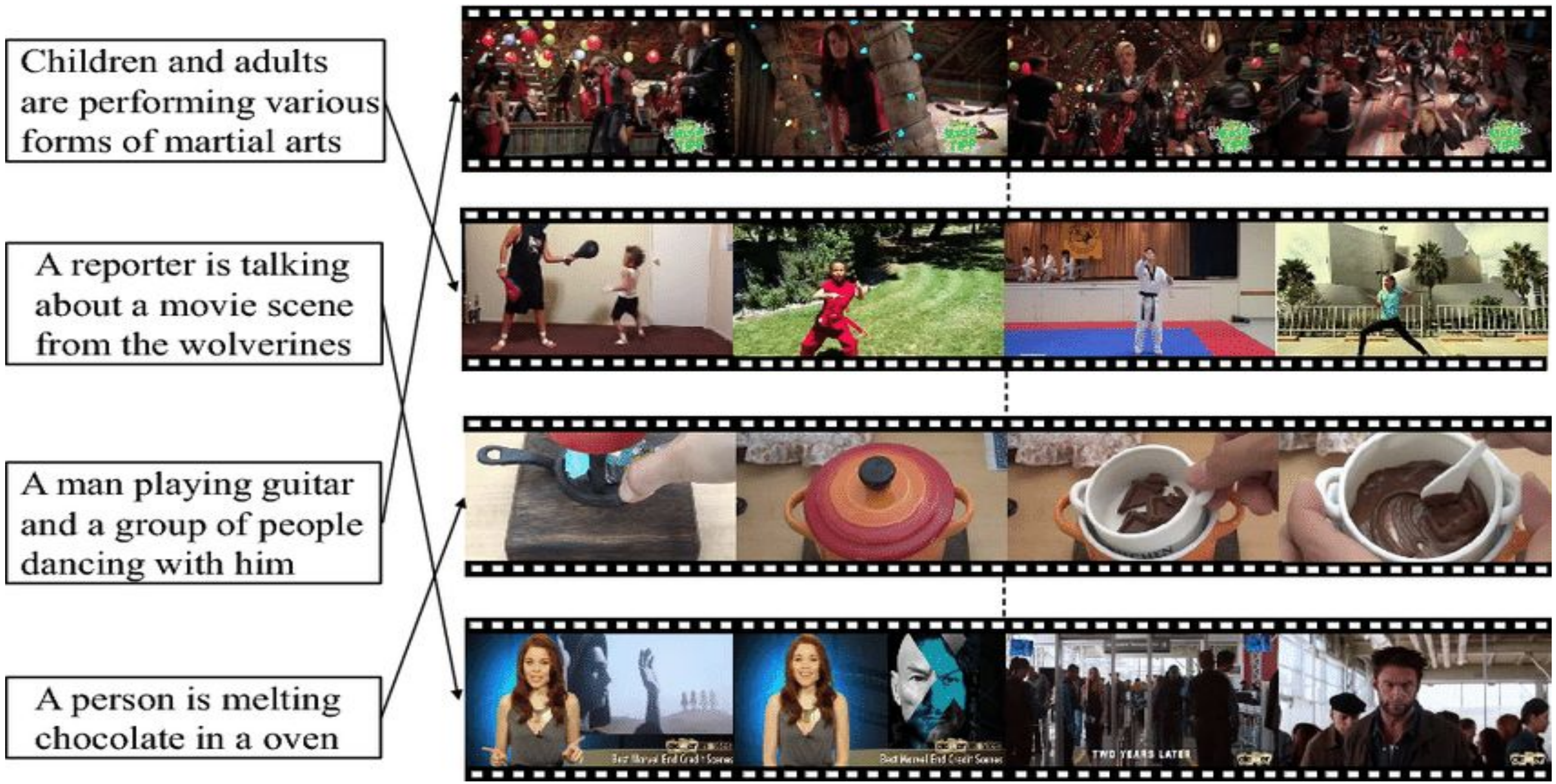
VIDEO-TEXT REPRESENTATION LEARNING

Yash Sharma – yss26 **Monaal Sanghvi – ms3185**

Jenil Jain – jj822 **Akshay Patil – avp119**

Introduction



- As we start dealing with multimodal data increasingly, using appropriate methods to deal with these different modalities becomes important.
- Video- text retrieval is an extremely relevant task in today's world where we deal with these two modalities widely.

