پیش‌بینی نقاط خیرگی چشم در صحنه‌های متحرک با استفاده از شبکه‌های عمیق

دانشجو: نیلوفر فریدنی

استاد راهنما: دکتر محمدعلی اخایی

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

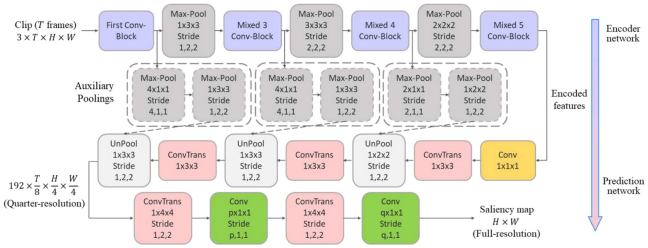
چشم انسان دریچه‌ای به سوی دنیای بیرون است و نقش اساسی در دریافت اطلاعات محیطی دارد. الگوهای دیداری، شامل حرکات چشم، نقشی کلیدی در پردازش اطلاعات بصری دارند. با وجود اهمیت بالای مطالعه الگوهای دیداری، ثبت و تحلیل دقیق آنها به دلیل هزینه‌بر بودن تجهیزات ردیابی چشم و پیچیدگی تحلیل داده‌ها، چالش‌برانگیز است. توانایی پیش‌بینی دقیق الگوهای دیداری بدون نیاز به تجهیزات ثبت داده‌ها، گامی تحول‌آفرین در زمینه‌های مختلف مانند بینایی ماشین، تعامل انسان و کامپیوتر، و فشرده‌سازی ویدیو خواهد بود. پروژه حاضر با هدف طراحی شبکه‌های عمیق برای پیش‌بینی مکان بعدی نقاط تمرکز دید انسان در تصاویر متحرک، در حال انجام است. این پروژه با پیش‌بینی مسیر حرکت چشم انسان در تصاویر متحرک با کمترین خطا، به دنبال ارتقای درک ما از نحوه تعامل انسان با محیط پیرامون و توسعه فناوری‌های نوین در حوزه‌های مختلف علمی و کاربردی است.

مدل پیشنهادی

مجموعه داده DHF1K

در این پروژه، مجموعه داده DHF1K برای آموزش و ارزیابی شبکه عصبی در پیش‌بینی مسیر حرکت چشم در تصاویر متحرک به کار گرفته شده است. این مجموعه داده، شامل ۱۰۰۰ کلیپ ویدیویی و ثبت حرکت چشم انسان در حین نمایش این ویدیوها است و به عنوان مجموعه داده معتبر در زمینه تحقیقات ردیابی چشم شناخته می‌شود. برای این مطالعه، ۵۰۰ ویدیو برای آموزش و ۱۰۰ ویدیو برای اعتبارسنجی استفاده شده است.

معماری شبکه پیشنهادی



تابع هزینه (Kullback-Leibler Divergence)

این تابع برای محاسبه اختلاف و تفاوت بین دو توزیع احتمال مورد استفاده قرار می‌گیرد. در اینجا، تابع KLD بین دو توزیع احتمال ورودی و هدف محاسبه می‌شود. این تابع معیاری از تفاوت میان دو توزیع احتمال است و مقدار آن نشان‌دهنده اندازه تفاوت میان توزیع مدل و توزیع واقعی است. در نتیجه، کمینه کردن مقدار KLD می‌تواند به ما کمک کند تا مدل را بهبود بخشیم و پیش‌بینی‌های معتبرتری داشته باشیم.

$$KLD(inp, trg) = \sum_i trg_i \log\left(\frac{trg}{inp_i + \epsilon}\right)$$

در نهایت برای اینکه دینامیک بین فریم‌ها رعایت شود تابع هزینه نهایی به شکل زیر تعریف شد:

$$Total\ Loss = KLDLoss + \alpha * TemporalConsistency$$

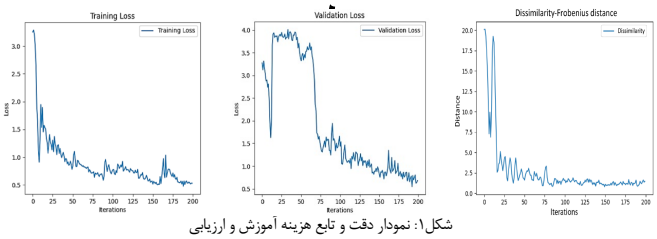
$$TemporalConsistency(inp, trg) = \frac{1}{N}\sum_i^N |inp_i - trg_i|$$

نتایج

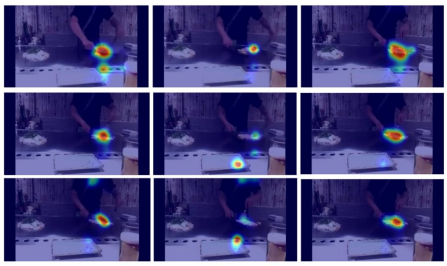
ساختار نهایی شبکه در هر بلوک به شکل زیر است:

لایه	ابعاد ورودی	ابعادی خروجی
SepConv3d	3×64×7×7	64×64×112×112
MaxPool3d	64×64×112×112	64×64×56×56
BasicConv3d	64×64×56×56	64×64×56×56
SepConv3d	64×192×56×56	192×64×56×56
MaxPool3d	192×64×56×56	192×64×28×28

برای آموزش شبکه از بهینه‌ساز SGD استفاده شده است. نرخ یادگیری در ابتدا ۰/۱ بوده و از مومنتوم ۰/۹ استفاده شده است. هم‌چنین از نرخ کاهش وزن 2×10^{-7} استفاده شده است. هم‌چنین پارامتر pile نیز به عنوان دسته‌ی داده ورودی برای آموزش تعریف شده است که برابر با ۵ است. در طی آموزش اندازه batch برابر با ۸ و پارامتر len temporal نیز به مفهوم تعداد فریم‌های ورودی برای هر آموزش برابر با ۳۲ در نظر گرفته شده است. در گزارش مربوط به نتایج، نمودار تابع هزینه و دقت به عنوان فاصله فروبنویوس دو فریم در تکرارهای برنامه در نظر گرفته شده است.



شکل ۱: نمودار دقت و تابع هزینه آموزش و ارزیابی



شکل ۲: نمونه‌های خروجی از فریم‌های یک ویدیو در مرحله تست

جمع بندی

در این پروژه، مدلی برای تشخیص حرکات چشم در ویدئو طراحی و پیاده‌سازی شد. از شبکه‌های عمیق کانولوشنی برای آموزش مدل استفاده شده و داده‌های آموزش از مجموعه داده DHF1K استخراج شده است. نتایج نشان می‌دهد که مدل پیشنهادی با دقت بالا می‌تواند الگوهای حرکت چشم را تشخیص داده و پیش‌بینی کند. در پایان، نقاط تمرکز بصری استخراج شده و ویدئوهای حاوی این اطلاعات تولید می‌شوند.

مراجع اصلی

1. **Wang, W., Shen, J., Guo, F., Cheng, M. M., & Borji, A. (2018). Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on computer vision and pattern recognition* (pp. 4894-4903).**
2. **Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., & Gao, S. (2018). Gaze prediction in dynamic 360 immersive videos. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5333-5342).**
3. **Akinyelu, A. A., & Blignaut, P. (2020). Convolutional neural network-based methods for eye gaze estimation: A survey. IEEE Access, 8, 142581-142605.**