**Title:** Sports Video Analysis on Large-Scale Data **Authors:** Dekun Wu, He Zhao, Xingce Bao, Richard P. Wildes (2022)

**https://arxiv.org/pdf/2208.04897v1**

### Key Concepts and Taxonomy

* **Domain:** Fine-grained sports video understanding.
* **Focus Tasks:**
  1. Video Captioning
  2. Fine-Grained Action Recognition
  3. Player Identification
* **Key Innovation:** Introduction of **NSVA**, a large-scale, publicly accessible NBA dataset with rich multimodal annotations and a unified model for the three tasks.
* **Architectural Basis:** Transformer-based pipeline using TimeSformer, Vision Transformers (ViT), and UniVL framework for multi-source feature fusion.

### Main Contributions and Findings

1. **Dataset – NSVA (NBA for Sports Video Analysis):**
   * Contains 32,019 clips from 132 NBA games (2018–2019 season).
   * Clips annotated with play-by-play captions, fine-grained action types (172 categories), and player identities (184 players).
   * Fully automated scraping and annotation pipeline; only 1,000 images manually annotated for training ball/basket detectors.
   * Enables scalable and reproducible research in sports video captioning and analysis.
2. **Model Architecture:**
   * Utilizes **TimeSformer** for capturing long-range spatiotemporal context.
   * Incorporates **fine-grained ViT-based features**: ball, basket, player with ball (PB), and courtline segmentation (position-aware, PA).
   * Combines features via a **cross-encoder** and unified **transformer decoder** for task-specific outputs.
3. **Performance:**
   * Outperforms baseline models (MP-LSTM, TA, Transformer, UniVL) on NSVA in captioning.
   * Ablation studies show incremental gains with each feature type (PB, BAL, BAS, PA).
   * Demonstrates NSVA’s potential in **hierarchical action recognition** and **player identification**.
   * Achieves 61.2% accuracy on coarse action recognition; 42.3% on fine-grained event labels.

### Limitations and Future Directions

* **Player Identification:**
  + Challenging due to blur, occlusions, and limited facial visibility.
  + Requires improved detection and tracking, possibly integrating facial or jersey recognition.
* **Feature Representation:**
  + Exploring advanced backbones (e.g., **Video Swin Transformer**) could enhance performance.
* **Multi-task Learning:**
  + Future work will apply **Prefix-Tuning** for jointly training captioning, recognition, and ID tasks.

**REFERENCE PAPERS**

**Detailed Summaries of References [1]–[50] from "Sports Video Analysis on Large-Scale Data"**

**[1] Aafaq et al., CVPR 2019**

<https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPR_2019/html/Aafaq_Spatio-Temporal_Video_Captioning_With_Object-Level_Semantic_Attribute_Enriched_Visual_CVPR_2019_paper.html>

* **Key Concepts and Taxonomy:** Spatio-temporal video captioning using visual encoding enriched with semantic attributes.
* **Main Contributions and Findings:** Proposed a method that captures high-level semantics and object interactions across time using attribute-based encodings.
* **Limitations and Future Directions:** Focused mainly on general videos, not specialized to sports domain.
* **Relevance to Primary Paper:** Inspired the use of enriched visual encoding for sports video captioning.

**[2] Aafaq et al., ACM Computing Surveys 2019**

<https://dl.acm.org/doi/10.1145/3301283>

* **Key Concepts and Taxonomy:** Survey on video description methods, datasets, and evaluation metrics.
* **Main Contributions and Findings:** Comprehensive analysis of techniques for generating textual descriptions from video.
* **Limitations and Future Directions:** Emphasizes generic datasets; lacks sports-specific insights.
* **Relevance to Primary Paper:** Provides foundational context for evaluating video captioning methods used in NSVA.

**[3] Banerjee and Lavie, ACL 2005**

<https://aclanthology.org/W05-0909.pdf>

* **Key Concepts and Taxonomy:** METEOR evaluation metric for machine translation.
* **Main Contributions and Findings:** Introduced METEOR to better correlate with human judgment.
* **Limitations and Future Directions:** Early metric; later improvements like CIDEr emerged.
* **Relevance to Primary Paper:** Used to evaluate captioning quality in NSVA.

