06/04/2025

CRCV REU Week 2

Eric Li



Overview



- Introduction
- Assignment 1
- Project 4

About Me

- Born in Shanghai, China
- This is my third year in the United States.



Lived in Los Angeles for 2 years.

Lived in Shanghai for 20 years.

Lived in Berkeley for 1 year.

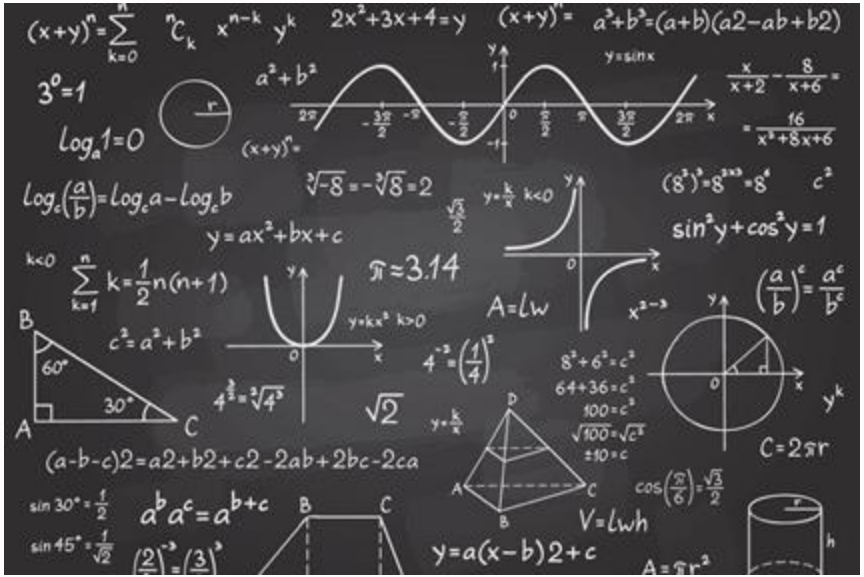
Academic:

- Used to major in Applied Mathematics at UC Berkeley, just switch to Data Science. (Was 4.0 before last Tuesday)
- Prior courses:

1. Calculus I&II
2. Linear Algebra & Differential Equations
3. Abstract Linear Algebra
4. Probability

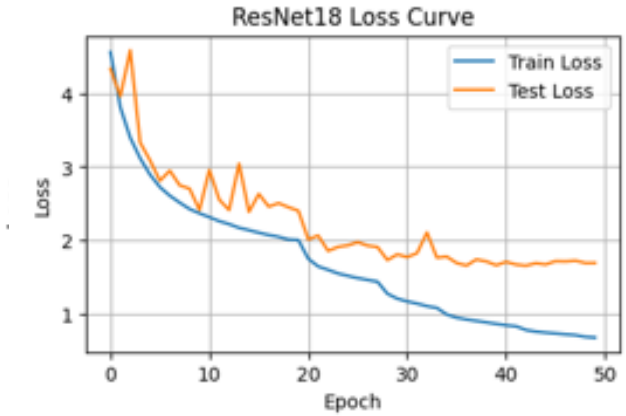
- Self-learning :

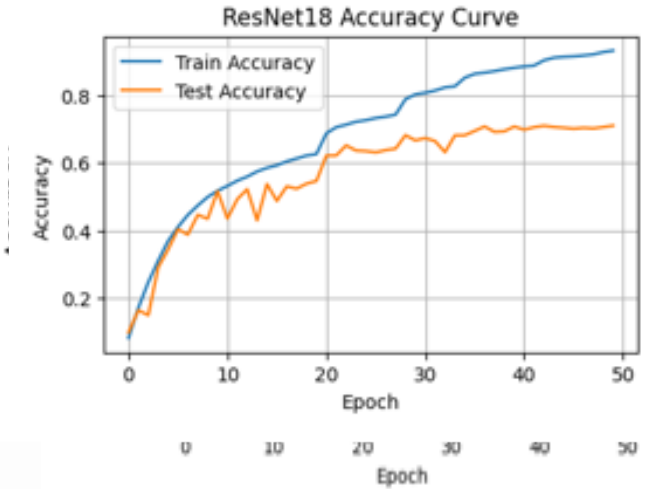
  <u>Computer Vision</u>
  – approximately 80% complete

