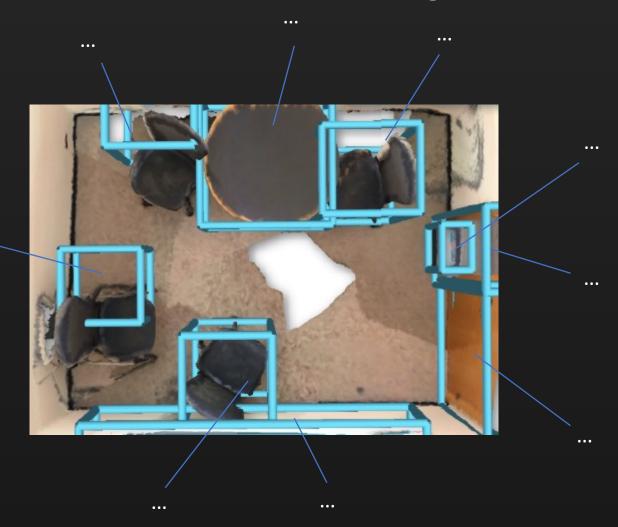
Dense Captioning for 3D Scenes with Transformers

- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

I. Scan2Cap: 3D Dense Captioning

This is a black office chair.
It is in the corner next to a black chair.



I. Scan2Cap

Point Cloud

Captions for the
Object Proposals

I. Scan2Cap: Architecture

Point Cloud

Object Detection

Module

Relational Graph Module Captioning Module

Captions for the Object Proposals

I. Scan2Cap: Architecture

PointNet++ **Voting Module** Relational Graph **Point Cloud** Module **Proposal Module Object Proposals** with Features Object Masks

Captioning Module

Captions for the Object Proposals

Object Detection Module

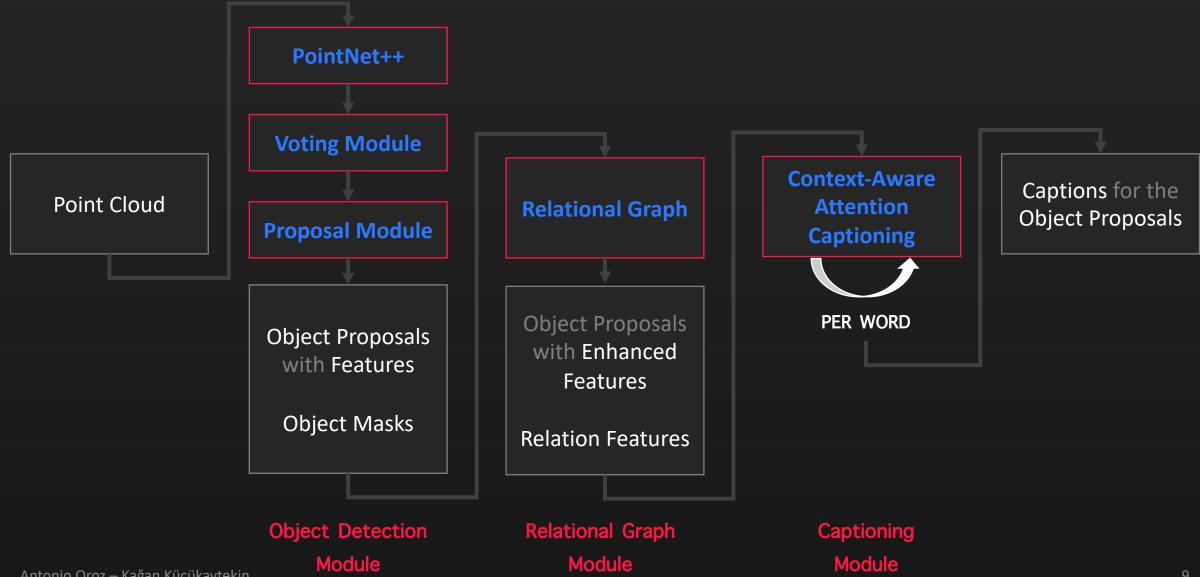
L Scan2Cap: Architecture

PointNet++ **Voting Module Point Cloud Relational Graph Proposal Module Object Proposals Object Proposals** with Enhanced with Features **Features Object Masks** Relation Features Relational Graph **Object Detection** Module Module

