CNN استفاده نمی کند. این کار آنها را مجبور به استفاده از لایه هایupsampling یاد شده می کند، به طور قابل توجهی افزایش پیچیدگی و زمان آموزش سیستم خود: تنظیم شبکه ما در PASCAL VOC 2012 حدود ۱۰ ساعت است، در حالی که آنها یک دوره آموزشی چند روزه (هر دو زمان در GPU مدرن) را گزارش می کنند.

۳,۲ کنترل اندازه فیلد پذیرنده و محاسبه تراکم تسریع کننده با شبکه های پیچیده

یکی دیگر از عناصر کلیدی در بازنگری شبکه ما برای محاسبه مقدار تراکم به صراحت کنترل اندازه فیلد پذیرنده شبکه است. جدیدترین روش های تشخیص تصویر مبتنی بر DCNN بر روی شبکه ها در وظیفه تقسیم بندی در مقیاسی بزرگ Imagenet آموزش دیده اند. این شبکه ها به طور معمول اندازه فیلد پذیرنده بزرگ دارد: در مورد

شبکه VGG-16 ما در نظر گرفتیم ، فیلد پذیرنده آن است۲۲۴ × ۲۲۴ (با صفر خالی) و ۴۰۴ × ۴۰۴ پیکسل اگر این شبکه به طور پیچیده اعمال شود. بعد از تبدیل این شبکه به یک کانولوشن کامل دیگر، اولین لایه کاملا متصل شده ۴،۰۹۶ فیلتر از اندازه فضایی 7×7 بزرگ و تنگنا محاسباتی در محاسبات نقشه نمره متراکم ما می شود.

ما این مسئله عملی را با زیرنمونه فضایی (با تخریب ساده) لایه FC تا اندازه فضایی ۴ × ۴ (یا ۳ × ۳) در ابتدا مورد توجه قرار داده ایم. این امر موجب کاهش فیلد پذیرنده پایین شبکه شده است در 128×128 (با کاراکتر صفر) یا ۳۰۸ × ۳۰۸ (در حالت پیچیده) و زمان محاسبه را کاهش داده است برای لایه FC اول با ۳-۲ بار. با استفاده از اجرای مبتنی بر caffe ما و تیتان GPU،در نتیجه شبکه VGG-derived بسیار کارآمد است: با دادن یک تصویر ورودی ۳۰۶ × ۳۰۶، آن امتیازات ویژگی متراکم ۳۹ × ۳۹ را تولید می کند در بالای شبکه با سرعت حدود ۸ فریم در ثانیه در طی آزمایش. این سرعت در طول تمرین ۳ فریم در ثانیه است. ما همچنین با موفقیت آزمایش کردیم با کاهش تعداد کانال ها در لایه های کاملا متصل از کمتر ۴،۰۹۶ تا ۱،۰۲۴، کاهش زمان محاسبه و رد حافظه بدون به خطر انداختن عملکرد به مراتب بیشتر است، همانطور که در ادامه دربخش ۵ آمده است. استفاده از شبکه های کوچکتر مانند Krizhevsky و همکاران. (۲۰۱۳) می تواند محاسبه ویژگی تراکم زمان آزمون میزان ویدیو را حتی در GPUهای سبک وزن اجازه دهد.

۴ بازیابی مرزی جزئی: فیلدهای تصادفی شرطی کاملا متصل و پیش بینی چند مقیاسی

۴,۱ شبکه های پیچیده عمیق و چالش محلی سازی

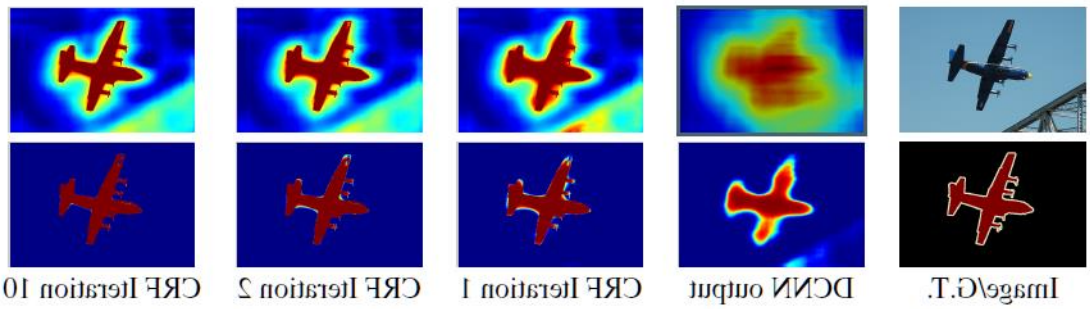
همانطور که در شکل ۲ نشان داده شده است، نقشه های نمرات DCNN می تواند به طور قابل اعتماد موقعیت دشوار و مجسم از اشیاء در یک تصویر را پیش بینی کند، اما برای نقطه گذاری پین نمای کلی دقیق خود کمتر مناسب است. یک همکاری طبیعی بین دقت طبقه بندی و دقت محلی سازی با شبکه های پیچیده وجود دارد: مدل های عمیق تر با لایه های حداکثر ادغام چندگانه بیشترین موفقیت را در وظایف طبقه بندی به اثبات رسانیدند. با این حال، فیلدهای پذیرنده بزرگ و تغییرناپذیر افزایش یافته آنها ، این مشکل موقعیت بدست آمده را از نمرات در سطح های خروجی بالا چالش انگیزتر خود را موجب می شود.