Dataset

- Youcook II

- It contains **2000** long untrimmed videos from **89** cooking recipes; on average, each distinct recipe has **22** videos.

- The videos are all in the third-person viewpoint.

Why COOT?

- Cooperative Hierarchical Transformer. Uses **different hierarchies** in text and video to generate best possible representation

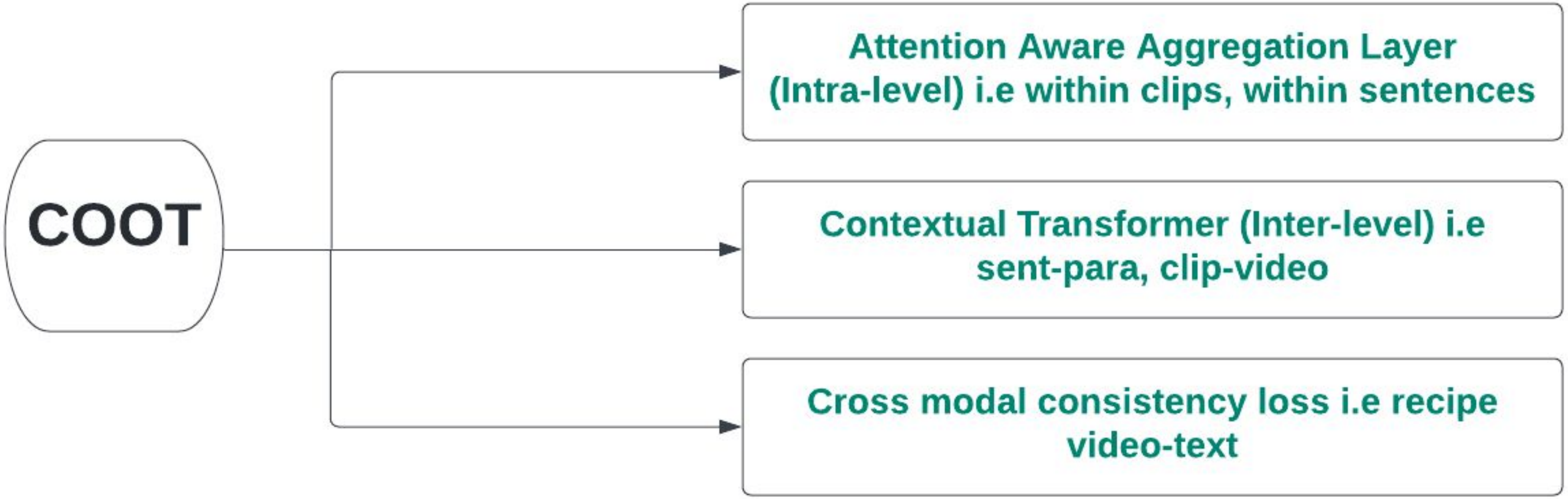
- Other multi-modal transformers like MDMMT use **too many pretrained expert models** for feature extraction.

- COOT model has only **10.6 million** parameters which **reduces** the training time to 3 hours on two GTX1080Ti GPU's