**[4] Bertasius et al., ICML 2021**

<https://arxiv.org/abs/2102.05095>

<https://github.com/facebookresearch/TimeSformer>

* **Key Concepts and Taxonomy:** Space-time attention via TimeSformer for video understanding.
* **Main Contributions and Findings:** Proposed transformer-only model for spatiotemporal video representation.
* **Limitations and Future Directions:** Focus on general videos, not specific to fine-grained sports.
* **Relevance to Primary Paper:** TimeSformer is the base video encoder in NSVA model.

**[5] Bi et al., ICCV 2021**

<https://www.google.com/search?q=https://openaccess.thecvf.com/content/ICCV2021/html/Bi_Procedure_Planning_in_Instructional_Videos_ICCV_2021_paper.html>

* **Key Concepts and Taxonomy:** Procedure planning using contextual modeling and policy learning.
* **Main Contributions and Findings:** Addressed instructional videos with complex temporal dependencies.
* **Limitations and Future Directions:** Does not cover sports or identity-aware captions.
* **Relevance to Primary Paper:** Provides evaluation metrics and planning approaches applicable to NSVA.

**[6] Brown et al., NeurIPS 2020**

<https://arxiv.org/abs/2005.14165>

* **Key Concepts and Taxonomy:** Few-shot learning with large-scale language models (GPT-3).
* **Main Contributions and Findings:** Showed the capability of LMs to perform tasks with minimal examples.
* **Limitations and Future Directions:** Lacks video input; text-only.
* **Relevance to Primary Paper:** Motivates the use of pretrained transformers for video-text tasks.

**[7] Caba Heilbron et al., CVPR 2015**

<https://www.google.com/search?q=https://openaccess.thecvf.com/content_cvpr_2015/html/Caba_Heilbron_ActivityNet_A_Large-Scale_2015_CVPR_paper.html>

<http://activity-net.org/>

* **Key Concepts and Taxonomy:** ActivityNet dataset for human activity understanding.
* **Main Contributions and Findings:** Created a large benchmark dataset for complex activity recognition.
* **Limitations and Future Directions:** Does not include professional sports.
* **Relevance to Primary Paper:** NSVA’s benchmark design is inspired by ActivityNet's structured format.

**[8] Carion et al., ECCV 2020**

<https://arxiv.org/abs/2005.12872>

<https://github.com/facebookresearch/detr>

* **Key Concepts and Taxonomy:** DETR—end-to-end object detection with transformers.
* **Main Contributions and Findings:** Introduced transformer architecture to object detection.
* **Limitations and Future Directions:** High computational cost.
* **Relevance to Primary Paper:** Encouraged use of transformers in NSVA’s object detection.

**[9] Carreira and Zisserman, CVPR 2017**

<https://arxiv.org/abs/1705.07750>

* **Key Concepts and Taxonomy:** Action recognition and Kinetics dataset.
* **Main Contributions and Findings:** Provided a strong benchmark and new model (I3D).
* **Limitations and Future Directions:** Limited domain specificity.
* **Relevance to Primary Paper:** Influences NSVA’s video encoding approach.

**[10] Chang et al., ECCV 2020**

<https://www.google.com/search?q=https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1831_ECCV_2020_paper.php>

* **Key Concepts and Taxonomy:** Instructional video procedure planning.
* **Main Contributions and Findings:** Modeled temporal structures in instructional videos.
* **Limitations and Future Directions:** Not adapted to sports.
* **Relevance to Primary Paper:** Basis for NSVA’s fine-grained action evaluation.

**[11] Chen and Dolan, ACL 2011**

<https://aclanthology.org/P11-1042.pdf>

* **Key Concepts and Taxonomy:** Parallel data collection for paraphrase evaluation.
* **Main Contributions and Findings:** Built a dataset from newswire headlines for sentence-level paraphrasing.
* **Limitations and Future Directions:** Not focused on video or multimodal data.
* **Relevance to Primary Paper:** Influences the handling of diverse textual captions in NSVA.

**[12] Chen et al., arXiv 2018**

<https://arxiv.org/abs/1805.00334>

* **Key Concepts and Taxonomy:** Dense-captioning events in video.
* **Main Contributions and Findings:** System report combining multiple features for event-level video captioning.
* **Limitations and Future Directions:** Prototype-level insights; lacks reproducibility.
* **Relevance to Primary Paper:** Provided captioning baselines for benchmarking NSVA.

