LEARNING CROSS-MODAL CONTEXT FOR VIDEO CAPTIONING

Cross-modal context interaction

Video captioning, which aims to automalicaly natural language descriptions for video content, is a c challeming task due to complex video data and the inffiticatie is natural language.

This document introduces a novel Multi-level Cross-modal Context (MCC) [illegible] designed to address the limtations the existing video captioning approaches.

The MCC model chararicized by its effectiv capture integate muti-level cross-modal contexts between visual visual and textual modallites, leading to more accural divericatied video descriptions.

AUTHORS / AFFILIATIONS

Problem Statement and Limitations of Existing Methods

Separate/ Superfical Interaction

- Visual & Textual features separately.
- Single, shallow interaction level

Limited Context Modeling

- Lack & fine-grained object detail.
- Struggle with global scene understanding.

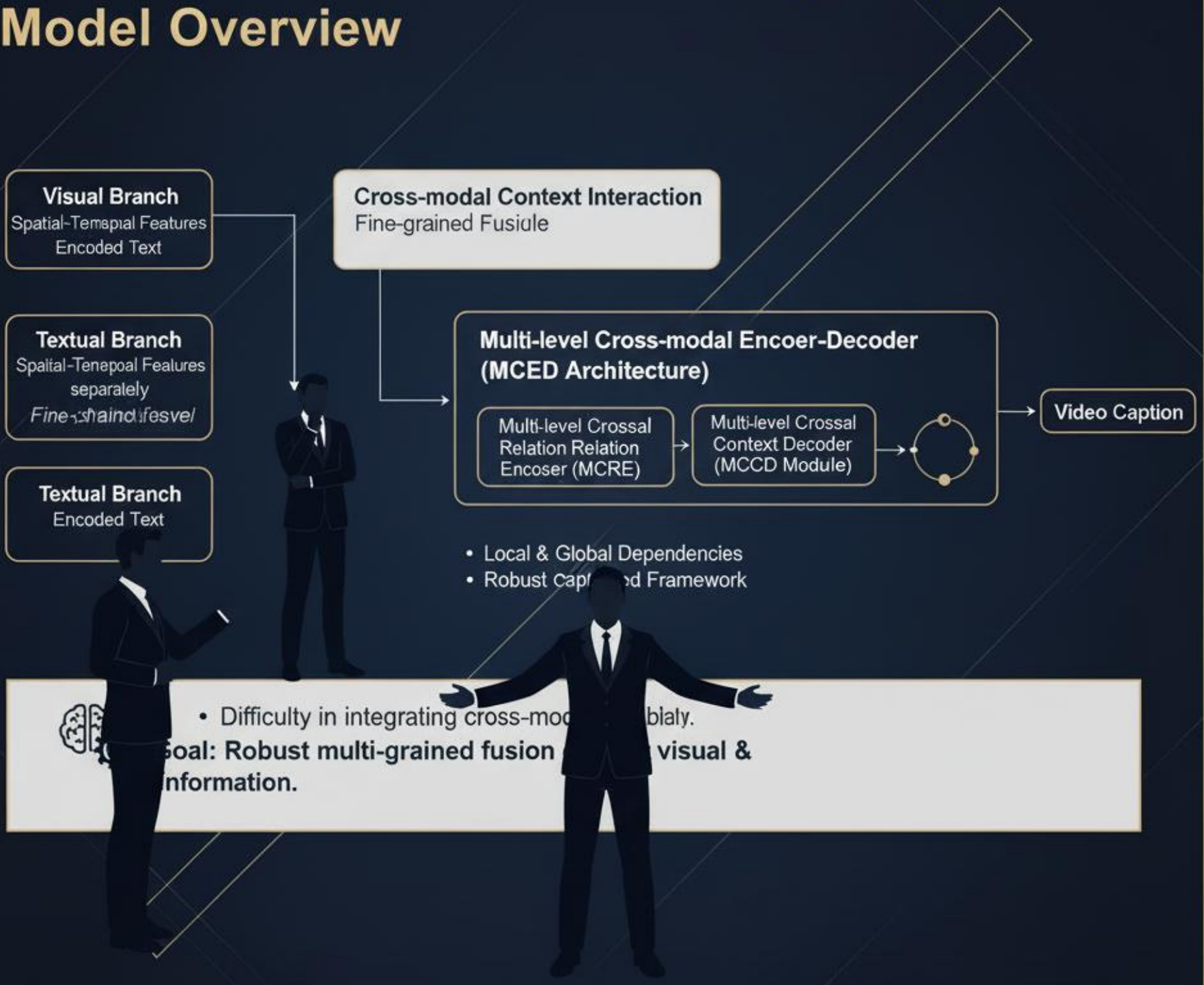
Generic & Repertive Captions

- Repettive pharses.
- Lacks diversity & naturalness.

- Difficulty in integrating cross-modal context globally & locally.

Goal: Robust multi-grained fusion of visual & visual & textual information.

Multi-level Cross-modal Context (MCC) Model Overview



Cross-modal Context Interaction (CCI) Module: Fine-graned Fusion

Module Overview

- Crucial for fine-grained fusion of visual & textual features. Tenglaich & TextusTexd & Textual ineraction

Process & Output

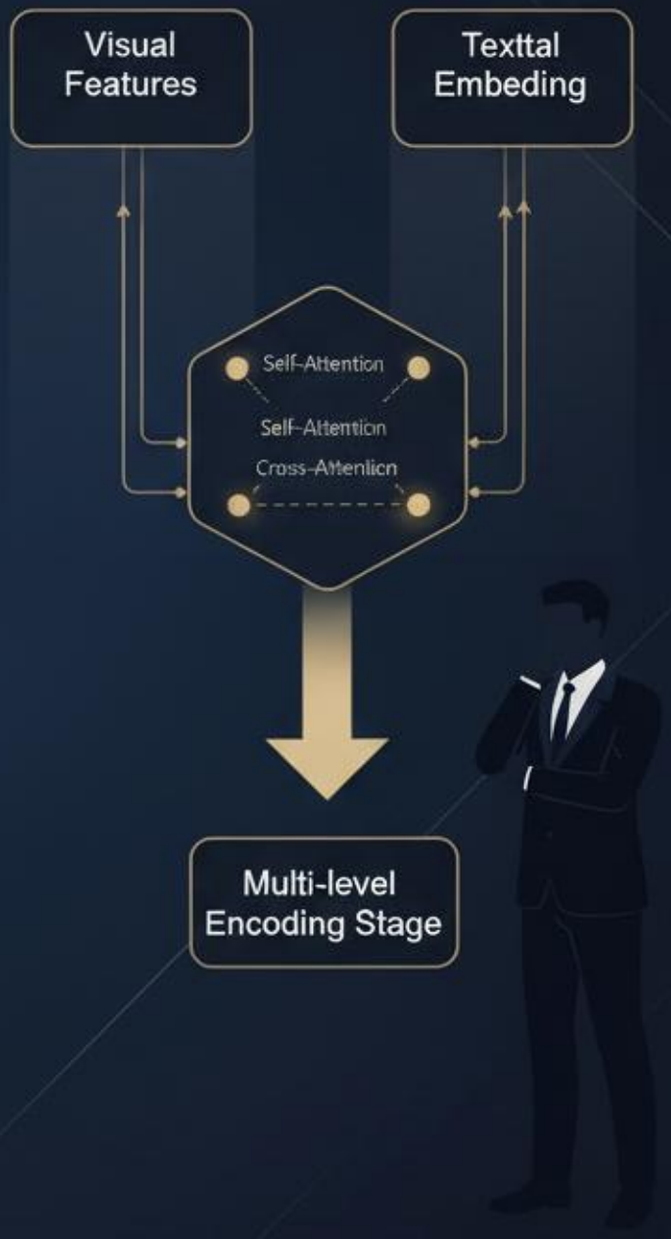
- Tisual features V(=ctual embeding E(T input. Uses Attention Mechanism

Series self-attention & cross-attention layers
Each modality informs & refines & scene othe other
Captures dynamic relationships at grandial

- Compules are queraldy for cross-mlext C, CCI grandiar level.