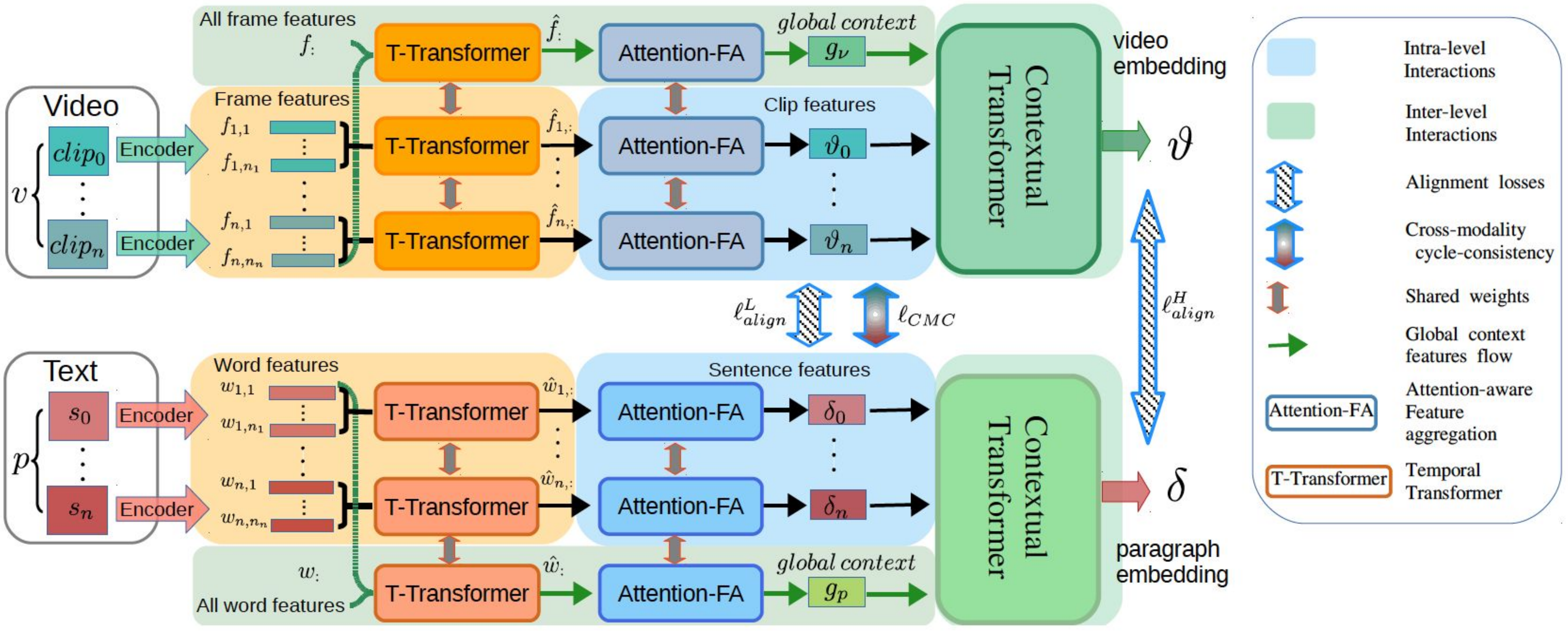
- **Step localisation** enforced by alignment losses

  Video-text alignment example: If a sentence contains the word soup, the most similar video frame to it must contain the picture of soup