Captioning Module

Captions for the **Object Proposals**

L Scan2Cap: Architecture



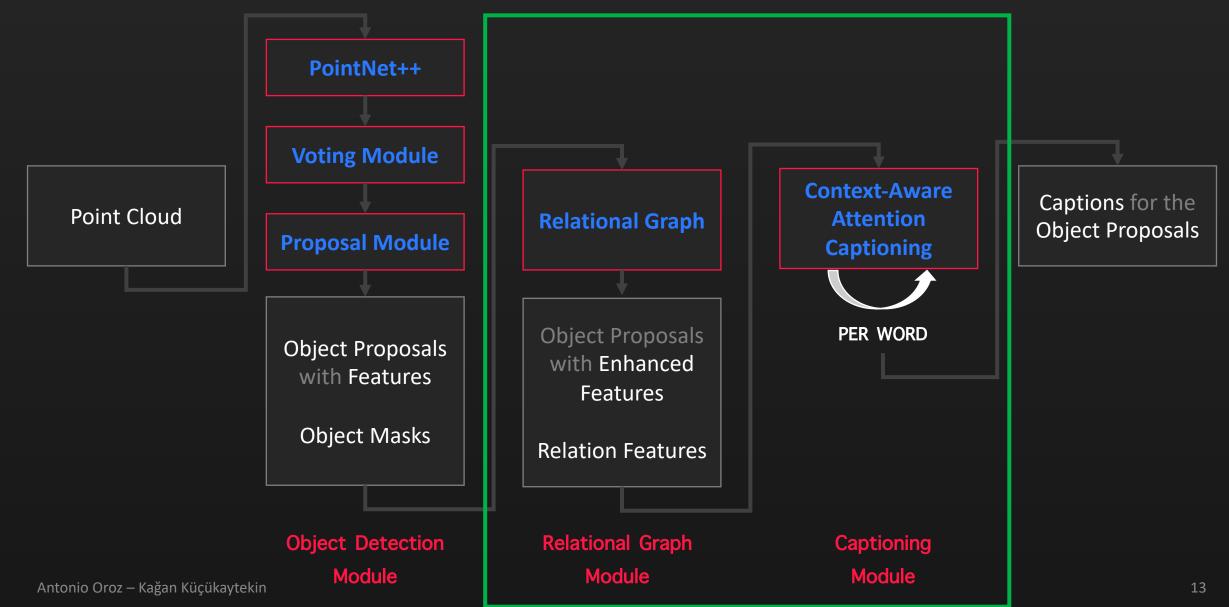
- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

II. Scan2CapMMT: Motivation for MMT

PointNet++ **Voting Module Context-Aware** Captions for the **Point Cloud Attention Relational Graph Object Proposals Proposal Module Captioning** PER WORD **Object Proposals Object Proposals** with Enhanced with Features **Features Object Masks** Relation Features Relational Graph **Object Detection** Captioning Module Module Module Antonio Oroz – Kağan Küçükaytekin 12