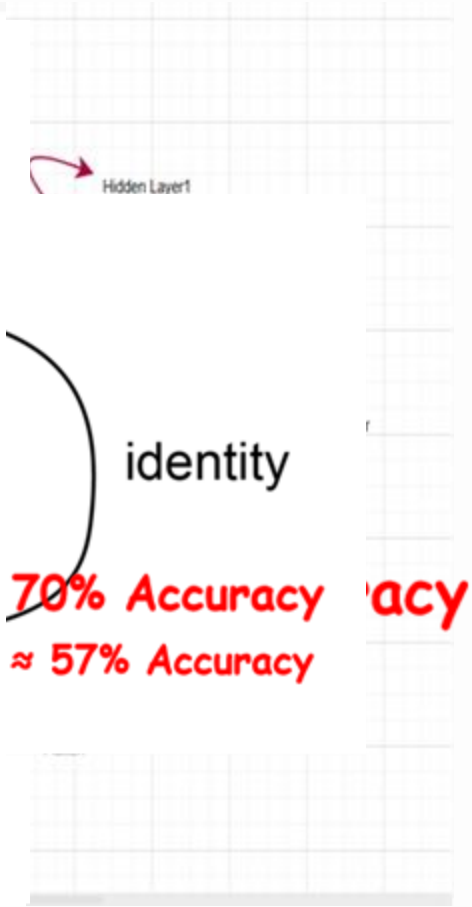


Assignment 1

- Dataset: CIFAR 100
- Methods:
  - Basic CNN, as Assignment 0
  - Traditional ResNet18
  - Modified ResNet18 with ImageDataGenerator







Project: Multi-modal Large Language Models for Video Anomaly Detection

Lead: Dr. Tian

- Past
- Present
- Future

arXiv:2404.01014v1 [cs.CV] 1 Apr 2024

Harnessing Large Language Models for Training-free Video Anomaly Detection

Luca Zanella[1] Willi Menapace[1] Massimiliano Mancini[1] Yiming Wang[2] Elisa Ricci[1,2]
University of Trento[1] Fondazione Bruno Kessler[2]
https://lucazanella.github.io/lavad/

Abstract

Video anomaly detection (VAD) aims to temporally locate abnormal events in a video. Existing works mostly rely on training deep models to learn the distribution of normality with either video-level supervision, one-class supervision, or in an unsupervised setting. Training-based methods are prone to be domain-specific, thus being costly for practical deployment as any domain change will involve data collection and model training. In this paper, we radically depart from previous efforts and propose ***LA****nguage-based* ***VAD*** *(LAVAD), a method tackling VAD in a novel,* training-free *paradigm, exploiting the capabilities of pre-trained large language models (LLMs) and existing vision-language models (VLMs). We leverage VLM-based captioning models to generate textual descriptions for each frame of any test video. With the textual scene description, we then devise a prompting mechanism to unlock the capability of LLMs in terms of temporal aggregation and anomaly score estimation, turning LLMs into an effective video anomaly detector. We further leverage modality-aligned VLMs and propose effective techniques based on cross-modal similarity for cleaning noisy captions and refining the LLM-based anomaly scores. We evaluate LAVAD on two large datasets featuring real-world surveillance scenarios (UCF-Crime and XD-Violence), showing that it outperforms both unsupervised and one-class methods without requiring any training or data collection.*



Figure 1. We introduce the first training-free method for video anomaly detection (VAD), diverging from state-of-the-art methods that are ALL training-based with different degrees of supervision. Our proposal, LAVAD, leverages modality-aligned vision-language models (VLMs) to query and enhance the anomaly scores generated by large language models (LLMs).