Major Components



COOT - Architecture



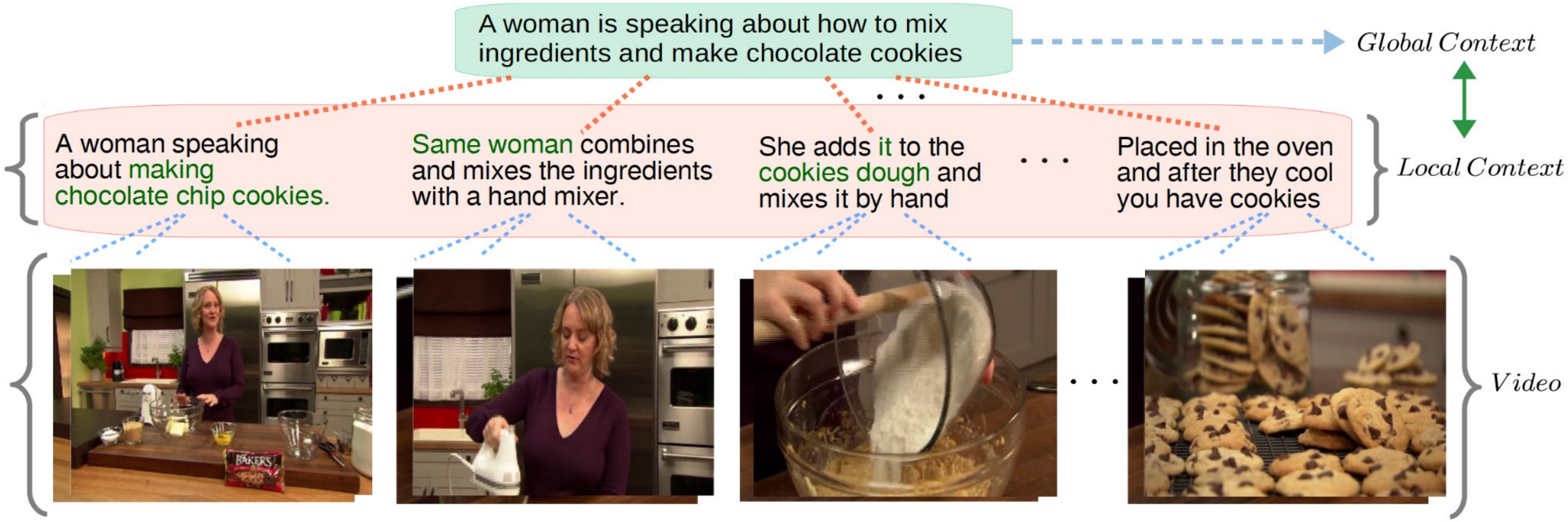
Brief overview

1. **Where are the features taken from?**
   - Video features (Frames) - from model pre-trained on HowTo100M dataset
   - Text features (Words) - Bert-based uncased

2. **What is the hierarchy in data?**
   - Words and frames, sentences and clips, paragraphs and videos. Each of these have different semantics

3. **How are the levels of hierarchy related in the transformer model?**
   - Input frame/word features into a T-transformer
   - Feed output into aggregation module to get sentence/clip level features.
   - Final video/text embeddings produced from contextual transformer.

Global and Local Context



- In the third sentence, to know the type of dough (cookie) the model should have information about the general context of the video (making chocolate cookies).
- Likewise, in the second sentence, to know that she is the "same woman", the model must be aware of the person's identity throughout the video.
- Therefore to get accurate representation we must know the **local and the global context.**

Loss Functions

- **Alignment Loss** (Goal is to push apart embeddings for negative samples)

P = Positive sample
N = negative sample
alpha = margin

$$L(\mathcal{P}, \mathcal{N}, \alpha) = max(0, \alpha + D(x, y) - D(x', y)) + max(0, \alpha + D(x, y) - D(x, y'))$$

Maximise distance wrt to negative video sample

Maximise distance wrt to negative text sample

where $D(x, y) = 1 - x^\intercal y / (\|x\| \|y\|)$ is the cosine distance of two vectors.

