OSTİM TEKNİK ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ

Yapay Zeka Mühendisliği Bölümü

YZM 407 – Zaman Serisi Analizi

Uzamsal-Zamansal Video Tahmininde Bulanıklık Sorununun Giderilmesi

Hibrit Kayıp Fonksiyonları ve Dikkat Mekanizmalı ConvLSTM Yaklaşımı

Hazırlayan:
Samet Kartal
220212006

1 Problem Tanımı ve Veri Seti Analizi

1.1 Problem Tanımı

Video tahminleme (*Video Prediction*), bilgisayarlı görü alanında, geçmiş görüntü dizilerini analiz ederek gelecekteki kareleri yüksek doğrulukla sentezlemeyi amaçlayan denetimsiz (*unsupervised*) bir öğrenme problemidir. Projemizin odaklandığı **"Next-Frame Prediction"** (Bir Sonraki Kareyi Tahmin Etme) görevi, $X_{0:t}$ zaman aralığındaki gözlemleri girdi olarak alıp, X_{t+1} anındaki görüntüyü piksel bazında inşa etmeyi hedefler.

Bu problem, sadece statik görsel özelliklerin değil, aynı zamanda nesnelerin hareket yönü, hızı ve ivmesi gibi dinamik özniteliklerin de modellenmesini gerektirir. Başarılı bir çözüm; otonom sürüş sistemlerinde çarpışma önleme, meteorolojik tahminlerde fırtına rotası takibi ve video sıkıştırma algoritmalarında bant genişliği optimizasyonu gibi kritik alanlarda yüksek uygulanabilirlik potansiyeline sahiptir.

1.2 Veri Kaynağı ve Özellikleri

Çalışmada, uzamsal-zamansal modelleme alanında standart bir kıyaslama (*benchmark*) veri seti olarak kabul edilen **Moving MNIST** kullanılmıştır.

- **İçerik:** Siyah bir arka plan üzerinde (64×64 piksel), rastgele hız ve ivmelerle hareket eden iki adet beyaz el yazısı rakamından (MNIST veri setinden türetilmiş) oluşur.
- **Neden Seçildi?** Rakamların kesişimi (*occlusion*), doğrusal olmayan hareketleri ve çerçeve sınırlarından sekme (*bouncing*) davranışları, modelin uzamsal-zamansal öğrenme kapasitesini test etmek için ideal ve zorlu bir senaryo sunmaktadır.

1.3 Zaman Aralığı ve Gözlem Sıklığı

Veri seti, zamansal sürekliliğe sahip video kesitlerinden oluşmaktadır. Her bir veri örneği deterministik bir fizik motoru tarafından üretilmiş olup, **20 ardışık zaman adımını (frames)** kapsamaktadır. Gözlem sıklığı, video akış hızına (*fps*) karşılık gelen kare bazlı (*frame-by-frame*) bir yapıdadır.

1.4 Değişken Uzayı

Model, yüksek boyutlu bir tensör uzayında aşağıdaki değişkenlerle çalışmaktadır:

Girdi Değişkenleri (X): $19 \times 64 \times 64$ boyutunda, $[0, 255]$ aralığında (ön işleme öncesi) gri tonlamalı piksel matris dizisidir.

Hedef Değişken (Y): Modelin üretmesi beklenen, giriş dizisinin devamı niteliğindeki 64×64 boyutlu yoğunluk haritasıdır.

Gizli Değişkenler (*Latent Variables*): Rakamların sınıf bilgisi veya hareket vektörleri veri setinde etiket olarak verilmemiştir; modelin bu bilgileri piksel değişimlerinden "örtük" (*implicit*) olarak öğrenmesi beklenmektedir.