Mathematical Foundation

$$A = \frac{\text{Attention}(Q, K, V)}{QK^{\top}} = \frac{\text{softmazk}[}{\sqrt{d_k}V}$$



Multi-level Cross-modal Encorer-Decoder (MCED) Architecture

The Multi-level Cross-modal Relation ahitecture is responsible processing processing pentieation the fused mose-fusded cross-modal contexts and generatir.he vide capctus.This three main anaotin components: Multi-level Cross-modal Detoder (MCRE), and oherly an lched, and vextual components:

1. Multi-level Cross-modal Relation Encorer (MCRE

- Leverages interactive context from Builds rich, multi-grained video representations
- Uses multiple encorer layers for seamatic levels.
- Self-attention for intra-attection for for inter-modal
- Hierrochichal encoding for globcd-details.

2. Multi-level Cross-modal Context Decoder (MCCD)

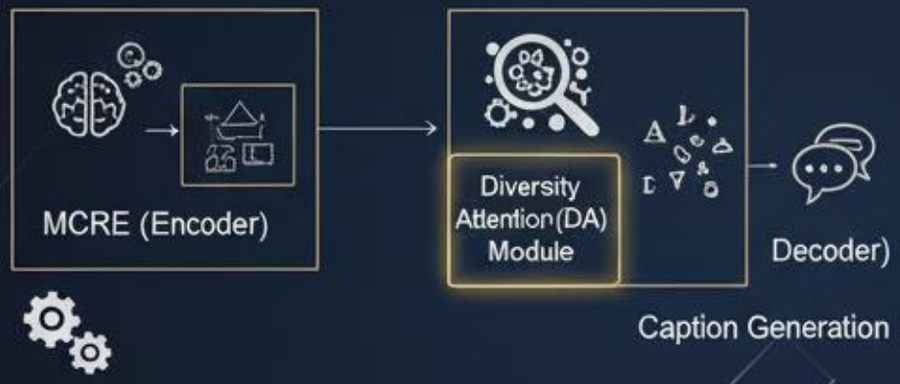
- Generates caption word-vordy-Conditined on Minted on MCRE representations.
- Dynamic attention selects relevrant cosavant info from different levels sel layers with
- masked self-rootad self-attention & cross-attection.

Multi-level Cross-modal Encorer-Decoder (MCED) Diversity Attention (Cont):

3. Diversity Attention: Fine-grained Fusion

3. Diversity Attention (DA) Module

MCED Architecture



- Dynamiacly adjusts attention weights.
- Encourage diversity in word choce.
- Prevents repeitive phrases.

- Modulates standard attention with diversity component (DA) word choce.
  (λ: diversity emıehsis hyperparatemter.

- Difficulty in integrating attentely
  Godust multi-grained visual fusion: $s' \; {}_{satn} + {}_{satn} = \lambda \, s^{attn} \; {}_{sdiv}\}$
  (λ: diversity emhehsis hyperparaimter.

Experimental Setup: Datasets and Evaluation Metrics

MSVD Datset

- 1,970 YouTube clips
  42 hours total
- 40 sentences/video (avg.).
  Split: 1200 Train, 100 Val, 670 Tes
- iverse content, high, high
  annotation varablity

MSR-VTT Datset

- 1,0000 video clips
- 20 English senterices/video
  Split: 6513 Train, 497 Val Test
- Wide categories: people, sports,
  Complex, longer
  descriptions

Evaluation Metrics

- CIDer (Semantic similarity)
- METEOR (N-gram, fluency,
  (Longes-L (Longest Common
  Subecqunce
- BLEU@1-4 (-Ngram ova