II. Scan2CapMMT: Motivation for MMT



II. Scan2Cap

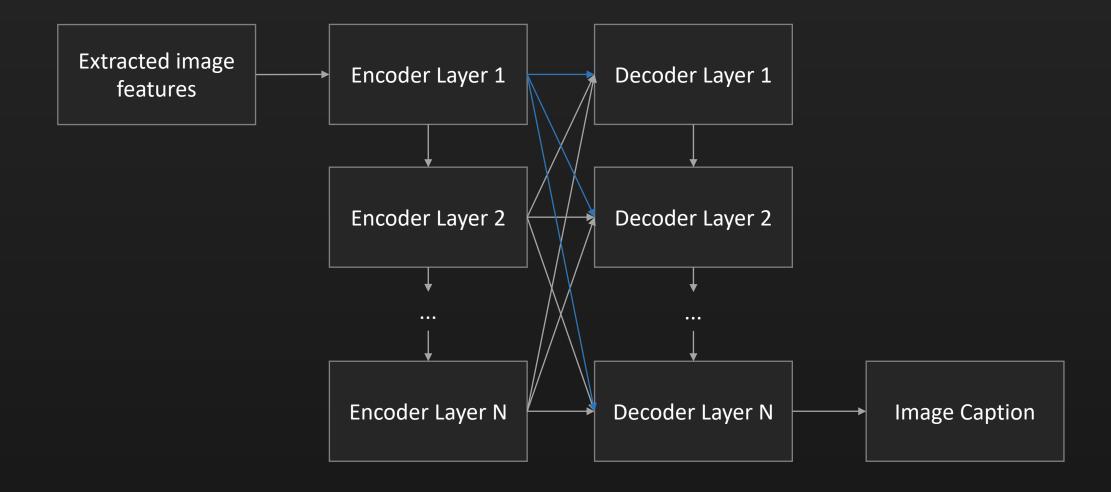
Point Cloud

Captions for the
Object Proposals

II. Meshed-Memory Transformer

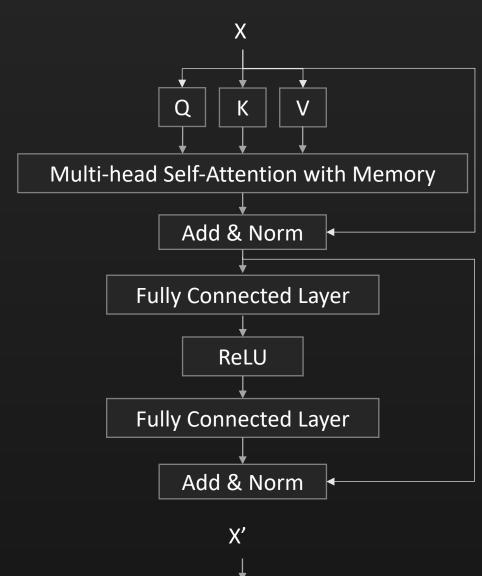
Extracted image features Image Caption

II. Meshed-Memory Transformer

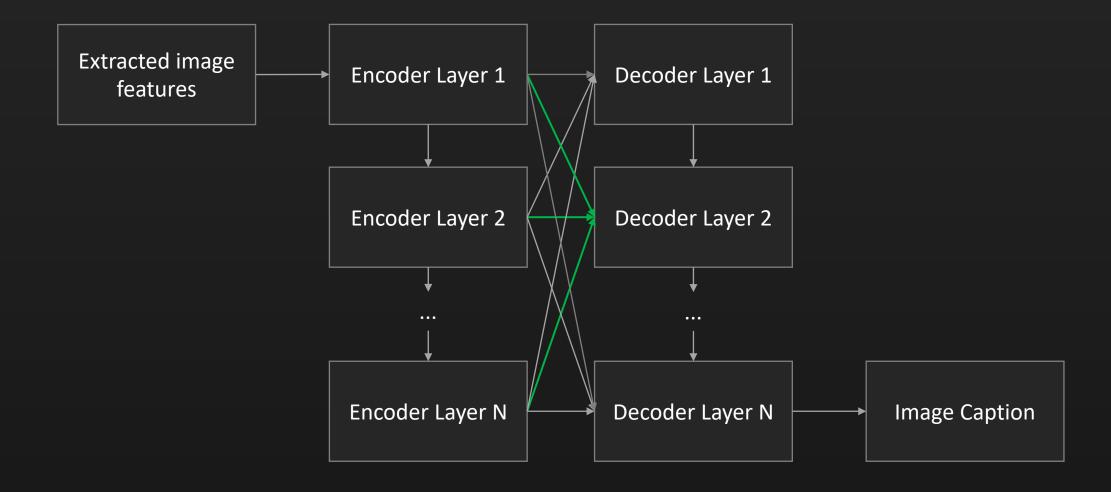


II. Meshed-Memory Transformer: Encoder

