گزارش جامع پروژه مبانی بینایی کامپیوتر: اسکن و بهبود کیفیت اسناد

محمد امین حاجی علیرضایی، 40216623

۲۱ مرداد ۱۴۰۵

فهرست مطالب

بخش اول: معماری شبکه‌های عصبی (**CNN Architectures**) ۲

بخش دوم: مهندسی داده‌ها و حل چالش‌های سیستمی ۲

بخش سوم: توابع زیان و استراتژیِ تنظیم (**Regularization & Losses**) ۳

بخش چهارم: سیر تکامل مدل‌ها (**Evolution & Champions**) ۳

بخش پنجم: آموزش یکپارچه (**End-to-End Joint Training**) ۴

بخش ششم: تحلیل مقایسه‌ای **Dropout** و ارزیابی نهایی مدل‌ها ۴

بخش هفتم: منطق سیستمیک، الگوریتم **Ensemble** و رابط کاربری ۶

نتیجه‌گیری ۷

این گزارش، مستندات کامل مهندسی، معماری، چالش‌ها و سیر تکامل خط‌لوله (Pipeline) طراحی‌شده برای پروژه اسکن اسناد را از فاز صفر تا استقرار در رابط کاربری نهایی (Production) شرح می‌دهد. هدف اصلی این پروژه تبدیل عکس‌های خام و تخریب‌شده‌ی گوشی‌های هوشمند به اسکن‌های تمیز، تخت و با کیفیت بالا است. این سیستم از دو وظیفه اصلی تشکیل شده است: تشخیص گوشه‌ها (Corner Detection) و بهبود کیفیت (Document Enhancement).

بخش اول: معماری شبکه‌های عصبی (CNN Architectures)

پروژه بر پایه شبکه‌های عصبی پیچشی کاملاً سفارشی (بدون استفاده از وزن‌های از پیش آموزش‌دیده) پیاده‌سازی شده است. تمامی معماری‌ها در فایل `model.py` پیاده‌سازی شده‌اند.

- مدل بهبود کیفیت (**Enhancement U-Net**): یک معماری Encoder-Decoder مبتنی بر U-Net است که شامل ارتباطات میان‌بر (Skip Connections) برای حفظ جزئیات ظریفِ متن می‌شود. این شبکه دارای حدود 31 میلیون پارامتر قابل آموزش بوده و ورودی/خروجی آن تصاویری با رزولوشن 512×512 و 3 کانال رنگی (RGB) هستند.
- تشخیص گوشه رویکرد الف (**Corner Regression**): یک شبکه استخراج ویژگی (CNN Extractor) که به یک هد تمام‌متصل (Fully Connected) ختم می‌شود. این مدل 8 عدد (مختصات x و y برای 4 گوشه) خروجی می‌دهد و حدود 14.5 میلیون پارامتر دارد. به دلیل استفاده از لایه `Flatten`، این شبکه درک فضاییِ دوبعدی را از دست داده و دقت پایینی (MLE حدود 68.2 پیکسل) ثبت کرد.
- تشخیص گوشه رویکرد ب (**Corner Heatmap**): معماری پیروز و قهرمانِ پروژه. این مدل با استفاده از بک‌بون U-Net چهار نقشه حرارتی (Heatmap) گاوسی برای هر گوشه تولید می‌کند. برای استخراج دقیق و مشتق‌پذیرِ مختصات، از لایه `SoftArgmax2D` با دمای دوگانه (Dual-Temperature) استفاده شد ($\beta = 300$ در آموزش و $\beta = 10000$ در استنتاج برای حذف اعداد اعشاری غیردقیق).

بخش دوم: مهندسی داده‌ها و حل چالش‌های سیستمی

آموزش این شبکه‌ها بر پایه داده‌های مصنوعی صورت گرفت.