کار اخیر دو جهت را برای حل این چالش محلی سازی دنبال کرده است. اولین رویکرد برای به دست آوردن اطلاعات از لایه های مختلف در شبکه پیچیده به منظور تخمین بهتر مرزهای شی این است (.Long et al،

2014؛ Eigen & Fergus، 2014). رویکرد دوم این است که یک نماینده سوپر پیکسل، اساسا محول کردن وظیفه محلی سازی را به روش تقسیم بندی سطح پایین به کار می برد. این مسیر با روش بسیار موفق اخیر Mostabi و همکارانش دنبال می شود. (۲۰۱۴).

در بخش ۴,۲، ما یک مسیر جایگزین جدید را بر اساس پیوند ظرفیت شناختی از DCNNs و دقت محلی سازی دقیق CRFs به طور کامل متصل شده و نشان می دهد که به طور قابل ملاحظه ای موفق در رسیدگی به چالش محلی سازی است، تولید نتایج تقسیم بندی معنایی دقیق و بازخوانی مرزهای شی در سطح جزئی که بسیار فراتر از روش های در دسترس است.

۴,۲ فیلدهای تصادفی شرطی کاملا متصل با محلی سازی دقیق



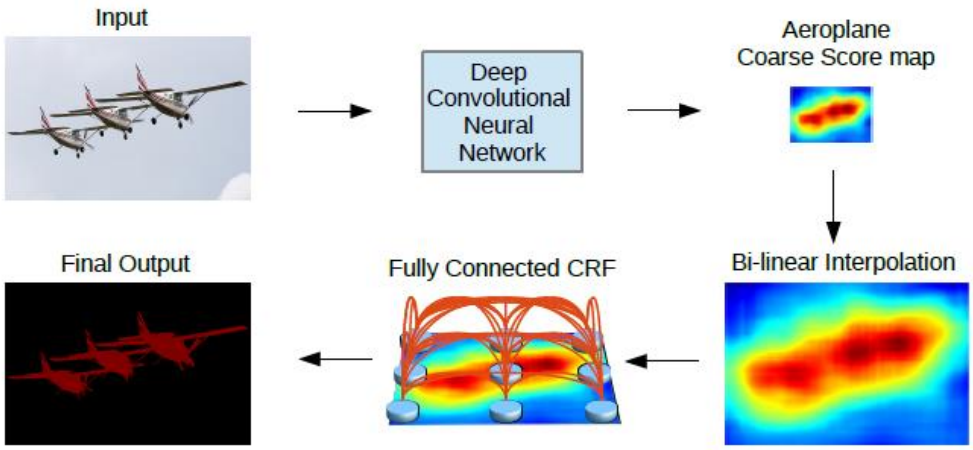
شکل ۲: نقشه نشان (ورودی قبل از عملکرد softmax) و نقشه باور (خروجی تابع softmax) برای هواپیما.ما پس از هر تکرار فیلد میانگین نمرات (ردیف ۱) و اعتبار (ردیف دوم) را نشان می دهیم.

خروجی آخرین لایه DCNN به عنوان ورودی استنتاج فیلد میانگین استفاده می شود. بهترین نمایش در رنگ.

به طور سنتی، زمینه های تصادفی شرطی (CRFs) برای نقشه های تقسیم بندی لرزه ای صاف استفاده شده است (Rother et al.، 2004؛ Kohli et al.، 2009). به طور معمول این مدل ها شامل شرایط انرژی هستند که گره های همسایه زوج، با توجه به تخصیص همان برچسب به پیکسل های فضایی پروکسیمال. به طور کافی، عملکرد اصلی این CRF های کوتاه برد برای پاکسازی پیش بینی های جعلی از طبقه بندی های ضعیف ساخته شده در بالای ویژگی های دستی مهندسی محلی است.

در مقایسه با این طبقه بندی های ضعیف، معماری مدرن DCNN، مانند آنچه ما استفاده کردیم در این کار تولید نقشه های امتیاز و پیش بینی های برچسب معنایی که به طور کیفی متفاوت هستند. همانطور که نشان داده شده

است در شکل ۲ نقشه های نمره به طور معمول کاملا صاف و نتایج طبقه بندی همگن ساخته شده هستند. در این رژیم، استفاده از CRF های کمینه می تواند زیان آور باشد، زیرا هدف ما باید بهبود ساختار محلی دقیق و نه بیشتر صافی آن با استفاده از پتانسیل حساس به کنتراست باشد.



شکل ۳: تصویر مدل. نقشه نمایی درشت از شبکه عصبی پیچیده عمیق (با

لایه های کاملا پیچیده) با یک درون یابی دو خطی نمونه برداری شده است. CRF به طور کامل متصل شده اعمال می شود

برای اصلاح نتایج تقسیم بندی. بهترین رنگ در حال مشاهده.

(Rother و همکاران، ۲۰۰۴) در ارتباط با CRF های محدوده محلی می توان به طور بالقوه محلی سازی را بهبود بخشید، اما هنوز هم ساختارهای نازک از دست می رود و به طور معمول نیاز به حل یک مشکل بهینه سازی گسسته گران قیمت دارد.