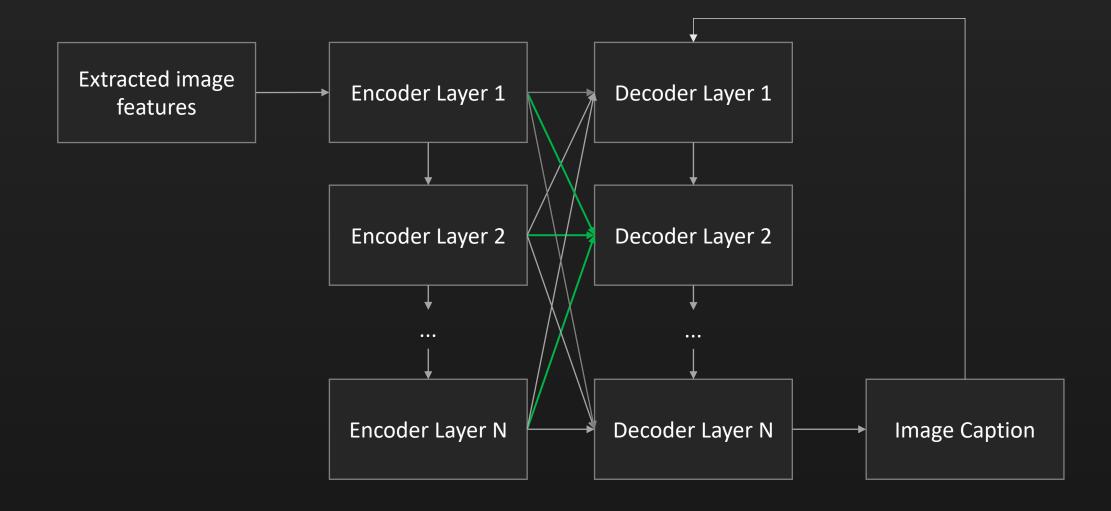
Encoder Layer 1



II. Meshed-Memory Transformer



II. Meshed-Memory Transformer



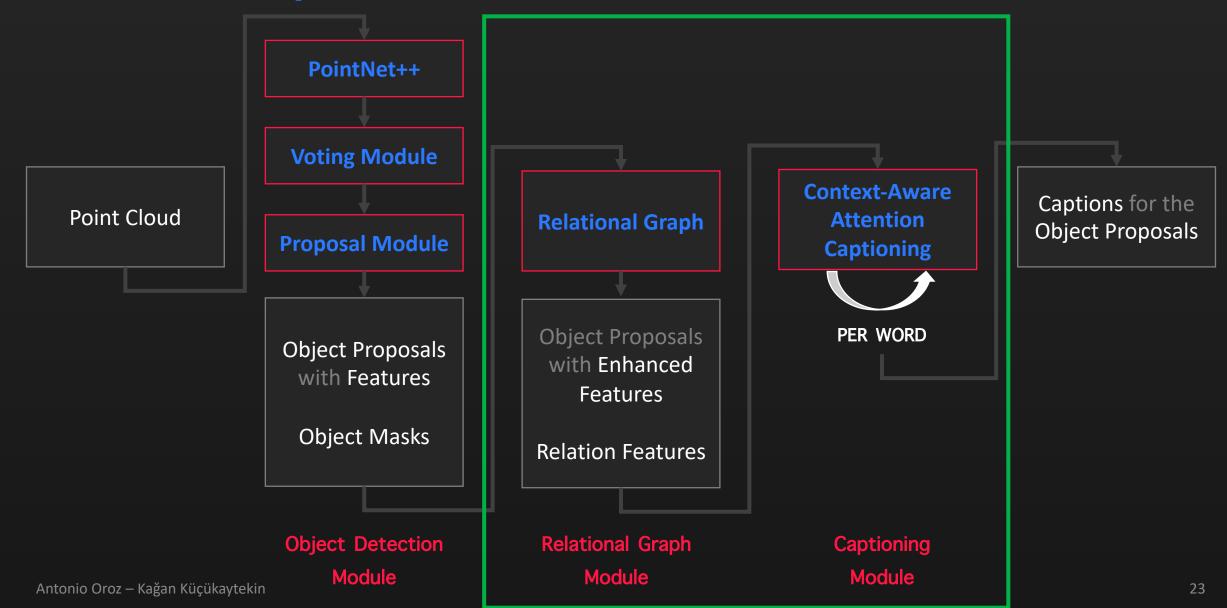
II. Meshed-Memory Transformer: Decoder

Decoder N+1 Decoder N Add & Norm FC ReLU Decoder Layer N Add & Norm Encoders Multi-Head Cross-Attention Add & Norm Masked Multi-Head Self-Attention Decoder N-1