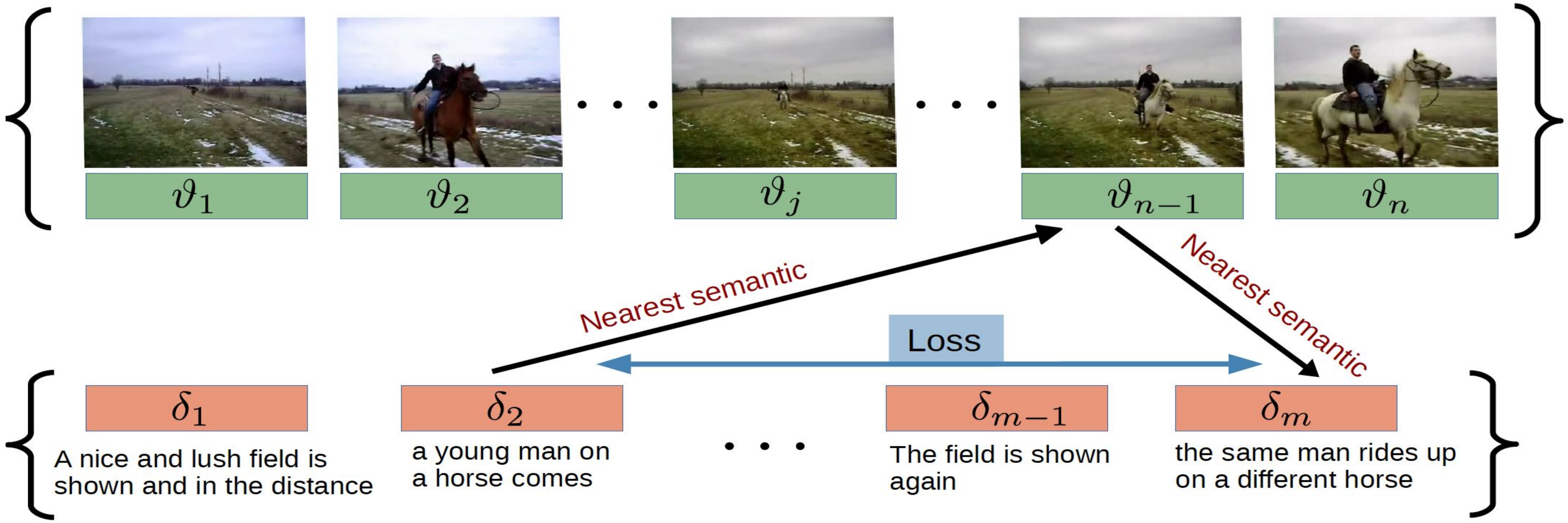
- **Cluster Loss**

$$\ell_{cluster} = \sum_{k \in \mathcal{D}, i, k' \neq k, i' \neq i} L((1,1), \{(\vartheta_i^k, \vartheta_{i'}^{k'}), (\delta_{i'}^{k'}, \delta_i^k)\}, \gamma) + \sum_{k \in \mathcal{D}, k' \neq k} L((1,1), \{(\vartheta^k, \vartheta^{k'}), (\delta^{k'}, \delta^k)\}, \eta)$$

- **Final Loss**

$$l_{final} = l_{align}^{\mathrm{L}} + l_{\mathrm{align}}^{\mathrm{H}} + l_{\mathrm{align}}^{\mathrm{g}} + l_{\mathrm{cluster}} + l_{\mathrm{CMC}}$$

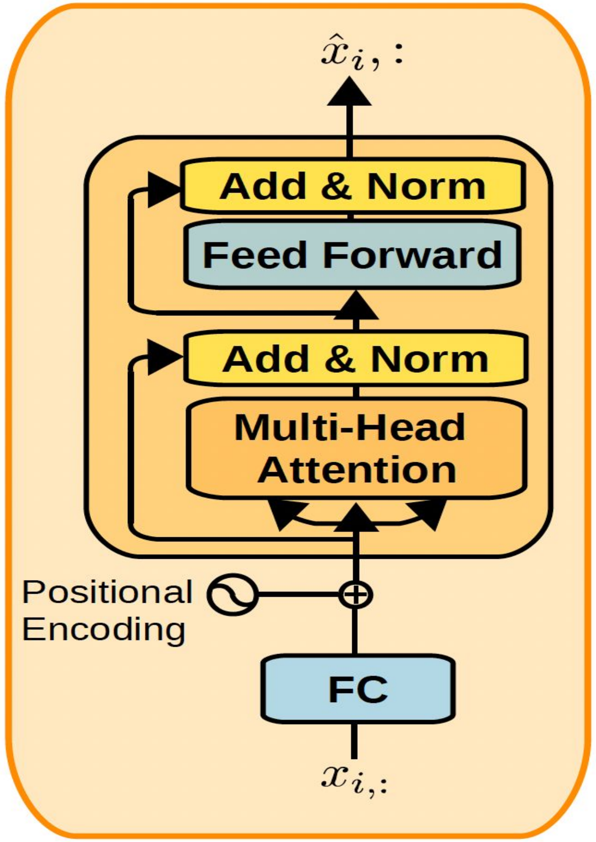
Cross Modal Cycle Consistency



- For a sentence s_i , we find its **nearest neighbor** in the clip sequence and again its neighbor in the sentence sequence.
- Semantically cycle consistent if and only if it cycles back to the original location.
- It **penalizes deviations** from cycle-consistency