$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j)$$

برای غلبه بر این محدودیت های CRF های کوتاه مدت، ما به طور کامل متصل به سیستم مان هستیمCRF مدل Kr¨ahenb¨uhl & Koltun (۲۰۱۱). این مدل کارکرد انرژی را به کار می گیرد

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w_m \cdot k^m(\boldsymbol{f}_i, \boldsymbol{f}_j), \text{ where } \mu(x_i, x_j) = 1 \text{ if } x_i \neq x_j, \quad (1)$$

که x تخصیص برچسب برای پیکسل ها است. ما از پتانسیل منحصر به فرد استفاده می کنیم $i(x_i) = - \log$ که $\theta P(x_i)$

$P(x_i)$ احتمال تخصیص برچسب به پیکسل i است که توسط DCNN محاسبه می شود. پتانسیل دوگانه هست:

و صفر، در غیر این صورت (به عنوان مثال، مدل Potts). برای هر جفت پیکسل های i و j در تصویر یک اصطلاح جفت جفت وجود دارد مهم نیست که چطور دور از یکدیگر قرار دارند، یعنی نمودار عامل مدل به طور کامل متصل است. هر K^m هست هسته گاوسی وابسته به ویژگی های (به عنوان f مشخص شده است) استخراج شده برای پیکسل های i و j است و توسط پارامتر W_m وزن می گیرد، ما شرایط دو جانبه و رنگی را می پذیریم ، به طور خاص ، هسته ها به صورت زیر هستند:

$$w_1 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}\right) + w_2 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}\right) \quad (2)$$

که هسته اول به هر دو موقعیت پیکسل (تفکیک شده به عنوان p) و تداخل رنگ پیکسل بستگی دارد (تفکیک شده به عنوان I)، و هسته دوم تنها به موقعیت های پیکسل بستگی دارد. پارامترهای بالا σ_α و σ_β وσ_γ"مقیاس" هسته های گاوسی را کنترل می کند.

مهم این است که این مدل میتواند به نتیجه استدلال احتمالی تقریبا کارآمد برسد (Kr¨ahenb¨uhl & Koltun, 2011).این پیام در جریان به روز رسانی تقریبی فیلد متوسط به طور کامل تجزیه $b(x) = \Pi_i b_i(x_i)$ را می توان به صورت پیچیده با یک هسته گاوس در فضای مشخصه بیان کند. الگوریتم های فیلترینگ با کیفیت بالا (آدامز و همکاران، ۲۰۱۰) این محاسبات را به طور قابل توجهی سرعت بخشید نتیجه یک الگوریتم که در عمل بسیار سریع است، کمتر از ۰٫۵ ثانیه به طور متوسط برای تصاویر Pascal VOCبا استفاده از اجرای عمومی در دسترس از(Kr¨ahenb¨uhl & Koltun، 2011).

۴٫۳ پیش بینی چند بعدی

به دنبال نتایج امیدوار کننده اخیر (Hariharan et al. 2014a؛ Long et al.، 2014) ما همچنین روش پیش بینی چند مرحله ای را برای افزایش دقت محلی سازی مرزی کشف کردیم. به طور مشخص، ما به تصویر ورودی و خروجی هر یک از اولین چهار لایه جمع کننده ماکزیمم دو لایه MLP (لایه اول: ۱۲۸ 3x3 فیلترهای کانولوشن، لایه دوم: ۱۲۸ 1X1 فیلترهای کانولوشنی) که نقشه ویژگی آن به نقشه ویژگی آخرین لایه اصلی شبکه متصل است را پیوست می کنیم. نقشه ویژگی مجموعه که به لایه softmax وارد می شود با کانال ۵ * ۱۲۸ = ۶۴۰

افزایش می یابد. ما فقط تنظیم می کنیم وزن جدید اضافه شده را،پارامترهای شبکه دیگر را به مقادیر آموخته شده توسط این روش از بخش ۳ حفظ کنید. همانطور که در بخش تجربی بحث شده است، معرفی این اتصالات مستقیم اضافی از لایه های با وضوح خوب عملکرد محلی سازی را بهبود می بخشد، با این حال این اثر به اندازه ای چشمگیر نیست که یک CRF کاملا متصل به دست آمده است.

Method	mean IOU (%)
DeepLab	59.80
DeepLab-CRF	63.74
DeepLab-MSc	61.30
DeepLab-MSc-CRF	65.21
DeepLab-7x7	64.38
DeepLab-CRF-7x7	67.64
DeepLab-LargeFOV	62.25
DeepLab-CRF-LargeFOV	67.64
DeepLab-MSc-LargeFOV	64.21
DeepLab-MSc-CRF-LargeFOV	68.70

(a)

Method	mean IOU (%)
MSRA-CFM	61.8
FCN-8s	62.2
TTI-Zoomout-16	64.4
DeepLab-CRF	66.4
DeepLab-MSc-CRF	67.1
DeepLab-CRF-7x7	70.3
DeepLab-CRF-LargeFOV	70.3
DeepLab-MSc-CRF-LargeFOV	71.6

(b)

جدول ۱: (a) عملکرد مدل های پیشنهاد شده ما در مجموعه PASCAL VOC 2012 'val' set (با آموزش در مجموعه ’train‘ "تقویت شده"). بهترین عملکرد با استفاده از هر دو ویژگی های مقیاس چندگانه و میدان دید بزرگ به دست می آید. (b) عملکرد مدل های پیشنهادی ما (با آموزش در مجموعه 'trainval' تقویت شده) در مقایسه با سایر روش های پیشرفته در مجموعه آزمون PASCAL VOC 2012.

۵ ارزیابی تجربی

مجموعه داده : ما مدل DeepLab ما را براساس معیار تقسیم بندی PASCAL VOC 2012 آزمایش کردیم.(Everinghamو همکاران، ۲۰۱۴)، متشکل از ۲۰ کلاس شیء پیش زمینه و یک کلاس پس زمینه. این مجموعه داده اصلی شامل ۱۴۶۴/ ۱، ۱۴۴۹/ ۱ و۱۴۵۶/ ۱؛ به ترتیب تصاویر برای آموزش، اعتبارسنجی و تست.مجموعه داده تکمیل شده است با حاشیه های اضافی که ارائه شده توسط Hariharan et al. (۲۰۱۱)نتیجه در تصاویر آموزشی۱۰/ ۵۸۲. این عملکرد اندازه گیری شده در دوره هایی از تقاطع پیکسل اتحادیه (IOU) در میان تقاطع ۲۱ کلاس.