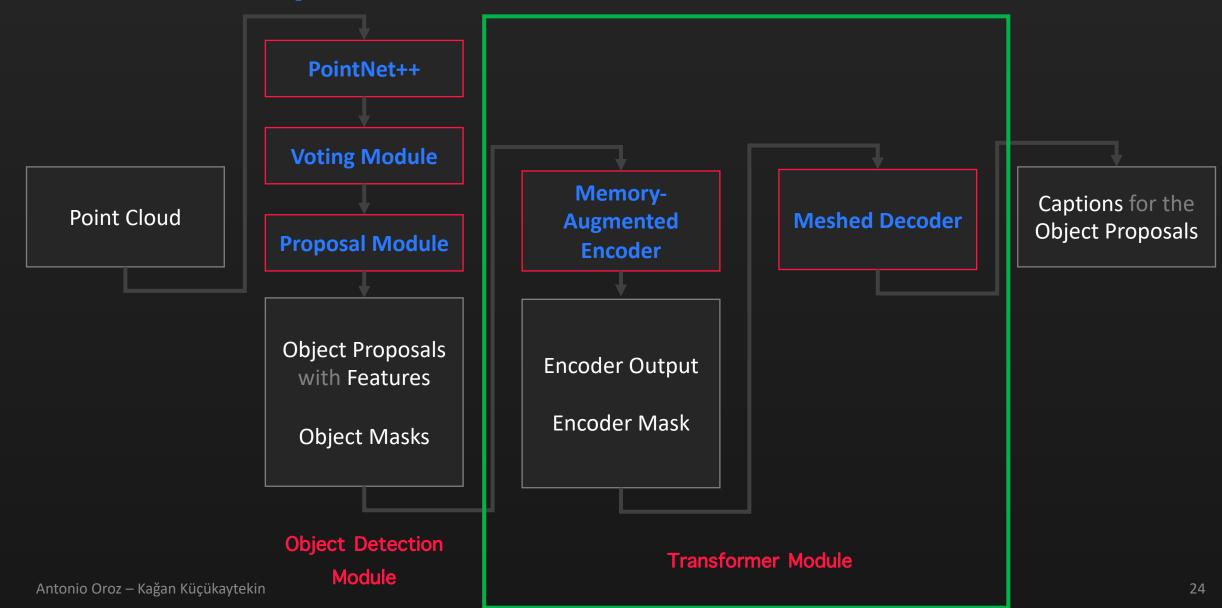
- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

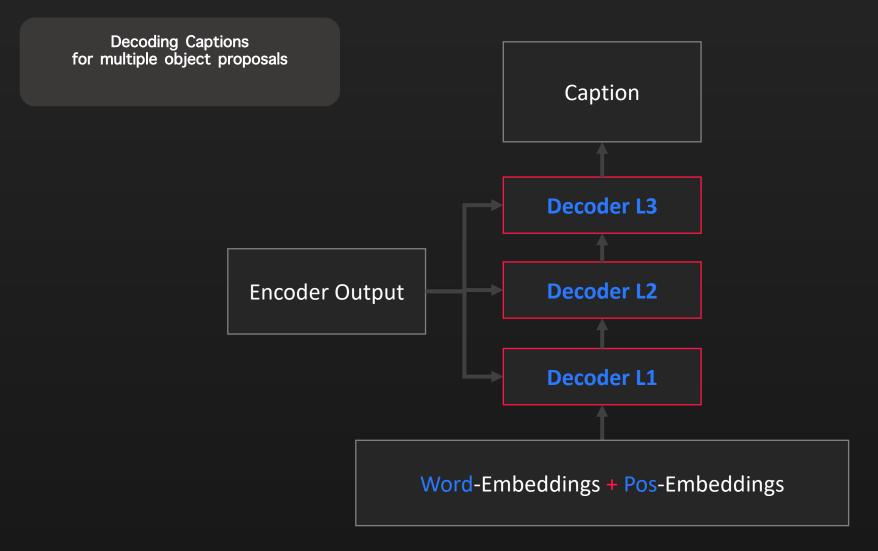
- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