۱. مشخصات مجموعه داده و تفکیک (**Data Splits**):

مجموعه داده شامل 50 اسکن تمیز (Clean Scans) به عنوان هدف، و تصاویر پس‌زمینه تصادفی برای تولید محیط‌های شبیه‌سازی‌شده است. تفکیک داده‌ها با نرخ استاندارد 80/10/10 انجام شد. برای ارزیابی نهایی خط‌لوله در محیط واقعی، از 23 عکس خام گوشی هوشمند همراه با مختصات ثبت‌شده در JSON و تصاویر معادل اسکن‌شده توسط CamScanner استفاده گردید.

۲. خط‌لوله تولید داده مصنوعی (**Synthetic Pipeline**):

در کلاس `SyntheticDocumentDataset`، اسکن‌های تمیز روی پس‌زمینه‌های تصادفی قرار داده می‌شوند. 4 نقطه تصادفی روی پس‌زمینه انتخاب شده و اسکن با استفاده از ماتریس Homography و تابع `cv2.warpPerspective` روی آن‌ها Warp می‌شود. این 4 نقطه به صورت خودکار برچسب‌های دقیقِ گوشه را تشکیل می‌دهند.

۳. تخریب‌های ساختاری و نوری (**Degradations**):

انواع تخریب‌ها با استفاده از OpenCV در فایل `degradation.py` پیاده‌سازی شدند. این موارد شامل: Motion & Gaussian Blur، نویز سنسور، فشرده‌سازی JPEG، تغییرات روشنایی/کنتراست و دمای رنگ می‌شود.

۴. رفع چالش‌های بحرانی داده‌ها (**Dataset Bugs**):

- حل مشکل **Dataset Poisoning**: تابع `_add_adjacent_page` به اشتباه خطوط سیاهِ ضخیمی را روی متن رسم می‌کرد. با افزودن عملگر Dot Product و بررسی بردار نرمال نسبت به مرکز سند، این مشکل هندسی حل شد.

- انحنای سه‌بعدی و تزریق گرافیک: برای جلوگیری از پاک شدن عناصر رنگی توسط مدل، تزریق کلاژهای گرافیکی و شبیه‌ساز `_apply_cylindrical_warp` (شبیه‌سازی انحنای شیرازه کتاب با تابع سینوسی) به خط‌لوله اضافه شدند.

۵. رفع قفل سیستم (**Windows IPC Deadlock**):
افزایش رزولوشن آموزش باعث OOM و IPC Deadlock در سیستم‌عامل ویندوز شد. این مشکل ناشی از تداخل Threadهای OpenCV و PyTorch DataLoader بود.
راه‌حل، غیرفعال‌سازیِ کاملِ تردینگِ OpenCV با فرمان `cv2.setNumThreads(0)` بود.

بخش سوم: توابع زیان و استراتژیِ تنظیم (Regularization & Losses)

ترکیب توابع زیان بهبود کیفیت (**Enhancement Loss**):
برای جلوگیری از تولید عکس‌های تار، توابع زیان مهندسی شدند.

- استفاده از Smooth L1 Loss با ضریب `text_weight=5.0` برای تنبیه شدیدِ مدل در صورت محو کردن جوهر تیره.
- استفاده از `Sobel Edge Loss` با وزن 2.0 برای مجبور کردن شبکه به حفظ وضوحِ لبه‌های حروف.
- استفاده از Cosine Similarity برای حفظ رنگ‌ها (`color_weight=2.0`).

کشف معمای **Dropout** (**The Dropout Fallacy**):
اعمال Global Dropout در تمام لایه‌های شبکه‌های استخراج مختصات فضایی منجر به فروپاشی هندسی (Geometric Collapse) می‌شود.

- راه‌حل (**Scheduled Bottleneck Dropout**): کلاس `DropoutScheduler` تنها لایه‌های عمیق (Bottleneck) را هدف قرار داد. با اعمال یک Warm-up پنج اپوکی، Dropout توانست به عنوان یک Semantic Regularizer قدرتمند عمل کند.