آموزش : ما ساده ترین شکل آموزش قطعی را اتخاذ کردیم، تجزیه مراحل آموزش های DCNN و CRF ، فرض بر این است که شرایط غیر معمول ارائه شده توسط DCNN در دوره آموزشی CRF ثابت می شود. برای آموزش DCNN ما از شبکه VGG-16 استفاده می کنیم که در ImageNet پیش آموزش دیده است. ما شبکه VGG-16 رادرمرتبه ی کارکرد طبقه بندی پیکسل VOC 21 با نزول شیب تصادفی در عملکرد تلفات آنتروپی متقاطع تنظیم می کنیم ، همانطور که در بخش ۳,۱ توضیح داده شده است. ما از یک دسته کوچک از ۲۰ تصاویر و نرخ یادگیری اولیه 0.001 (0. 01 برای لایه طبقه بندی نهایی) استفاده می کنیم ، ضرب کردن نرخ یادگیری با ۰,۱ در هر ۲۰۰۰ تکرار. ما از دور حرکت 0.9 و افت وزن 0.0005 استفاده می کنیم.

پس از اینکه DCNN تنظیم شد، ما پارامترهای CRF به طور کامل متصل را در مدل معادله (۲) در کنار این خطوط از Kr¨ahenb¨uhl & Koltun (۲۰۱۱) بررسی می کنیم . ما از مقادیر پیش فرض $w_2 = 3$ و $\sigma_\gamma = 3$ استفاده می کنیم و ما جستجو می کنیم بهترین مقادیر را از w_1، σ_α و σ_β توسط اعتبار سنجی متقاطع در یک زیر مجموعه ای کوچک از مجموعه اعتبار سنجی (از ۱۰۰ تصویر استفاده می کنیم). ما طرح جستجوی دقیق را بکار می بریم.به طور خاص، محدوده جستجو اولیه از پارامترها هست

$w_1 \in [5 ؛ ۱۰]$ ، $\sigma_\alpha \in [۵۰۰ :۱۰ :۱۰۰]$ و $\sigma_\beta \in [۳ :۱ :۱۰]$ (نماد MATLAB)، و سپس ما اندازه گام های جستجو در اطراف بهترین ارزش های چرخش اول را اصلاح می کنیم. ما تعداد میانگین تکرارهای زمینه را برای ۱۰ بار برای همه آزمایش های گزارش شده ثابت کردیم.

ارزیابی بر روی مجموعه اعتبار سنجی: ما اکثر ارزیابی هایمان را در مجموعه 'PASCAL 'val اجرا کردیم، آموزش مدل ما را بر روی مجموعه 'PASCAL 'train افزوده شده. همانطور که در قسمت (a)۱ نشان داده شده است ، ترکیبCRF به طور کامل متصل به مدل ما (مشخص شده توسط DeepLab-CRF) افزایش عملکرد قابل توجهی را به ارمغان می آورد، در حدود ۴٪ بهبود بیش از DeepLab. ما یادآوری می کنیم که کار Kr¨ahenb¨uhl & Koltun(۲۰۱۱) نتایج ۲۷/۶ ٪ از TextonBoost (Shotton و همکاران، ۲۰۰۹) را به ۲۹/۱٪ بهبود داد که این بهبود باعث می شود ما در اینجا (از ۵۹/۸٪ به ۶۳/۷٪) موثرترین را گزارش کنیم.

با توجه به نتایج کیفی، ما مقایسه های بصری بین DeepLab و DeepLab-CRF ارائه می کنیم در شکل ۷. استفاده از یک CRF کاملا متصل به طور قابل توجهی نتایج را بهبود می بخشد، به مدل اجازه می دهد به طور دقیق مرزهای شی پیچیده را جذب کند.

ویژگی های چندگانه : ما همچنین از ویژگی های لایه های میانبری مانند Hariharan و همکاران (a ۲۰۱۴) بهره می بریم ؛ .Long et al. (2014) همانطور که در قسمت (a)۱ نشان داده شده است ، اضافه کردن ویژگی های چند بعدی به

Method	kernel size	input stride	receptive field	# parameters	mean IOU (%)	Training speed (img/sec)
DeepLab-CRF-7x7	7×7	4	224	134.3M	67.64	1.44
DeepLab-CRF	4×4	4	128	65.1M	63.74	2.90
DeepLab-CRF-4x4	4×4	8	224	65.1M	67.14	2.90
DeepLab-CRF-LargeFOV	3×3	12	224	20.5M	67.64	4.84

جدول ۲: اثر میدان دید. ما عملکرد (پس از CRF) و سرعت آموزش را در مجموعه 'PASCAL VOC 2012 'val نشان می دهیم

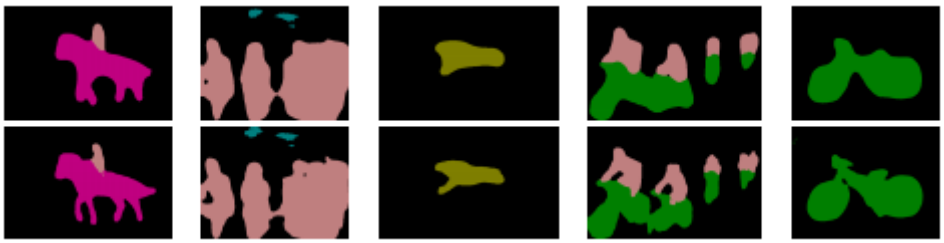
به عنوان تابع (۱) اندازه هسته اولین لایه کامل متصل شده، (۲)

مقدار گام ورودی به کار رفته در الگوریتم atrous.