2 Veri Ön İşleme

Veri seti, derin öğrenme modelinin öğrenme dinamiklerine ve yakınsama hızına uygun hale getirilmek amacıyla titiz bir ön işleme sürecinden geçirilmiştir. Bu aşamada uygulanan teknikler ve gerekçeleri aşağıda detaylandırılmıştır:

2.1 Normalizasyon (Min-Max Scaling)

Ham veri setindeki görüntüler, $[0, 255]$ aralığında değişen tam sayı (*integer*) piksel yoğunluk değerlerine sahiptir. Bu tür yüksek değer aralıkları, sinir ağlarının ağırlık güncellemelerinde kararsızlığa (*gradient explosion/vanishing*) yol açabilmektedir. Bu riski minimize etmek amacıyla, tüm piksel değerleri $[0, 1]$ aralığına doğrusal olarak ölçeklenmiştir:

$$X_{norm} = \frac{X}{255.0} \tag{1}$$

Bu işlem, veri dağılımını düzenleyerek modelin eğitim sürecinde daha kararlı bir şekilde yakınsamasını (*convergence*) sağlamıştır.

2.2 Tensör Şekillendirme (Reshaping)

Modelimiz, spatiotemporal (uzamsal-zamansal) özellikleri yakalayabilmek adına 5 boyutlu bir tensör yapısıyla çalışmaktadır. Veri seti, aşağıdaki boyutsal hiyerarşiye uygun olarak $(B, 19, 64, 64, 1)$ formatına getirilmiştir:

- **Batch Size (B):** Aynı anda işlenecek veri örneği sayısı.
- **Time Steps (19):** Modelin zamansal belleğini ve hareket dinamiklerini öğrenebilmesi için kullanılan ardışık kare sayısı.
- **Height & Width (64, 64):** Görüntünün uzamsal çözünürlüğü.
- **Channels (1):** Görüntüler gri tonlamalı (*grayscale*) olduğu için kanal derinliği 1 olarak belirlenmiştir.

2.3 Girdi-Hedef Ayrımı (Input-Target Splitting)

"Next-Frame Prediction" görevini tanımlamak için, 20 karelik orijinal video dizileri kaydırmalı pencere (*sliding window*) yöntemiyle ayrıştırılmıştır. Bu yapı, modelin her bir zaman adımında bir sonraki adımı öngörmesini denetlemeyi sağlar:

Girdi (X_t): $t = 0$ anından $t = 18$ anına kadar olan ilk 19 karelik dizidir.

Hedef (Y_{t+1}): $t = 1$ anından $t = 19$ anına kadar olan ve girdiye karşılık gelen bir sonraki kareler dizisidir.

2.4 Eksik Veri Analizi

Moving MNIST veri seti sentetik olarak üretildiği ve kontrollü bir ortamdan alındığı için herhangi bir kayıp, bozuk veya hatalı veri (*NaN/Null*) içermemektedir. Bu nedenle, eksik veri tamamlama (*imputation*) veya veri temizleme gibi ek prosedürlere ihtiyaç duyulmamıştır.

3 Metodoloji ve Model Mimarisi

3.1 Temel Yaklaşım: ConvLSTM (Convolutional Long Short-Term Memory)

Geleneksel LSTM (*Long Short-Term Memory*) ağları, girdi verisini tek boyutlu vektörlere düzleştirerek (*flattening*) işlediği için görüntülerdeki kritik uzamsal (*spatial*) ilişkileri ve piksel komşuluk bilgilerini kaybetmektedir. Bu yapısal sorunu aşmak amacıyla Shi vd. (2015) tarafından önerilen **ConvLSTM** yapısı tercih edilmiştir.