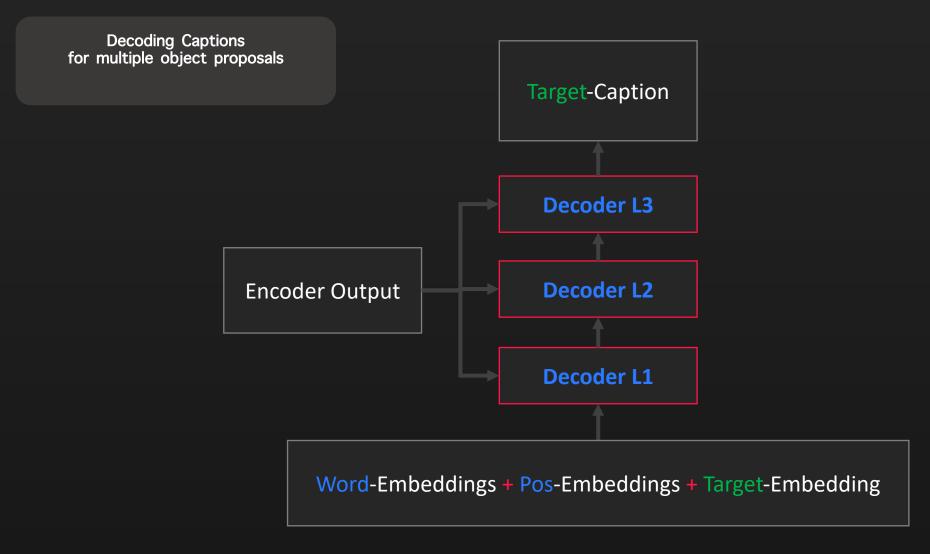
III. Scan2CapMMT: Initial Architecture



III. Scan2CapMMT: With MMT

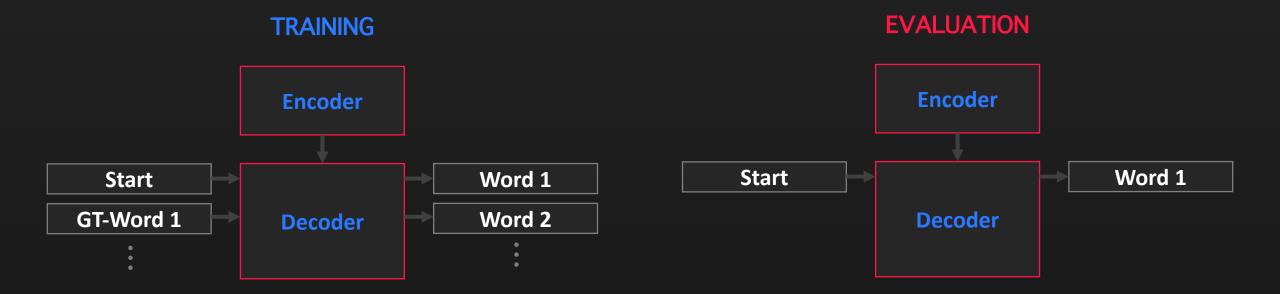






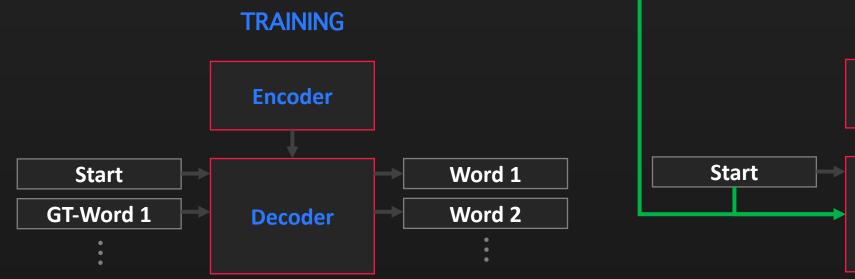
Decoding Captions for multiple object proposals