مدل DeepLab ما (به عنوان DeepLab-MSc نامیده می شود) عملکرد را حدود ۱/۵٪ بهبود می بخشد و علاوه بر این،تاسیس CRF به طور کامل متصل شده (به عنوان DeepLab-MSC-CRF نامیده شده) به میزان ۴٪ بهبود میبابد.مقایسه کیفی بین DeepLab و DeepLab-MSc در شکل ۴ نشان داده شده است.ویژگی های چند مقیاس می تواند کمی مرزهای شیء را اصلاح کند.

فیلد نمایش الگوریتم "انعطاف پذیری" که ما استفاده می کنیم، به ما اجازه می دهد تا به طور داوطلبانه فیلد نمایش (FOV) مدل ها با تنظیم گام ورودی را کنترل کنیم، همانطور که در شکل ۱ نشان داده شده است. در Tab. 2، ما با چندین اندازه هسته و گام ورودی در اولین لایه کامل متصل شده آزمایش کردیم. روش،-DeepLab CRF-7x7، اصلاح مستقیم از شبکه VGG-16 است که در آن اندازه هسته ۷×۷ و گام ورودی ۴ است. این مدل

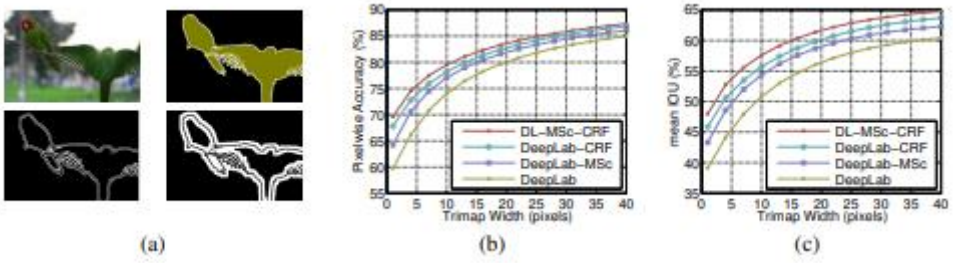
عملکرد ٪۶۷/۶۴ در مجموعه "val" را ارائه می دهد، اما این نسبتا کند است(در طول تمرین ۱/۴۴ عکس در هر ثانیه). ما سرعت مدل را تا ۲/۹ عکس در ثانیه با کاهش اندازه هسته به ۴×۴ بهبود داده ایم. ما با دو نوع مختلف شبکه با اندازه های مختلف FOV آزمایش کرده ایم ، DeepLab-CRF و DeepLab-CRF-4x4؛ دومی دارای FOV بزرگ (به عنوان مثال گام ورودی بزرگ) و عملکرد بهتر را به دست می آوریم. در نهایت، ما اندازه هسته ۳×۳ و گام ورودی ۱۲ را به کار می بریم، و همچنین اندازه های فیلتر را از ۴۰۹۶ به ۱۰۲۴ برای دو لایه جدید تغییر می دهیم. جالب توجه است، مدل نتیجه DeepLab-CRF-LargeFOV با عملکرد CRF-DeepLab 7x7 گران مطابقت دارد. در همان زمان،این ۳۶ /۳ بار سریعتر اجرا می شود و پارامترهای قابل ملاحظه کمتری دارد(۲۰،۵ مگابایت به جای ۱۳۴،۳ مگابایت) .عملکرد چندین مدل مختلف در تابلو۱ خلاصه شده است. ، سود بهره برداری ویژگی های چند مقیاس و بزرگ FOV را نشان می دهد.



شکل ۴: شامل ویژگی های چند مقیاس، تقسیم بندی مرز را بهبود می بخشد. ما نتایج بدست آمده از DeepLab و DeepLab-MSc در ردیف اول و دوم به ترتیب نشان می دهیم. بهترین مشاهده در رنگ.

میانگین پیکسل IOU در امتداد محدوده شیئ برای تعیین دقت مدل پیشنهادی در نزدیکی مرزهای شی، ما دقت تقسیم بندی را با یک آزمایش مشابه با Kohli و همکاران ارزیابی می کنیم.(۲۰۰۹)؛ Kr¨ahenb¨uhl & Koltun (۲۰۱۱). به طور خاص، از برچسب "void" که در مجموعه val مشخص شده است استفاده می کنیم.که معمولا در اطراف مرزهای شیء رخ می دهد. ما میانگین IOU را برای آن پیکسل هایی که در داخل یک باند باریک (نامیده می شود trimap) از برچسب های "void" واقع شده است محاسبه کردیم. همانطور که در شکل ۵ نشان داده شده است، بهره برداری ویژگی های چند مقیاس از لایه های متوسط و تصفیه نتایج تقسیم بندی با CRF طور کامل متصل شده به طور قابل توجهی نتایج اطراف مرزهای شی را بهبود می بخشد.

مقایسه با State-of-art در شکل ۶، ما از لحاظ کیفیت مدل پیشنهادیمان ، مدل CRF-DeepLab را با دو مدل پیشرفته: FCN-8s (Long et al.2014)، و TTI-Zoomout-16 (Mostabi و همکاران، ۲۰۱۴) در مجموعه "val" (نتایج از مقالات آنها استخراج شده)مقایسه کردیم. مدل ما قادر به ترسیم مرزهای شیء پیچیده است.



شکل ۵: (a) برخی از نمونه های trimap (بالا سمت چپ: تصویر. بالا سمت راست: زمین واقعی. پایین سمت چپ: trimap

از ۲ پیکسل. پایین سمت راست: trimap از ۱۰ پیکسل). کیفیت نتیجه تقسیم بندی در یک گروه در اطراف مرزهای شی برای روش های پیشنهادی. (b) دقت پیکسل. (c) پیکسل به معنای IOU.