Temporal Transformer

Temporal Transformer

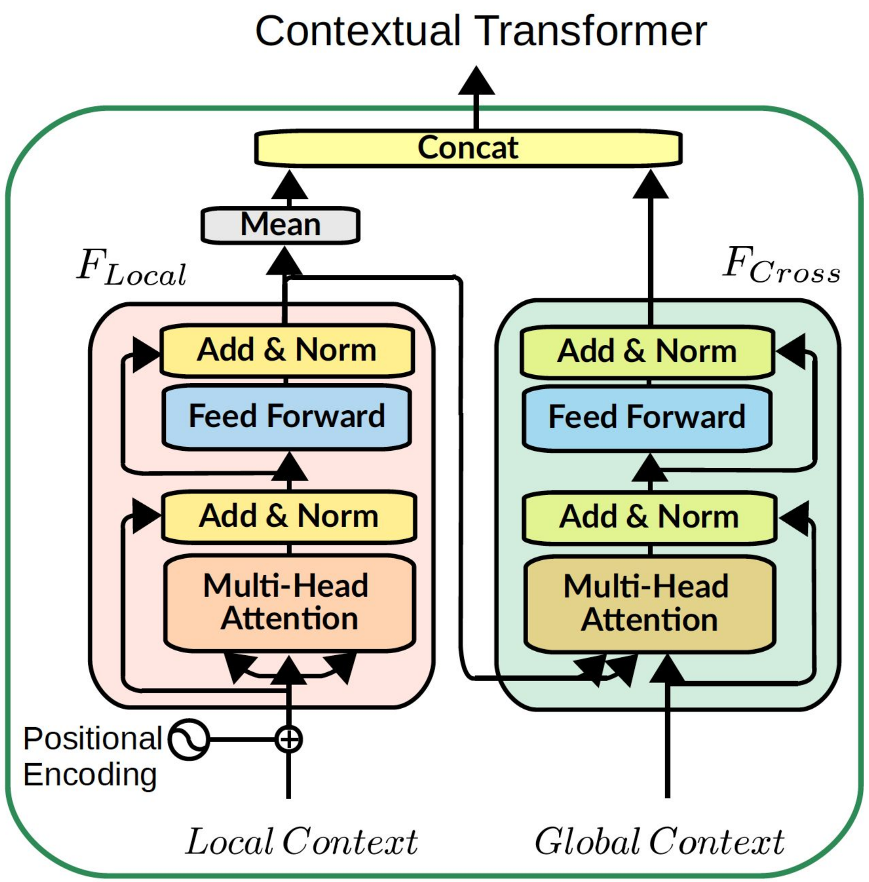


- Captures relationship between **frame/word features**
- Two temporal transformers for **recipe text, and video**
- For a video, encode frames to get **frame level features**. Encode text to get word level features
- Pass features through an **Attention aware feature aggregation layer**.

Attention Aware Feature Aggregation Layer

- Produces **clip/sentence level features**
- Standard feature fusion methods consider each feature independently by average pooling or max pooling.
- Hence, they **miss** the relationship between features to highlight the relevant features.
- In cooking videos, **objects on the table are more important** than objects in the background. Therefore we need to attend to those objects more.
- Hence the use of **attention aware layer for feature aggregation**

Contextual Transformer



- Produces final **video and text embeddings**

- Compute **key (K)-value(V) pairs** based on these embeddings and query(Q) based on the global context.

Results

Retrieval Type	R@1	R@5	R@10	R@50	Median Rank
Video-Paragraph	0.810	0.958	0.978	0.996	2.2
Paragraph-Video	0.783	0.963	0.978	0.996	2.3
Clip-Sentence	**0.159**	**0.395**	**0.512**	**0.782**	**74.4**
Sentence-Clip	**0.169**	**0.406**	**0.525**	**0.780**	**73.2**

Conclusion

- We observe that the retrieval metrics for the global context are much better than those for the local context. This might be true because the global context captures more information.
- Ongoing work: Fine-tuning CLIP (Results to be included in the report for a small subset of YouCook2)