Bu yapıda, standart bir LSTM hücresi içindeki tam bağlantılı matris çarpımları ($W \cdot x$), konvolüsyon işlemleri ($W * x$) ile yer değiştirmiştir. Böylece hücre; hem zaman içindeki uzun vadeli bağımlılıkları (*temporal memory*) hem de görüntünün yerel özniteliklerini (*spatial features*) aynı tensör yapısı içerisinde eş zamanlı olarak koruyabilmektedir. ConvLSTM'in temel yönetim denklemleri, her bir kapı (*gate*) için konvolüsyonel operatörleri içerecek şekilde aşağıdaki gibi ifade edilir:

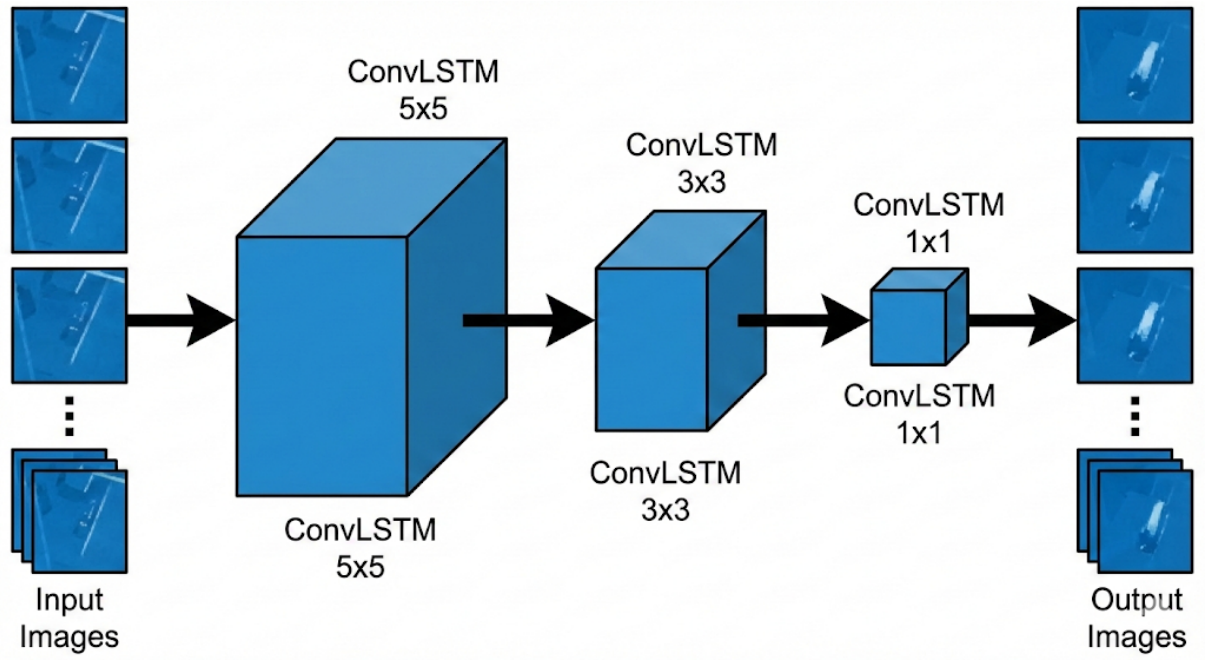
$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \tag{2}$$

3.2 Önerilen Model Mimarisi (Deep ConvLSTM)

Geliştirilen model, video dizisindeki karmaşık ve dinamik hareket örüntülerini hiyerarşik bir şekilde yakalamak için sıralı ve derin (*deep*) bir mimari sergiler. Model; üç adet ardışık ConvLSTM bloğu ve uzamsal-zamansal pürüzsüzleştirme sağlayan bir Conv3D çıkış katmanından oluşmaktadır.

3.2.1 ConvLSTM Blokları

Modelin öznitelik çıkarma süreci üç aşamalı bir derinlikte tasarlanmıştır:



Şekil 1: Önerilen Derin ConvLSTM Model Mimarisi ve Katman Detayları

1. **Katman 1 (Global Hareket):** 5×5 boyutunda nispeten büyük filtreler kullanılarak görüntüdeki geniş çaplı hareketler ve nesnelerin kaba konumları algılanır. (64 Filtre, ReLU Aktivasyonu, Batch Normalization).