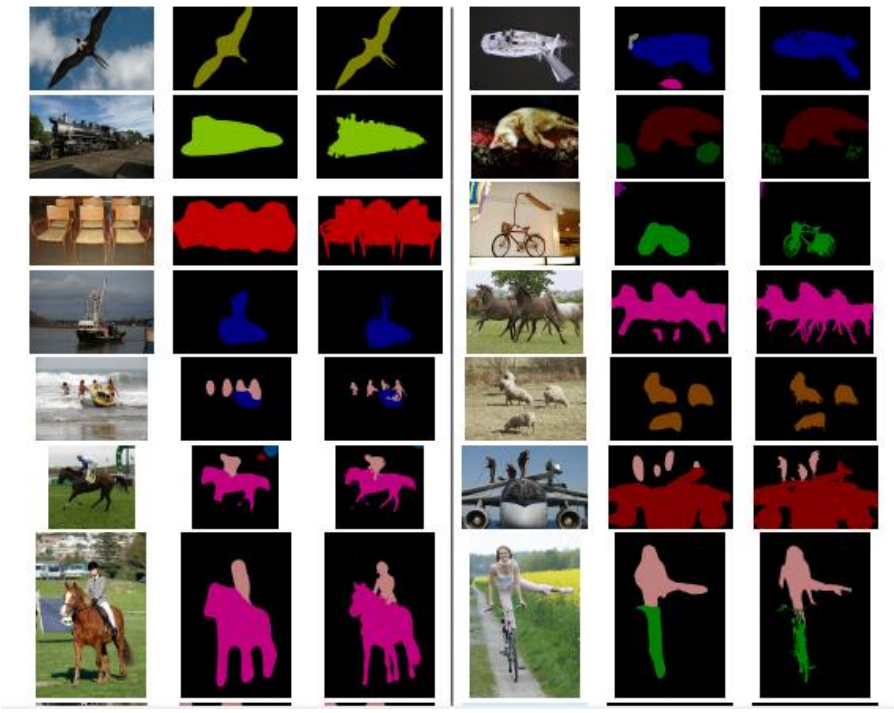
شکل ۶: مقایسات با مدل های state-of-the-art در مجموعه val. ردیف اول: تصاویر. سطر دوم: حقایق زمین ردیف سوم: سایر مدل های اخیر (سمت چپ: FCN-8s، راست: TTI-Zoomout-16). ردیف چهارم: DeepLab-CRF ما. بهترین رنگ مشاهده شده.

بازتولید پذیری ما روش های پیشنهادی را با گسترش Caffe framework عالی انجام داده ایم (Jia و همکاران، ۲۰۱۴). ما کد منبع، فایل های پیکربندی و مدل های آموزش یافته مان را به اشتراک می گذاریم که اجازه می دهیم باز تولید نتایج در این مقاله را در وب سایت همراه

https://bitbucket.org/deeplab/deeplab-public

نتایج مجموعه آزمون قرار دادن مجموعه گزینه های مدل ما را بر روی مجموعه اعتبار سنجی، ما گونه های مدلمان را در مجموعه آزمون رسمی "PASCAL VOC 2012" ارزیابی می کنیم. همانطور که در Tab۳ نشان داده شده است ، مدل های DeepLab-CRF و DeepLab- MSC-CRF ما عملکرد را به ترتیب از ۶۶/۴ ٪ و ۶۷/۱ ٪ به معنی IOU[1] می رساند. مدل های ما همه دیگر مدل های پیشرفته تر را اجرا نمی کند (به ویژه (۲۰۱۴ TTI-Zoomout-16 (Mostajabi et al ،

FCN-8s (Long et al. 2014) وMSRA-CFM (Dai et al. 2014). هنگامی که ما FOV را از مدل ها افزایش می دهیم، DeepLab-CRF-LargeFOV عملکرد از ۷۰/۳ ٪ بازدهی می دهد، همانند DeepLab-CRF-7×7 در حالی که سرعت آموزش آن سریع تر است. علاوه بر این، بهترین مدل ما، DeepLab-MSc-CRF-LargeFOV،با بهره گیری از ویژگی های چند منظوره و FOV بزرگ، بهترین عملکرد ۷۱/۶٪ را به دست می آورد.