Measures quality: n-gram ovalap, seantrip, semantic simidarity, fluency

Main Results and Comparison on MSVD and MSR-VTT

MSVD Dataset

Metric	MSVD	METEOR	Previous SOTA	Brevious
CIDer	98.2%	95.2%	31.1%	43.2%
METEOR	35.3%	35.7%	34.1%	45.3%
ROUGE-L	59.3%	59.3%	57.78	48.2%
BLEU@4	45.1%	56.2%	56.2%	43.2%

MSR-VTT Basaeet

Model	CIDer	METEOR	S2.2-L	MSU-VT
MCC Model	47.6%	31.1%	52.2%	31.6%
Transformer	44.2	38.4	28.5	28.0
HCRN	45.8	28.8	50.9	28.1
X-LAN	46.5	20.9	51.11	30.5

- The MCC model consistently achieved the of-state or highly compettlince acrose both MSVD globally, demonisrating accurate and diverse video capbally.

The MCC model consistently achieved strerte of uretetive baslines... our MCC model our MSV-VTT detıT datesets, to model suprases X-LAN by **+1.2%** in **METEOR** and demonsating its effecteumess in signfrcantly Transforrer by **+2.7%** in **METOR** and ŝ **+3.4%** in €3.4% in CIDE.

Ablation Study: Impact Impact of Key Components

Component Removed

	CIDer (%) Removed	METEOR (%)	Key Insight
Baseline (w/o MCED)	43.1	Critical role of multi-processing	Repettive pharses.
w/o MCRE	27.8	Encoder builds rich cross-modal	Encoder levention
w/o MCCD	45.7	cross- modal represenations	Lacks caption diversity & quality
w/ DA	47.0	Decoder leverages multi-level context	Lacks diversity & ity
MC Model:	47.6	State-of-state performance	

Each component is vital for inenus & locally.
Each component ior MCC model's superior performance

Qualittive Analysis and Diversity Improvement

MCC Model: Richer Descriptions

More Accurate & Diverse: Captures specific details.

Baseline: "A person is playing guitar."

MCC: "A man strumming acoustic guitar, focuing intently on the music."

Diversity Attention (DA) Module

Encourges Broader
Avoids repettion

Encourges Broader Vocabuluy:

Dynamic Attention Adjustment:
Balances accuracy & variation.

More Natural & Human-like

Figure 3: Qualitative Improvements

BASELINE: "A person is playing guitar."

MCC + DA: "A man is acoustic guitar, focuing on the music."

Conclusion and Future Work

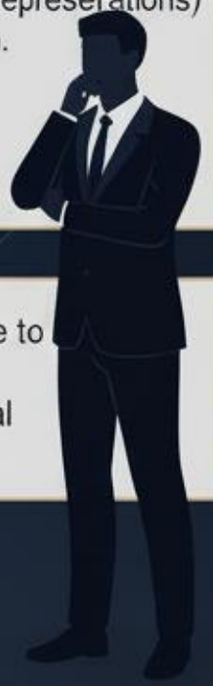
Key Contributions

1. Cross-modal Context Interaction (CCI): Enables fine-grained fusion of visual and textual features.

2. Multi-level Cross-modal Encoer-Decorder (MCED): it Comparining MCRE and MCCD, it processss of multiple multiple semantic levels for robuse stath CIDer on 97.6% for robust encoding ad decoding.

Future Work

- Extend to incoporate external knowledge for richer semarli. mdeasto
- Explore sophisticetate spat-temeentical reasoning (e.ıp. graph-based represerations) video video content).



Experimental results on MSVD and MSRTT datasets demonstated demonstate to performance, achievimg sttate up MS.2% CIDer on MSEr and MSR CIDe: Ablattion studies confimed con..ridies confimed MCC onsmer each critical bompamaklon by produice accurate, detaiel, and diverse descriptions.