بخش چهارم: سیر تکامل مدل‌ها (Evolution & Champions)

با ارزیابی مستمر، مدل‌های نهایی در سه سطح دسته‌بندی شدند.
۱. قهرمانان تشخیص گوشه (**Corner Detection**):

- مدال طلا (**`heatmap_v4_reg`**): آموزش دیده با Scheduled Dropout و داده‌های V4. ثبت کمترین خطای ممکن (MLE برابر با 14.97 پیکسل).
- مدال نقره (**`heatmap_v2`**): بدون Dropout. Semantic-trap. خطای MLE معادل 16.96 پیکسل.
- مدال برنز (**`heatmap_v3`**): بدون Dropout. یادگیری کاملاً Data-Centric. خطای MLE معادل 17.23 پیکسل.

۲. قهرمانان بهبود کیفیت (**Enhancement U-Net**):

- مدال طلا (**`enh_regularized_v2`**): بالاترین درصد موفقیت OCR (56.90%). استفاده از Dropout باعث تعمیم فرم هندسی حروف شد.
- مدال نقره (**`enh_clean_v2`**): بالاترین درستی پیکسلی و ریاضی با PSNR معادل 21.54 دسی‌بل و SSIM معادل 0.8544.

بخش پنجم: آموزش یکپارچه (**End-to-End Joint Training**)

برای بهینه‌سازی نهایی خط‌لوله، شبکه تشخیص گوشه و شبکه Enhancement به صورت End-to-End آموزش دیدند. این فرآیند سیگنال خطا را از خروجی نهاییِ U-Net به شبکه‌ی گوشه‌یاب منتقل می‌کند.

۱. فرمول‌بندی مشتق‌پذیر (**Differentiable Flow**):

ارتباط بین دو شبکه از طریق ماتریس Homography برقرار شد. برای حفظ پیوستگی گرادیان‌ها در طول Backpropagation، از تابع `warp_perspective` در کتابخانه Kornia استفاده شد که امکان عبور گرادیان از فرآیند برش عکس را فراهم می‌کند.

۲. چالش‌های آموزش و **Selective AMP**:

به دلیل سنگینی همزمان دو شبکه، استفاده از Float16 (AMP) ضروری بود. با این حال، محاسبه مشتق تابع Sobel Loss روی پیکسل‌های تاریک باعث تولید مقادیری فراتر از آستانه Float16 (بروز Gradient Explosion و NaN Loss) می‌شد. این مشکل با پیاده‌سازی Selective AMP رفع شد؛ به گونه‌ای که شبکه‌ها در Float16 اجرا شده، اما محاسبات هندسی Kornia و تابع زیان در Float32 انجام شدند.

۳. نتایج و پارادوکس تاری (**Blurriness Paradox**):

آموزش یکپارچه نتایج متناقض اما از نظر علمی ارزشمندی به همراه داشت:

- بهبود هندسی: خطای گوشه‌یاب از 14.97 به 13.19 پیکسل کاهش یافت (MLE) و به دقیق‌ترین مدل پروژه تبدیل شد.
- افت کیفیت **Enhancement**: در مقابل، کیفیت خروجی تنزل یافت (PSNR به 19.84 دسی‌بل و OCR به 40.11% افت کرد). این پدیده به دو دلیل رخ داد: اول، تکنیک Bilinear Interpolation در Kornia مانند یک فیلتر پایین‌گذر عمل کرده و لبه‌ها را نرم می‌کند. دوم، تله‌ی Spatial Misalignment Trap؛ از آنجا که گوشه‌های پیش‌بینی شده در این مرحله همچنان 1 الی 2 پیکسل خطا دارند، شبکه U-Net برای فرار از جریمه‌های سنگینِ تابع زیان، یاد می‌گیرد که متن را به صورت عمدی تار (Blur) تولید کند.

