۱۱) توضیح سیستم.

در این شبکه به نظر می رسد که یک سیستم video captioning است. و به این صورت عمل می کند که ابتدا یک لایه 3D conv روی تصاویر می زند و ویژگی ها و تناب با وزن های خودش را استخراج می کند و این عملیات را برای هر فریم پشت سر هم انجام می دهد. پس خروجی لایه 3D conv را به یک CNN می دهد و در نهایت به کمک لایه mean pooling این ویژگی ها را ترکیب می کند. در واقع در این جا به کمک 3D conv ویژگی هایی که به طول زمان ارتباط دارند را استخراج می کند و پس بردار ویژگی استخراج شده را به عنوان ورودی CNN می دهد و در نهایت چون فریم های پشت سر هم اطلاعات شبیه به هم و قابل ترکیب دارند که به کمک یک لایه mean pooling سیاهیک اطلاعات را استخراج می کند.