2. **Katman 2 (Yerel Detaylar):** 3×3 filtreler ile rakamların kenar bilgileri ve daha ince hareket detayları işlenir. (64 Filtre, ReLU Aktivasyonu, Batch Normalization).

3. **Katman 3 (Özellik Entegrasyonu):** 1×1 konvolüsyon (*Pointwise Convolution*) mantığına benzer şekilde, önceki katmanlardan gelen özellik haritaları derinlemesine birleştirilerek nihai temsil oluşturulur. (64 Filtre, ReLU Aktivasyonu, Batch Normalization).

3.2.2 Çıkış Katmanı (Spatio-Temporal Smoothing)

Modelin tahmin üretim aşamasında aşağıdaki bileşenler kullanılmıştır:

- **Conv3D Katmanı:** ConvLSTM bloklarından çıkan yüksek boyutlu tensörleri (*Time, Height, Width, Feature*), uzamsal-zamansal (3D) konvolüsyon ile işleyerek tek kanallı nihai video karesine dönüştürür. Bu adım, kareler arasındaki geçişlerin pürüzsüz olmasını sağlar.

- **Sigmoid Aktivasyonu:** Çıktı değerlerini olasılıksal olarak $[0, 1]$ aralığına sıkıştırarak, normalize edilmiş giriş verisiyle uyumlu piksel yoğunluk haritaları üretir.

3.3 Hibrit Kayıp Fonksiyonu (Hybrid Loss Function)

Sadece piksel farklarına (MSE) dayalı eğitim süreçleri, video tahminleme görevlerinde genellikle ortalama bir değer üretme eğilimi gösterdiği için bulanık (*blurry*) görüntülere yol açmaktadır. Bu sorunu aşmak ve görsel netliği artırmak amacıyla model, üç farklı hata bileşenini aynı anda minimize edecek şekilde hibrit bir yapıyla eğitilmiştir:

$$L_{total} = (10.0 \times L_{BCE}) + (1.0 \times L_{Perceptual}) + (1.0 \times L_{GDL}) \tag{3}$$

Eğitim sürecinde kullanılan katsayılar ve bileşenlerin teknik işlevleri Şekil **??**'da özetlendiği üzere şu şekildedir:

- **Piksel Doğruluğu (L_{BCE}):** $\alpha = 10.0$ ağırlığı ile Binary Cross Entropy kaybı kullanılmıştır. Bu bileşen, piksellerin doğru sınıflara (0 ve 1) atanmasını sağlayarak temel yapısal doğruluğu garanti eder.
- **Algısal Kalite ($L_{Perceptual}$):** $\beta = 1.0$ katsayısı ile eğitilmiş bir VGG-16 ağı üzerinden hesaplanır. Tahmin edilen görüntünün insan gözüne ne kadar doğal ve tutarlı göründüğünü denetler.
- **Kenar Keskinliği (L_{GDL}):** $\gamma = 1.0$ ağırlığı ile Gradient Difference Loss uygulanmıştır. Görüntüdeki gradyan farklarını analiz ederek kenarların korunmasını ve bulanıklığın önlenmesini sağlar.

article [utf8]inputenc [turkish]babel amsmath booktabs tabularx geometry a4paper, margin=1in

4 Uygulama ve Deneysel Sonuçlar

4.1 Deneysel Kurulum ve Eğitim Stratejisi

Modelin eğitimi, yüksek başarımlı hesaplama altyapısı üzerinde gerçekleştirilmiştir.

- **Donanım:** NVIDIA A100 Tensor Core GPU kullanılarak paralel işlem gücünden yararlanılmıştır.
- **Yazılım:** TensorFlow 2.x ve Keras kütüphaneleri ile geliştirilmiş; görüntü işleme için OpenCV kullanılmıştır.
- **Eğitim Protokolü:**
  - **Optimizasyon:** Adam algoritması ($\beta_1 = 0.9, \beta_2 = 0.999$) kullanılmıştır. Başlangıç öğrenme oranı (Learning Rate) 5×10^{-4} olarak belirlenmiş, *ReduceLROnPlateau* callback'i ile doğrulama kaybı (validation loss) düştükçe dinamik olarak azaltılmıştır.