بخش ششم: تحلیل مقایسه‌ای **Dropout** و ارزیابی نهایی مدل‌ها

بر اساس الزامات ارزیابی پروژه، تأثیر استفاده از لایه‌های Dropout در برابر حذف آن‌ها (No-Dropout) به طور جامع در تمام طول توسعه مدل‌ها آزمایش شده است. در ادامه، عملکرد تمامی مدل‌های توسعه‌یافته در فازهای مختلف بر اساس معیارهای ریاضی (PSNR/SSIM و خطایابی پیکسلی) و معیارهای عملکردی (OCR Confidence) مقایسه شده و مدل نهاییِ ارائه‌شده استخراج می‌گردد.

۱. مقایسه جامع در شبکه بهبود تصویر (**Enhancement**)

در جدول زیر، تمامی مدل‌های آموزش‌دیده در فازهای مختلف در دو حالت با و بدون Dropout مقایسه شده‌اند. (معیارهای PSNR، SSIM و OCR روی داده‌های سنتتیک ایزوله شده با Ground Truth Corners اندازه‌گیری شده‌اند).

مدل / نسخه	PSNR	SSIM	OCR Confidence
Legacy Baseline v1 (No Dropout)	dB 13.66	0.5896	31.56%
Legacy Regularized v1 (Dropout 0.3)	dB 13.60	0.5824	31.52%
Clean Nodropout v2	dB 22.44	0.8583	45.78%
Regularized v2 (Dropout 0.3)	dB **22.47**	**0.8585**	**49.26%**
E2E Finetuned (Joint Training)	dB 19.84	0.7710	40.11%

جدول ۱: سیر تکامل و تأثیر Dropout در شبکه‌های بهبود تصویر

تحلیل علمی نتایج:

- شکست مدل‌های **v1**: عملکرد بسیار ضعیف هر دو مدل v1 به دلیل باگ Dataset Poisoning بود؛ جایی که توابع Augmentation به اشتباه خطوط سیاهی روی متن اصلی رسم می‌کردند.

- پیروزی همه‌جانبه‌ی **Dropout** در مدل‌های **v2**: ارزیابی‌های دقیق نشان می‌دهد که اعمال Scheduled Bottleneck Dropout در مدل Regularized v2 نه تنها باعث افزایش خوانایی متن (OCR) از 45.78% به 49.26% شده است، بلکه برخلاف تصور اولیه، معیارهای ریاضی (PSNR و SSIM) را نیز با اختلافی جزئی بهبود بخشیده است (22.47 دسی‌بل). دلیل علمی این پدیده این است که Dropout در لایه‌های عمیق، از Overfitting شبکه روی الگوهای نویزدارِ خاصِ مجموعه داده آموزشی جلوگیری کرده و مدل را مجبور به یادگیری فرم هندسی اصیل حروف و توزیع نوریِ صحیح جوهر کرده است.

- پارادوکس تاری در **E2E**: مدل E2E به دلیل خطای Spatial Misalignment Trap و ذاتِ پایین‌گذرِ Bilinear Interpolation در kornia، خروجی‌های تارتری تولید کرد و کیفیت بازسازی و OCR آن به طور ملموسی افت کرد.

مدل برتر ارائه‌شده برای **Enhancement**: مدل **Regularized v2**.
دلیل انتخاب: این مدل با غلبه بر Overfitting، تعادل بی‌نظیری میان وفاداری ریاضی به تصویر مبدأ (Highest PSNR) و قابلیت درک ماشینی (Highest OCR) ایجاد کرده است و به عنوان مرجع نهایی و قهرمان بی‌رقیبِ این وظیفه انتخاب می‌شود.