supervision of both normal and abnormal videos) [11, 13, 15, 24, 28, 35], one-class (*i.e.* only normal videos) [18, 20, 21, 25, 37, 38], and unsupervised (*i.e.* unlabeled videos) [30, 31, 40]. While more supervision leads to better results, the cost of manual annotation is prohibitive. On the other hand, unsupervised methods assume abnormal videos to constitute a certain portion of the training data, a fragile assumption in

Goal:

1. Identifying the limitations of existing frame-to-text prompting video data

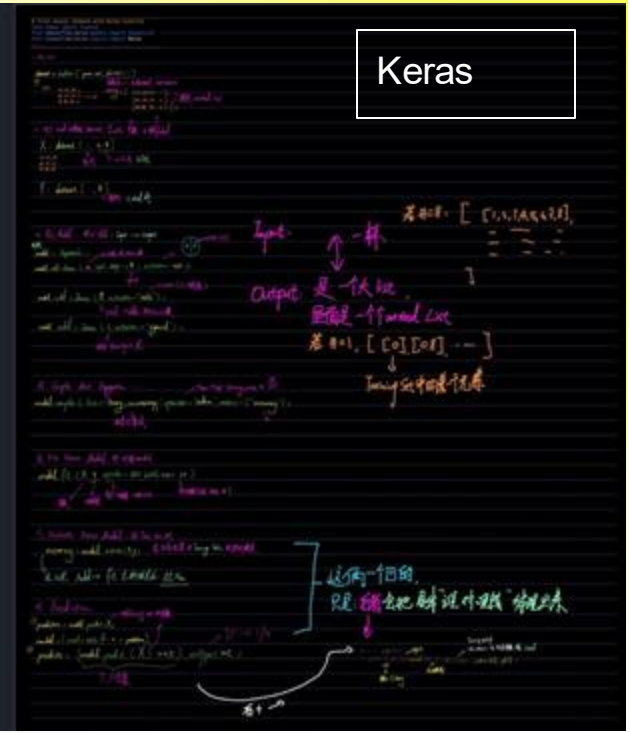
1. Extending MLLM or LLM by incorporating multi-modal embeddings alongside text to enrich scene representations.

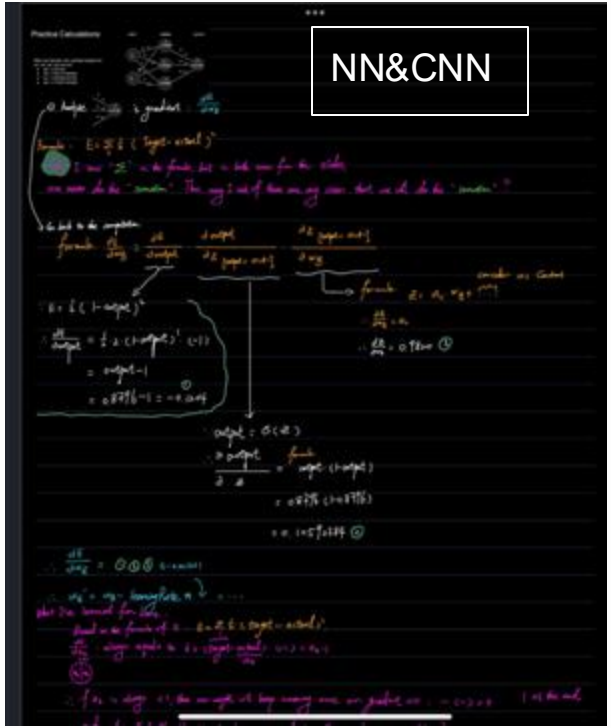
1. Benchmarking the resulting M³VAD framework against leading VAD methods on UCF-Crime and XD-Violence, aiming to match or exceed unsupervised and weakly-supervised performance without training.

