University of Mosul
College of Engineering

Video Prediction using 3D Deep Convolutional Neural Networks

Zahraa Talal Abed Ali

Ph.D. Thesis in Computer Engineering/ Computer Engineering

Supervised by

Prof. Dr. Shefa Abdul Rahman Dawwd

2024 A.D. 1445 A.H.

Abstract

Video prediction (VP) is a part of the important field of computer vision and video processing. Video prediction models extract the content of the future frame depending on the current and the past frames. However, these models must extract the spatial and temporal features in parallel. This is considered a big challenge in many researches. Extracting spatial and temporal feature operations is a very complicated process and needs a huge amount of data and large and complex operations, which are considered big challenges in any video processing. Many types of research apply two-dimensional convolutional, RNN, and CONVLSTM layers to build the VP models. However, these models suffer from extracting the temporal features efficiently and the classical VP models fail to extract the long-term temporal features. In this work, the basic architecture of deep learning models such as autoencoders (AE), variational autoencoders (VAE), and generative adversarial networks (GAN) is integrated with the 3D-CONV layers to design efficient video prediction models. The AE architecture integrates with 3D-CONV, 3D spatial max pooling, and temporal max pooling layers to design the proposed 3D-AE and 3D-EAE models. The third model was proposed to achieve a more stable VP model. The suggested temporal sampling function helped to increase the performance of the 3D-VAE model. The fourth model (3D-VGAN) was introduced to decrease the blurry prediction as much as possible. This model achieved better performance compared with other approaches. The results of the proposed 3D-VGAN were 39.96 for PSNR, 2. 616 * 10^{-4} for MSE, and 0.973 for SSIM based on the Cityscapes dataset. One of the big

challenges is a large hardware resources are utilized to implement all the proposed models. The absolute least weight value of the pruning technique is proposed to decrease the resources and maintain the performance as best as possible. The pruning technique truncated 29.7% of the total FLOP of the proposed 3D-VGAN model.

جامعة الموصل
كلية الهندسة

التنبؤ الفيديوي باستخدام الشبكات العصبية اللافوفية العميقة الثلاثية الابعاد

أطروحة دكتواره تقدمت به الطالبة

زهراء طلال عبد علي

الى

مجلس كلية الهندسة في جامعة الموصل كجزء من متطلبات نيل شهادة الدكتوراه في

هندسة الحاسوب

بإشراف

الأستاذ الدكتور شفاء عبد الرحمن داؤد

1445 هـ
2024 م

المستخلص

يعد التنبؤ بالفيديو (VP) جزءًا من المجالات المهمة في تطبيقات الرؤية الحاسوبية والمعالجات الفيديوية. حيث تقوم نماذج التنبؤ بالفيديو باستخراج محتوى الإطار المستقبلي اعتمادًا على الإطارات الحالية والسابقة. ومع ذلك، يجب على هذه النماذج استخراج السمات المكانية والزمانية بالتوازي. ويعتبر هذا تحديا كبيرا في العديد من الأبحاث. حيث تعتبر عمليات إستخراج الخصائص المكانية والزمانية هي عمليات معقدة للغاية وتحتاج إلى كمية هائلة من البيانات وعمليات رياضية معقدة، والتي تعتبر تحدياً كبيرإ في أي معالجة فيديوية. تعتمد العديد من الأبحاث على الطبقات تلافيفية ثنائية الأبعاد وRNN وCONVLSTM لبناء نماذج التنبوءات الفيديوية , ومع ذلك، تعاني هذه النماذج من صعوبة إستخلاص السمات الزمنية بكفاءة، وتفشل نماذج VP الكلاسيكية في استخلاص السمات الزمنية طويلة المدى. في هذا العمل، تم دمج البنية الأساسية لنماذج التعلم العميق مثل أجهزة التشفير التلقائي (AE)، وأجهزة التشفير التلقائي المتغيرة (VAE)، وشبكات الخصومة التوليدية (GAN)مع الطبقات اللافوفية الثلاثية الابعاد لتصميم نماذج فعالة للتنبؤ بالفيديو. حيث تتكامل بنية AE مع الطبقات اللافوفية الثلاثية الابعاد ، لتصميم نماذج 3D–AE و3D–EAE المقترحة. وكذلك تم إقتراح النموذج الثالث لتحقيق نموذج VP أكثر استقرارًا. ساعدت وظيفة أخذ العينات الزمنية المقترحة على زيادة أداء نموذج 3D–VAE . و كذلك تم تقديم النموذج الرابع (3D–VGAN)لتقليل التنبؤ الضبابي قدر الإمكان. حقق النموذج الاخير أداء أفضل مقارنة بالطرق الأخرى. كانت

نتائج 3D-VGAN المقترحة 39.96 لـ PSNR و0.973 لـ SSIM بناءً على مجموعة بيانات Cityscapes . أن أحد أهم التحديات الكبيرة التي واجهت هذا العمل هو اســـتخدام موارد الأجهزة الكبيرة لتنفيذ جميع النماذج المقترحة.ولهذا تم اقتراح القيمة الأقل وزنًا على الإطلاق لتقنية التقليم لتقليل الموارد والحفاظ على الأداء بأفضــل شـكل ممكن. حيث اقتطعت تقنية التقليم 29.7% من إجمالي FLOP لنموذج 3D-VGAN المقترح.