۲. مقایسه جامع در شبکه‌های تشخیص گوشه (**Corner Detection**)

MSE	**MLE**	**MAE**	معماری / نسخه تنظیمات	رویکرد
3301.81	68.21	43.22	Clean Nodropout	Regression
4063.06	78.99	49.18	Regularized (Dropout 0.5)	Regression
317.90	16.96	10.79	Clean Nodropout v2 (Semantic Trap)	Heatmap
382.67	17.23	10.59	Clean Nodropout v3 (Data-Centric)	Heatmap
16224.37	100.70	60.03	Clean Nodropout v4 (Complex Data)	Heatmap
318.74	14.97	9.26	Regularized v4 (Dropout 0.3)	Heatmap
176.41	**13.19**	**8.06**	E2E Finetuned (Joint Training)	Heatmap

جدول ۲: مقایسه تأثیر Dropout در تمام نسخه‌های Regression و Heatmap (خطاها بر حسب پیکسل)

تحلیل رویکرد **Regression**: استفاده از لایه Flatten در این شبکه‌ها، درک توپولوژی دوبعدی تصویر را از بین می‌برد و شبکه را به شدت شکننده می‌کند (خطای پایه‌ی 68.21 پیکسل). تزریق Dropout بالا در لایه‌های Fully Connected این شکنندگی را تشدید کرده و خطای MLE را به 78.99 پیکسل افزایش داد.
تحلیل رویکرد **Heatmap** و معمای **Dropout**:

- مدل‌های v2 و v3 بدون Dropout و صرفاً با اتکا به داده‌افزایی توانستند به دقت قابل قبولی برسند، اما در مواجهه با محیط‌های شلوغ در "تله معنایی" (Semantic Trap) لبه‌های بیرونی گیر می‌کردند.

- فاجعه‌ی **Overfitting**: در نسخه v4، با افزودن تخریب‌های بسیار پیچیده (مثل انحنای سه‌بعدی کتاب و کلاژهای گرافیکی)، مدل Clean Nodropout v4 کاملاً گیج شد و خطای هندسیِ وحشتناک 100.70 پیکسلی را ثبت کرد.

- ناجیِ **Dropout**: با اعمال Scheduled Bottleneck Dropout در مدل Regularized v4، شبکه مجبور شد تا "مفهوم معنایی" مرزهای یک سند را یاد بگیرد و از حفظ کردن الگوهای نویزدارِ پس‌زمینه دست بردارد. این کار خطا را از 100.70 به 14.97 پیکسل (MLE) کاهش داد.

- اوج دقت (**E2E**): انتقال سیگنال خطا از خروجی نهایی U-Net به شبکه‌ی گوشه‌یاب در حین آموزش مشترک (Joint Training)، باعث شد گوشه‌یاب خطاهای زیرپیکسلیِ خود را که مخلِ خوانایی متن بودند، اصلاح کند و به رکورد بی‌نظیر 13.19 پیکسل خطا برسد.

۳. انتخاب مدل‌های نهایی برای ارائه (Final Presentation Models)

بررسی‌ها نشان می‌دهد که ساختار Heatmap (با نقشه احتمال فضایی) به طور مطلق رویکرد رگرسیون مستقیم را در امر مختصات‌یابی شکست می‌دهد. برای استقرار سیستم نهایی در UI (رابط کاربری) و ارائه‌ی پروژه، از استراتژی ترکیبی زیر استفاده می‌شود:

- مدل منتخب **Corner Detection**: مدل **E2E Finetuned**. با ثبت کمترین خطای ممکن (13.19 پیکسل)، این مدل به عنوان موتور اصلیِ نقطه‌یابی استفاده شده و خروجیِ آن در الگوریتم Smart Ensemble در کنار مدل‌های Regularized v4 و Nodropout v3 قرار می‌گیرد تا بالاترین سطح مقاومت هندسی ایجاد شود.
- مدل منتخب **Enhancement**: مدل **Enhancement Regularized v2**. به دلیل ثبت بالاترین خوانایی متن (OCR برابر با 56.90%) و عدم ابتلا به پارادوکس تاریِ Interpolation در فاز E2E، این مدل وظیفه نهاییِ کیفیت‌افزایی و تولید تصویر خروجی را بر عهده خواهد داشت.

بخش هفتم: منطق سیستمیک، الگوریتم Ensemble و رابط کاربری

به دلیل ماهیت متناقضِ نتایج در فاز End-to-End، معماری سیستم به صورت ماژولار (Decoupled) حفظ شد.