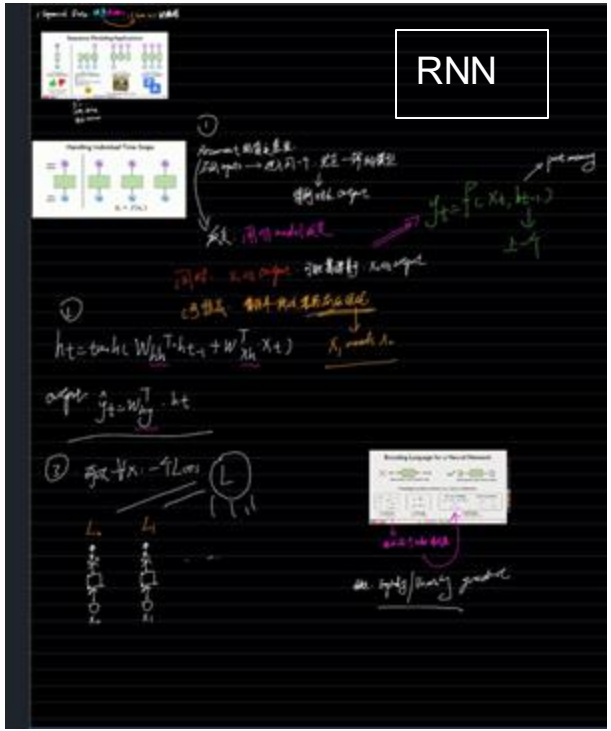
This week's Plan

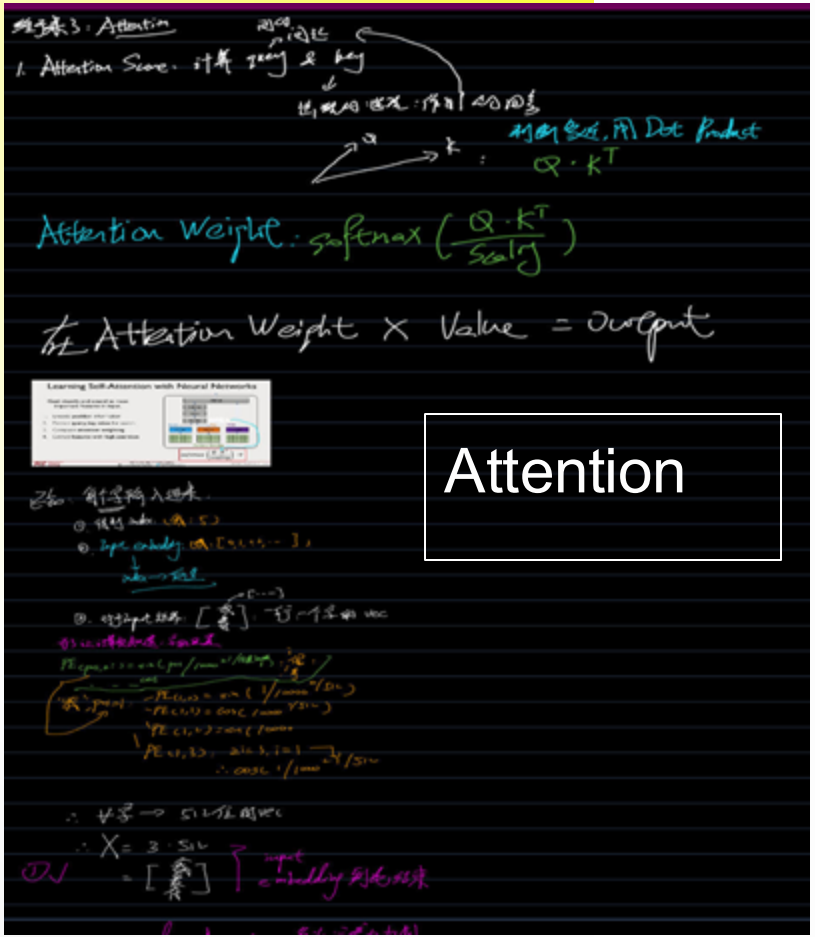
Today	Thursday	Friday
Suppose to finish all the Computer Vision materials. However, I spent more than 8 hours continuously trying to fully understand Transformers yesterday, but I failed to have a good command of it since I couldn't understand the process of Multi-head Attention. (Read the related slides from Stanford CS231n (no videos available and the lecturer prefer to teach it combined with RNN); MIT 6.S191: Only covered shallow understanding of attention can't help me understand the Transformer; Berkeley CS 182: Only formulas without any explanations. Without videos, it's really hard to understand) Topics remaining: 50% Transformers; Video Understanding (Video classification, 3D CNNs); Self-supervised Learning ; Vision and Language. (4 lectures left.) (Will be done by Thursday's evening)	**Daytime**: Finish studying all the remaining classes **Evening**: Start to read papers: 1. A Survey of Single-Scene Video Anomaly Detection; 2. Harnessing Large Language Models for Training-free Video Anomaly Detection	1. Reproducing the experiment by running the code from the second paper. 1. Meet with Dr.Tian to make weekend's plan and next week's plan.