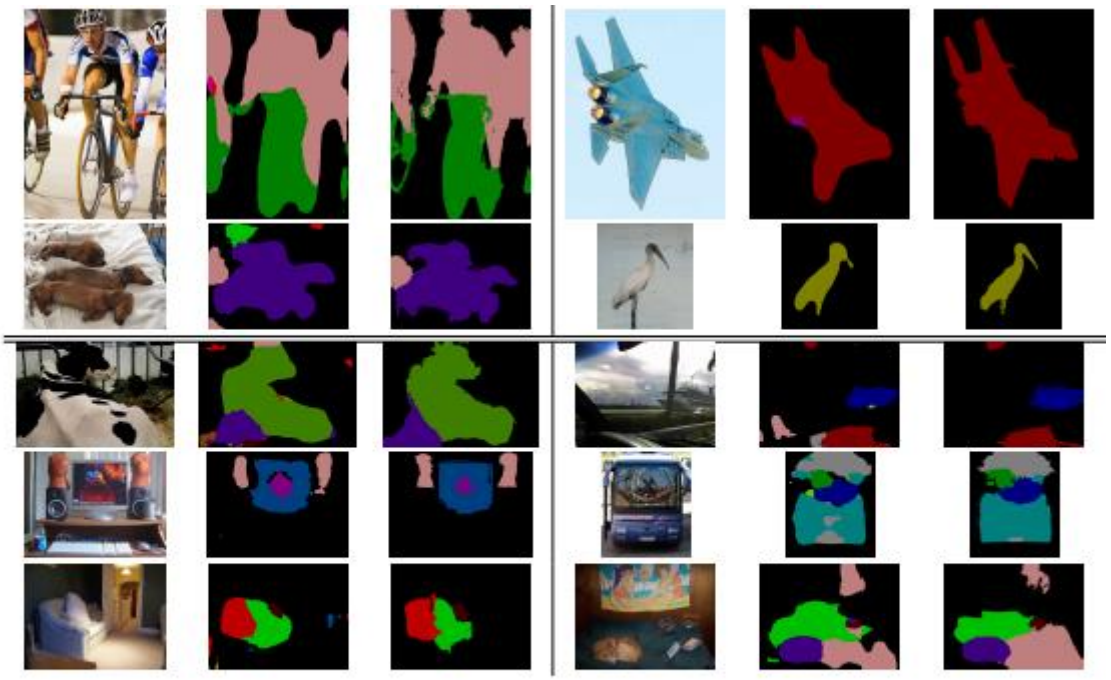


Figure 7: Visualization results on VOC 2012-val. For each row, we show the input image, the segmentation result delivered by the DCNN (DeepLab), and the refined segmentation result of the Fully Connected CRF (DeepLab-CRF). We show our failure modes in the last three rows. Best viewed in color.

شکل ۷: نتایج تجسم در VOC 2012-val. برای هر ردیف، تصویر ورودی را نمایش می دهیم .نتیجه تقسیم بندی توسط DCNN (DeepLab) ارائه شده و نتایج تقسیم بندی تصفیه شده از CRF (DeepLab-CRF) کاملا متصل. حالتهای شکست ما در سه ردیف آخر نمایش داده می شود. بهترین نمایش در رنگ

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
MSRA-CFM	-	75.7	26.7	69.5	48.8	65.6	81.0	69.2	73.3	30.0	68.7	51.5	69.1	68.1	71.7	67.5	50.4	66.5	44.4	58.9	53.5	61.8
FCN-8s	-	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
TTI-Zoomout-16	89.8	81.9	35.1	78.2	57.4	56.5	80.5	74.0	79.8	22.4	69.6	53.7	74.0	76.0	76.6	68.8	44.3	70.2	40.2	68.9	55.3	64.4
DeepLab-CRF	92.1	78.4	33.1	78.2	55.6	65.3	81.3	75.5	78.6	25.3	69.2	52.7	75.2	69.0	79.1	77.6	54.7	78.3	45.1	73.3	56.2	66.4
DeepLab-MSc-CRF	92.6	80.4	36.8	77.4	55.2	66.4	81.5	77.5	78.9	27.1	68.2	52.7	74.3	69.6	79.4	79.0	56.9	78.8	45.2	72.7	59.3	67.1
DeepLab-CRF-7x7	92.8	83.9	36.6	77.5	58.4	**68.0**	84.6	**79.7**	83.1	29.5	**74.6**	59.3	78.9	76.0	82.1	80.6	**60.3**	81.7	49.2	**78.0**	60.7	70.3
DeepLab-CRF-LargeFOV	92.6	83.5	36.6	**82.5**	62.3	66.5	**85.4**	78.5	**83.7**	30.4	72.9	**60.4**	78.5	75.5	82.1	79.7	58.2	82.0	48.8	73.7	63.3	70.3
DeepLab-MSc-CRF-LargeFOV	**93.1**	**84.4**	**54.5**	81.5	**63.6**	65.9	85.1	79.1	83.4	**30.7**	74.1	59.8	**79.0**	**76.1**	**83.2**	**80.8**	59.7	**82.2**	**50.4**	73.1	**63.7**	**71.6**

جدول ۳: برچسب گذاری IOU (٪) در مجموعه آزمون PASCAL VOC 2012، با استفاده از مجموعه آموزشی برای آموزش.

۶ بحث

کار ما ترکیبی از ایده های شبکه های عصبی پیچیده عمیق و زمینه های تصادفی شرطی به طور کامل متصل است، ارائه یک روش جدید قادر به تولید پیش بینی های دقیق معنایی و نقشه های تقسیم بندی مفصل است ، در حالی که از نظر محاسباتی کارآمد باشد. نتایج تجربی ما نشان می دهد که روش پیشنهادی به طور قابل ملاحظه

ای پیشرفت داشته این state-of-art در وظیفه تقسیم بندی تصویر معنایی PASCAL VOC 2012 در حال رقابت است.

در مدل ما چندین جنبه وجود دارد که ما قصد داریم آن را اصلاح کنیم، مثلا به طور کامل این دو اجزای اصلی را با هم ادغام کنیم (CNN و CRF) و تمام سیستم را در حالت end-to-end آموزش دهیم، شبیه بهKr¨ahenb¨uhl & Koltun (۲۰۱۳)؛ چن و همکاران (۲۰۱۴)؛ ژنگ و همکاران (۲۰۱۵). ما همچنین قصد داریم با مجموعه داده های بیشتر آزمایش کنیم و روشمان را با منابع دیگر داده مانند نقشه های عمیق و یا فیلم ها امتحان کنیم. به تازگی ما آموزش مدل پیشنهادی را با حاشیه نویسی ضعیف تحت نظارت دنبال کرده ایم، به شکل محدود کردن جعبه یا برچسب سطح تصویر (.Papandreou et al، 2015).

در سطح بالایی، کار ما در تقاطع شبکه های عصبی پیچیده و مدل های گرافیکی احتمالی قرار دارد ما قصد داریم تا بیشتر در مورد تعامل میان این دو طبقه قدرتمند از روش ها تحقیق کنیم و پتانسیل خود را برای به چالش کشیدن وظایف دیداری کامپیوتر کشف کنیم.