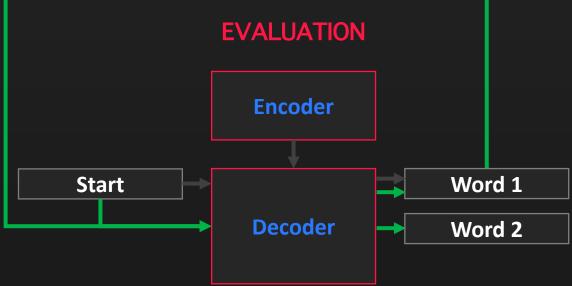
Caption-Generation in Training and Evaluation



Decoding Captions for multiple object proposals

Caption-Generation in Training and Evaluation





- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

- I. Scan2Cap
- **II.** Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

IV. Insights & First Results:

- Parameters: Scan2Cap MMT **7,830,308** vs **6,175,612** Scan2Cap
- Dropout: 0
- Weight Decay: 0
- Learning Rate: Changed from 1e-3 to 1e-4

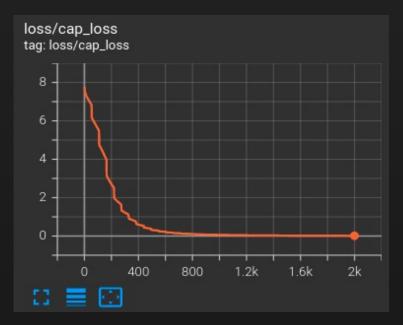
Antonio Oroz – Kağan Küçükaytekin

31

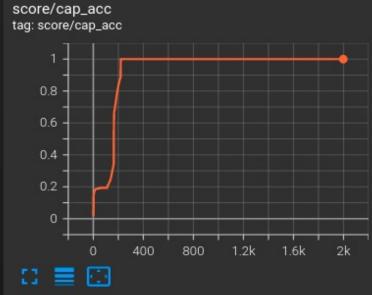
V. Insights & First Results: Overfitting Results