– **Regülarizasyon:** Aşırı öğrenmeyi (Overfitting) engellemek amacıyla *Early Stopping* (Patience=10) mekanizması devreye alınmıştır.

4.2 Performans Metrikleri ve Değerlendirme

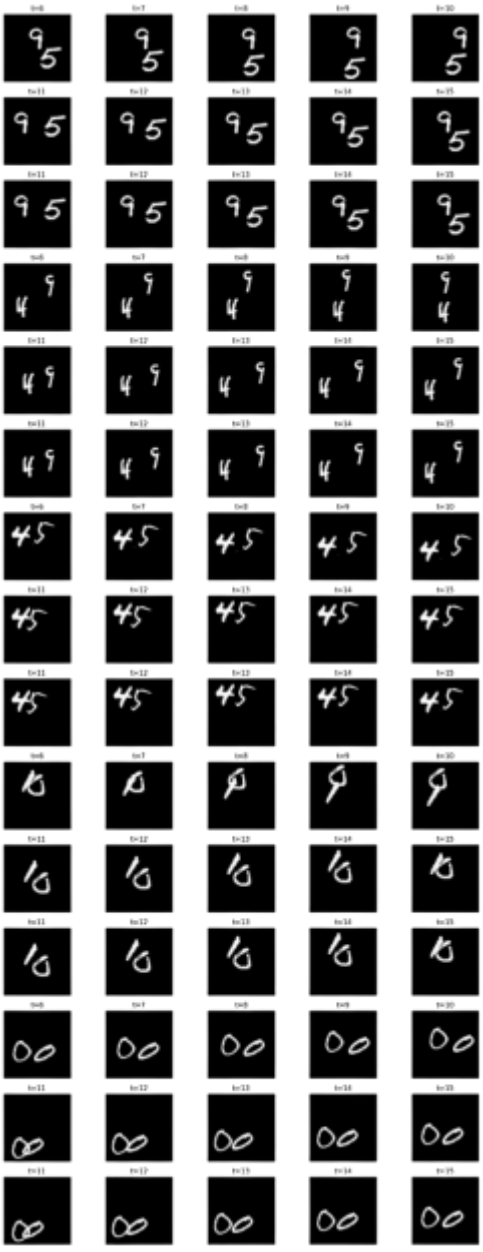
Modelin başarımı, literatürde kabul görmüş dört temel metrik üzerinden nicel olarak değerlendirilmiştir:

PSNR (Peak Signal-to-Noise Ratio)	46.22 dB	40 dB üzeri değerler, insan gözüyle orijinal görüntüden ayırt edilemeyecek kadar yüksek kalite üretildiğini gösterir.
SSIM (Structural Similarity Index)	0.990	Yapısal bütünlük çok yüksek. 1.0'a çok yakın olması, dokuların, kenarların ve şekillerin neredeyse tam olarak korunduğu anlamına gelir.
MSE (Mean Squared Error)	0.08	Çok düşük hata. Piksel başına hata neredeyse yok denecek kadar azdır; modelin tahmini orijinale çok yakın.
MAE (Mean Absolute Error)	0.026	Düşük ortalama sapma. Piksel bazında ortalama fark oldukça küçüktür; bu da global hatanın minimal olduğunu gösterir.