تعهدات

این کار بخشی از حمایتARO 62250-CS ، NIH Grant 5R01EY022247-03، پروژه اتحادیه اروپاRECONFIG FP7-ICT-600825 و پروژه اتحادیه اروپا MOBOT FP7-ICT-2011-600796. ما همچنین از حمایت شرکت NVIDIA با اهدای GPU های مورد استفاده برای این تحقیق سپاسگزاریم. ما می خواهیم از بازرسان ناشناس برای نظرات مفصل و بازخورد سازنده تشکر کنیم.

REFERENCES

Adams, A., Baek, J., and Davis, M. A. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*, 2010.

Arbeláez, P., Pont-Tuset, J., Barron, J. T., Marques, F., and Malik, J. Multiscale combinatorial grouping. In *CVPR*, 2014.

Bell, S., Upchurch, P., Snavely, N., and Bala, K. Material recognition in the wild with the materials in context database. *arXiv:1412.0623*, 2014.

Carreira, J. and Sminchisescu, C. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *PAMI*, 2012.

Carreira, J., Caseiro, R., Batista, J., and Sminchisescu, C. Semantic segmentation with second-order pooling. In *ECCV*, 2012.

Chen, L.-C., Schwing, A., Yuille, A., and Urtasun, R. Learning deep structured models. *arXiv:1407.2538*, 2014.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

Chen, X. and Yuille, A. L. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014.

Cogswell, M., Lin, X., Purushwalkam, S., and Batra, D. Combining the best of graphical models and convnets for semantic segmentation. *arXiv:1412.4313*, 2014.

Dai, J., He, K., and Sun, J. Convolutional feature masking for joint object and stuff segmentation. *arXiv:1412.1283*, 2014.

Delong, A., Osokin, A., Isack, H. N., and Boykov, Y. Fast approximate energy minimization with label costs. *IJCV*, 2012.

Eigen, D. and Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *arXiv:1411.4734*, 2014.

Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserma, A. The pascal visual object classes challenge a retrospective. *IJCV*, 2014.

Farabet, C., Couprie, C., Najman, L., and LeCun, Y. Learning hierarchical features for scene labeling. *PAMI*, 2013.

Geiger, D. and Girosi, F. Parallel and deterministic algorithms from mrfs: Surface reconstruction. *PAMI*, 13(5):401–412, 1991.

Geiger, D. and Yuille, A. A common framework for image segmentation. *IJCV*, 6(3):227–243, 1991.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

Giusti, A., Ciresan, D., Masci, J., Gambardella, L., and Schmidhuber, J. Fast image scanning with deep max-pooling convolutional neural networks. In *ICIP*, 2013.

Gonfaus, J. M., Boix, X., Van de Weijer, J., Bagdanov, A. D., Serrat, J., and Gonzalez, J. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010.

Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., and Malik, J. Semantic contours from inverse detectors. In *ICCV*, 2011.

Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. Hypercolumns for object segmentation and fine-grained localization. *arXiv:1411.5752*, 2014a.

Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. Simultaneous detection and segmentation. In *ECCV*, 2014b.

He, X., Zemel, R. S., and Carreira-Perpindn, M. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

Kohli, P., Ladicky, L., and Torr, P. H. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009.

Kokkinos, I., Deriche, R., Faugeras, O., and Maragos, P. Computational analysis and learning for a biologically motivated model of boundary detection. *Neurocomputing*, 71(10):1798–1812, 2008.

Krähenbühl, P. and Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

Krähenbühl, P. and Koltun, V. Parameter learning and convergent inference for dense random fields. In *ICML*, 2013.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2013.

Ladicky, L., Russell, C., Kohli, P., and Torr, P. H. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proc. IEEE*, 1998.

Lempitsky, V., Vedaldi, A., and Zisserman, A. Pylon model for semantic segmentation. In *NIPS*, 2011.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. *arXiv:1411.4038*, 2014.

Lucchi, A., Li, Y., Boix, X., Smith, K., and Fua, P. Are spatial and global constraints really necessary for segmentation? In *ICCV*, 2011.

Mallat, S. *A Wavelet Tour of Signal Processing*. Acad. Press, 2 edition, 1999.

Mostajabi, M., Yadollahpour, P., and Shakhnarovich, G. Feedforward semantic segmentation with zoom-out features. *arXiv:1412.0774*, 2014.

Papandreou, G., Kokkinos, I., and Savalle, P.-A. Untangling local and global deformations in deep convolutional networks for image classification and sliding window detection. *arXiv:1412.0296*, 2014.

Papandreou, G., Chen, L.-C., Murphy, K., and Yuille, A. L. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. *arXiv:1502.02734*, 2015.

Rother, C., Kolmogorov, V., and Blake, A. Grabcut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv:1312.6229*, 2013.

Shotton, J., Winn, J., Rother, C., and Criminisi, A. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. *arXiv:1409.4842*, 2014.

Tompson, J., Jain, A., LeCun, Y., and Bregler, C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *NIPS*, 2014.

Uijlings, J., van de Sande, K., Gevers, T., and Smeulders, A. Selective search for object recognition. *IJCV*, 2013.

Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., and Yuille, A. Towards unified depth and semantic prediction from a single image. In *CVPR*, 2015.

Yadollahpour, P., Batra, D., and Shakhnarovich, G. Discriminative re-ranking of diverse segmentations. In *CVPR*, 2013.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

Zhang, N., Donahue, J., Girshick, R., and Darrell, T. Part-based r-cnns for fine-grained category detection. In *ECCV*, 2014.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. Conditional random fields as recurrent neural networks. *arXiv:1502.03240*, 2015.