1 SAMPLE 1 SCENE

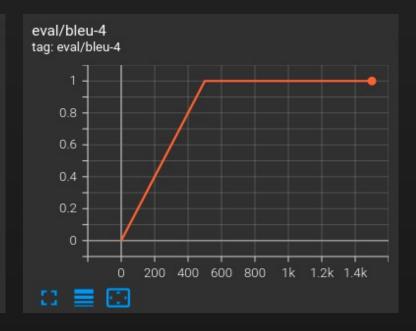
Caption Loss



Caption Accuracy



BLEU-4 Score

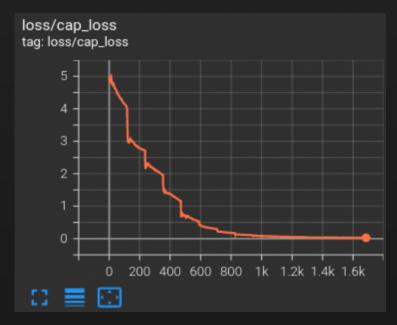


V. Insights & First Results: Overfitting Results

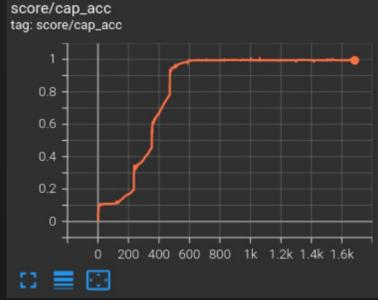
1 SAMPLE 1 SCENE

N SAMPLES 1 SCENE

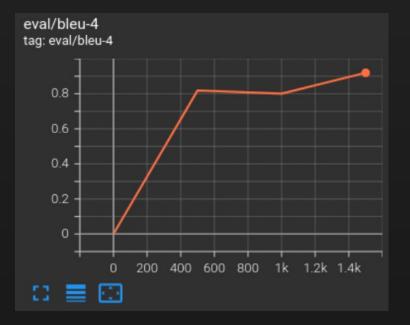
Caption Loss



Caption Accuracy



BLEU-4 Score



V. Insights & First Results: Overfitting Results

1 SAMPLE 1 SCENE

N SAMPLES 1 SCENE

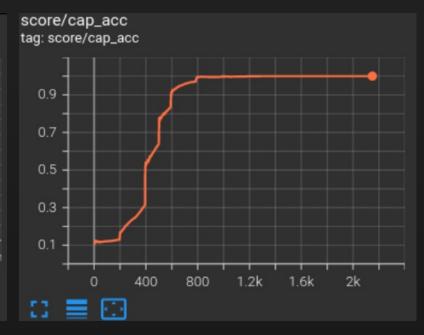
N SAMPLES M SCENES

Caption Loss loss/cap_loss tag: loss/cap_loss

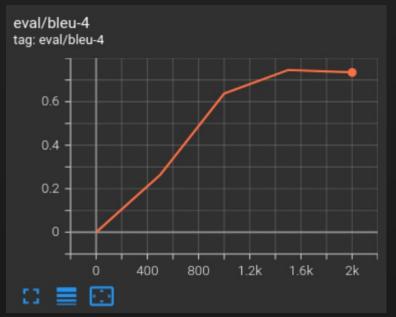
1.2k

1.6k

Caption Accuracy



BLEU-4 Score

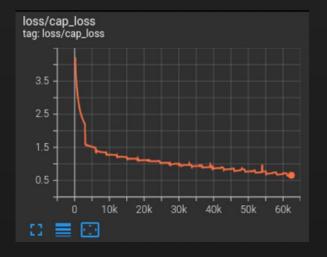


Antonio Oroz – Kağan Küçükaytekin

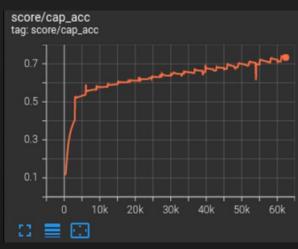
V. Insights & First Results: Training on the whole Dataset

Losses & Accuracies

Caption Loss



Caption Accuracy



V. Insights & First Results: Training on the whole Dataset

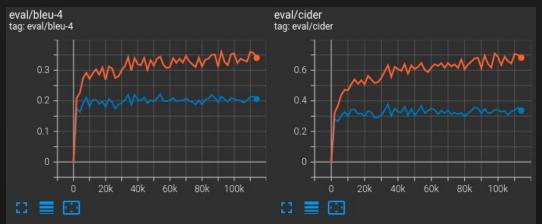
Losses & Accuracies

Evaluation

Scan2CapMMT



Scan2Cap



- I. Scan2Cap
- **II.** Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

- I. Scan2Cap
- II. Meshed-Memory Transformer
- III. Scan2CapMMT
- IV. Insights & First Results
- V. Next Steps

BEAM SEARCH

Instead of generating one sentence for an object proposal, generate multiple sentences in parallel and choose the final sentence with log propobabilities.

BEAM SEARCH

REINFOCEMENT LEARNING

After pretraining on the Cross-Entropy loss, use Reinforcement Learning with CIDEr-D as a reward to train the model.

BEAM SEARCH

REINFOCEMENT LEARNING

HYPERPARAMETER TUNING

Internal Dimensions of MMT

Decoder-/Encoder-Layers

Learning Rate

Number of Proposals

Schedules

Weight Decay

• • •

BEAM SEARCH

REINFOCEMENT LEARNING

HYPERPARAMETER TUNING

GROUP-FREE TRANSFORMER

Replace the current detection module with the Group-Free 3D Object Detection via Transformers module proposed by Liu et al.

Antonio Oroz – Kağan Küçükaytekin

42

BEAM SEARCH

REINFOCEMENT LEARNING

HYPERPARAMETER TUNING

GROUP-FREE TRANSFORMER

AoA

MMT currently uses Dot-Product Attention which we could replace with Attention on Attention

- I. Scan2Cap
- **II.** Meshed-Memory Transformer
- III. Scan2Cap with MMT
- IV. Insights & First Results
- V. Next Steps



THANK YOU FOR YOUR ATTENTION: D