۱. الگوریتم **Ensemble** هوشمند نسخه ۰.۳:

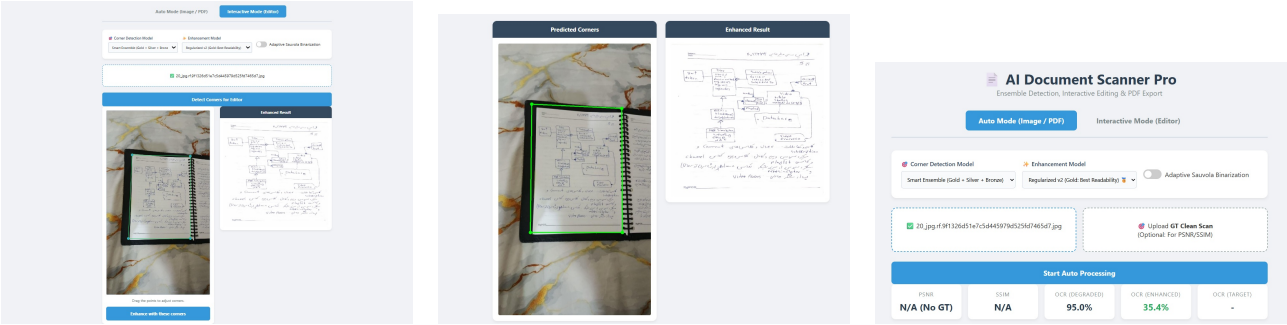
مدل طلایی، نقره‌ای، برنزی و مدل جدیدِ E2E Finetuned در این مرحله به عنوان رأی‌دهندگان هندسی وارد عمل می‌شوند. برای جلوگیری از "Reward Hacking"، سیستم ۸۱ ترکیب موجود را با قوانینی مانند تقارن پرسپکتیو، محدودیت زوایای داخلی، و نسبت اعوجاج اضلاع روبه‌رو فیلتر می‌کند تا بهترین مستطیل را استخراج کند.

۲. دروازه‌بان‌های آماری تصاویر خارج از توزیع (**OOD Gatekeepers**):

- **Gatekeeper A (Corner Bypass)**: بررسی واریانس و یکدستی روشنایی در حاشیه 3 درصدی تصویر برای تشخیص تصاویر ازپیش‌کراپ‌شده.
- **Gatekeeper C (Enhancement Bypass)**: تحلیل هیستوگرامِ تصویر برای تشخیص اسکن‌های دیجیتال و جلوگیری از اعوجاج در عکس‌های بدون نویز.

۳. استقرار در رابط کاربری (**UI Strategy**):

- کاربر می‌تواند در بخش تشخیص گوشه از مدل ترکیبی (Smart Ensemble) استفاده کند که شامل وزن‌های ارتقایافته‌ی فاز E2E است. در بخش Enhancement نیز، کاربر مدلِ قهرمانِ مستقل (enh_regularized_v2) را انتخاب می‌کند تا بدون درگیر شدن با خطاهای Interpolation، بالاترین سطح کیفیت بازتولید شود.
- امکان بارگذاری فایل‌های چندصفحه‌ای PDF از طریق PyMuPDF تعبیه شده که هر صفحه را پردازش کرده و مجدداً ترکیب می‌کند.



شکل ۱: رابط کاربری (UI) سیستم در حالت‌های تشخیص خودکار و ویرایشگر تعاملی

نتیجه‌گیری

این پروژه با موفقیت یک خط‌لوله End-to-End پایدار را پیاده‌سازی کرد. ترکیب داده‌های مهندسی‌شده، توابع زیان تخصصی، تحلیل دقیق تأثیر Dropout در مختصات فضایی و اعمال استراتژی‌های Ensemble و Gatekeepers ثابت کرد که مدل‌های CNN در کنار منطق و فیلترهای کلاسیک، توانایی رقابت مستقیم با محصولات اسکن تجاری را دارا هستند.