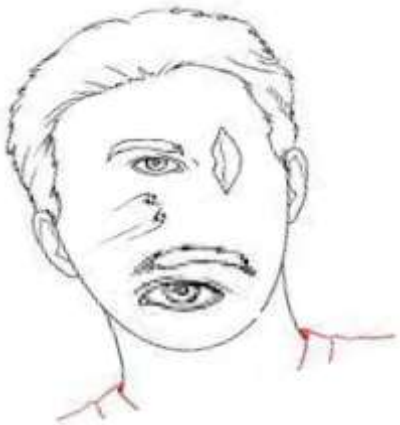
نام مدرس: دکتر محمدی **دستیاران آموزشی مرتبط:** احمدی - فرهمند **مهلت تحویل:** ۱۳ /۰۳/۱۴۰۲	**تمرین سری ششم** **درس یادگیری عمیق**	

۱- الف) عملکرد شبکه‌های کانولوشنی و شبکه‌ها مبتنی بر توجه را برای استخراج ویژگی در تصاویر، با یکدیگر مقایسه کنید و توضیح دهید هرکدام از آنها به استخراج چه دسته‌ای از ویژگی‌ها می‌پردازند.

ب) به نظر شما به‌ازای ورودی زیر، هر کدام از شبکه‌های مبتنی بر کانولوشن و مبتنی بر توجه جهت مسئله دسته‌بندی چهره انسان بودن یا نبودن، چگونه عمل می‌کنند:



۲- یک توجه چند سر additive با ۳ سر را در نظر بگیرید. ابعاد key ، query و value را به ترتیب ۱۰، ۲۰، ۳۰ در نظر بگیرید. فرض کنید هر کدام از سرها به ابعاد ۱۰۰ تبدیل شوند. همچنین در نظر داشته باشید که خروجی نهایی ۵۰ می‌باشد. با فرض اینکه دنباله ورودی ۶۴ تایی باشد ، تعداد پارامترها را مشخص کنید.

۳- یک معماری Vision Transformer (VIT) را در نظر بگیرید که به عنوان ورودی، یک تصویر ۳ بعدی با ابعاد $128 \times 128 \times 128$ و تعداد ۴ کانال را دریافت می‌کند. در صورتی که از patch sizeیی به ابعاد $16 \times 16 \times 16$ و لایه مخفی به ابعاد ۷۶۸ استفاده شده باشد، ابعاد positional embedding را محاسبه کنید.

۴- مقاله Swin Transformer: Hierarchical Vision Transformer using Shifted Windows را مطالعه کنید و به سوالات زیر پاسخ دهید:

راهنمایی: می‌توانید از لینک استفاده کنید.

الف) این معماری به دنباله حل چه چالشی می‌باشد.

ب) تفاوت بلوک‌های MSA و W-MSA در چیست؟ همچنین توضیح دهید که دلیل ساخت این نوع بلوک جدید چه بوده است.

ج) توضیح دهید که چرا از پنجره‌های shift یافته در این معماری استفاده شده است.

۵- موارد خواسته شده را در نوت‌بوک پیوست شده انجام دهید.

لطفا سند قوانین انجام و تحویل تمرین های درس را مطالعه و موارد خواسته شده را رعایت فرمایید

موفق و سلامت باشید