Some study notes:

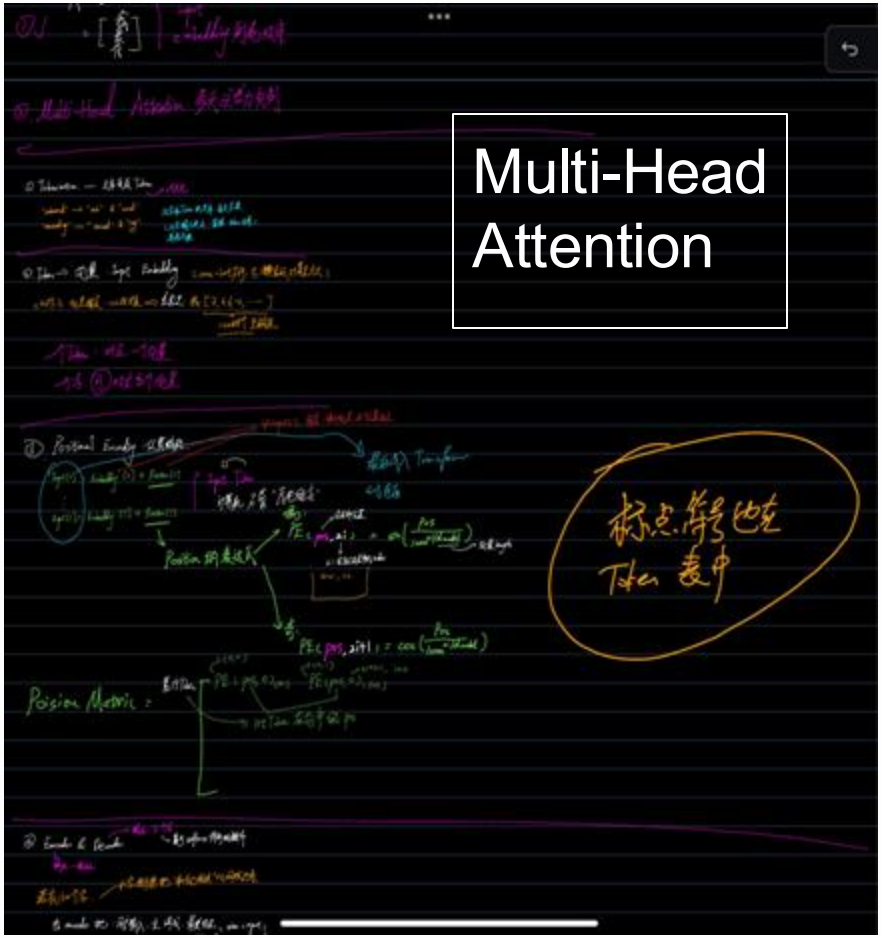







Attention Weight: softmax $\left(\frac{Q \cdot K^T}{\text{Scalg}}\right)$
在 Attention Weight × Value = output

Attention


Multi-Head Attention
标点符号也在 Token 表中

Thank you!