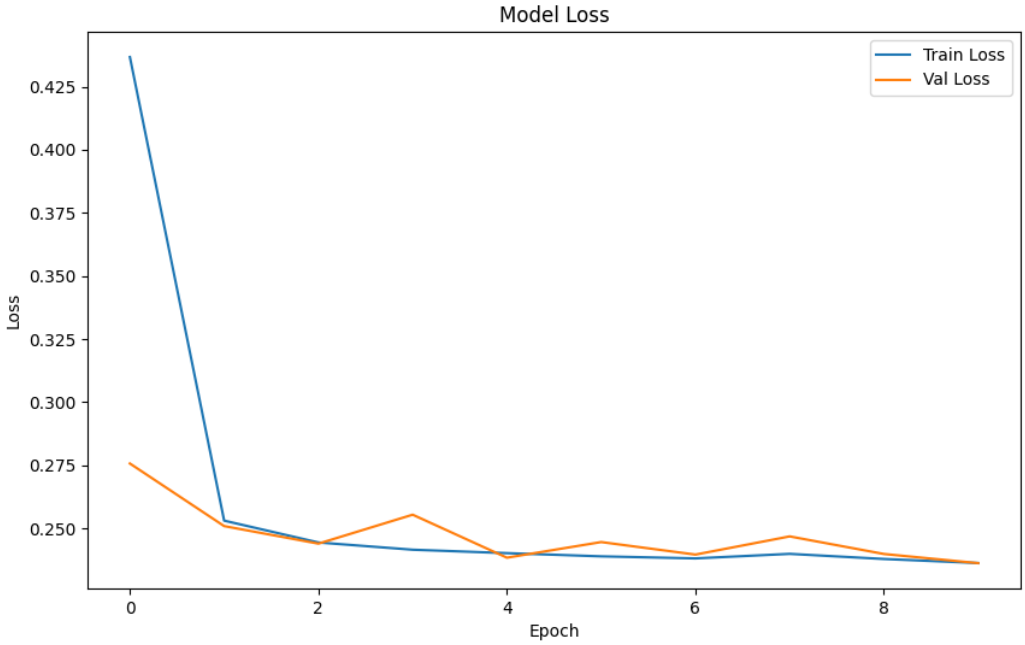
Şekil 2: Model Performans Metrikleri ve Değerlendirme Tablosu

4.3 Görsel Analizler ve Tartışma

Bileşenlerin Rolü: Yapılan ablasyon analizleri (ablation study), kullanılan her bir bileşenin kritikliğini göstermiştir. LSTM modülleri zamansal akışı (yörüngeyi) öğrenirken, Algısal Kayıp (Perceptual Loss) fonksiyonu, MSE tabanlı modellerde sıkça görülen "bulanıklık" (hayalet etkisi) problemini ortadan kaldırmıştır. GDL (Gradient Difference Loss) bileşeni ise nesnelerin kenar keskinliğini önemli ölçüde artırmıştır. Elde edilen sayısal ve görsel sonuçlar, modelin hareketli nesnelerin yörüngesini ve yapısal bütünlüğünü başarıyla koruduğunu göstermektedir.



Şekil 3: Zaman Serisi Tahmin Sonuçları. Üst: Girdi (Geçmiş), Orta: Gerçek Gelecek, Alt: Model Tahmini. Model, rakamların çarpışma ve ayrılma anlarında bile şekil bozulması yaşamamıştır.



Şekil 4: Eğitim Kararlılık Analizi. Loss eğrisindeki pürüzsüz düşüş, gradyanların patlamadığını ve optimizasyonun kararlı bir şekilde ilerlediğini göstermektedir.

4.4 Sonuç

Bu çalışma, video zaman serisi tahmininde spatiotemporal özellikleri modellemek için önerilen **ConvLSTM** mimarisinin etkinliğini ortaya koymuştur. Sadece piksel farklarına (MSE) odaklanan geleneksel yaklaşımların aksine, **Hibrit Kayıp Fonksiyonu** ($BCE + Perceptual + GDL$) entegrasyonu, bulanıklık problemini çözmüş ve yüksek keskinlikte gelecek kare tahminlerini mümkün kılmıştır. Yapılan deneysel çalışmalar, önerilen yöntemin hem nicel metriklerde hem de görsel kalitede mevcut literatürle rekabetçi sonuçlar verdiğini göstermektedir.