**[13] Damen et al., ECCV 2018**

<https://arxiv.org/abs/1804.02748>

<https://epic-kitchens.github.io/>

* **Key Concepts and Taxonomy:** Egocentric video dataset—EPIC-Kitchens.
* **Main Contributions and Findings:** Introduced a massive first-person action video benchmark.
* **Limitations and Future Directions:** Focused on kitchen and daily activities.
* **Relevance to Primary Paper:** Methodology and fine-grained action taxonomy useful for NSVA design.

**[14] Devlin et al., NAACL 2019**

<https://arxiv.org/abs/1810.04805>

* **Key Concepts and Taxonomy:** BERT—deep bidirectional transformer for language understanding.
* **Main Contributions and Findings:** Pretrained masked language model achieving state-of-the-art results.
* **Limitations and Future Directions:** Requires task-specific finetuning.
* **Relevance to Primary Paper:** BERT vocabulary and structure are used in NSVA’s language decoder.

**[15] Dosovitskiy et al., ICLR 2021**

<https://arxiv.org/abs/2010.11929>

* **Key Concepts and Taxonomy:** Vision Transformer (ViT) for image recognition.
* **Main Contributions and Findings:** Demonstrated transformers can match CNNs in vision tasks.
* **Limitations and Future Directions:** Needs large datasets to perform well.
* **Relevance to Primary Paper:** Used to extract object-level and positional features in NSVA.

**[16] Feichtenhofer et al., ICCV 2019**

<https://github.com/facebookresearch/SlowFast>

<https://arxiv.org/abs/1812.03982>

* **Key Concepts and Taxonomy:** SlowFast networks for video recognition.
* **Main Contributions and Findings:** Dual-pathway network for fine-grained temporal modeling.
* **Limitations and Future Directions:** Not integrated with transformers.
* **Relevance to Primary Paper:** Related to temporal modeling techniques considered in NSVA.

**[17] Isola et al., CVPR 2017**

<https://arxiv.org/abs/1611.07004>

* **Key Concepts and Taxonomy:** Image-to-image translation using conditional GANs (pix2pix).
* **Main Contributions and Findings:** Framework to map one image domain to another using paired data.
* **Limitations and Future Directions:** Needs aligned image pairs for training.
* **Relevance to Primary Paper:** Used for courtline segmentation in NSVA.

**[18] Jhamtani et al., ACL 2018**

<https://aclanthology.org/P18-1033.pdf>

* **Key Concepts and Taxonomy:** Move-by-move commentary generation for chess.
* **Main Contributions and Findings:** NLP model trained on online chess forums to generate commentary.
* **Limitations and Future Directions:** Domain-specific; limited visual input.
* **Relevance to Primary Paper:** Motivation for identity-aware commentary in sports.

**[19] Jocher et al., YOLOv5 GitHub 2021**

<https://github.com/ultralytics/yolov5>

* **Key Concepts and Taxonomy:** YOLOv5 object detection framework.
* **Main Contributions and Findings:** High-speed real-time object detection system.
* **Limitations and Future Directions:** Not end-to-end transformer-based.
* **Relevance to Primary Paper:** Backbone for player and ball detection in NSVA.

**[20] Kay et al., arXiv 2017**

<https://arxiv.org/abs/1705.06950>

* **Key Concepts and Taxonomy:** Kinetics dataset for human action recognition.
* **Main Contributions and Findings:** Created a large-scale benchmark for action classification.
* **Limitations and Future Directions:** Limited to coarse-grained actions.
* **Relevance to Primary Paper:** Provided baseline inspiration for NSVA's dataset.

**[21] Kim et al., ECCV 2020**

<https://www.google.com/search?q=https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1739_ECCV_2020_paper.php>

* **Key Concepts and Taxonomy:** Learning temporal alignment for procedural videos.
* **Main Contributions and Findings:** Proposed alignment-aware modeling between visual segments and language tokens.
* **Limitations and Future Directions:** Focused on cooking videos, lacks multi-agent interactions.
* **Relevance to Primary Paper:** Informs NSVA’s modeling of step-wise action sequences in sports.

**[22] Kuehne et al., CVPR 2014**

<https://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/>

<https://www.google.com/search?q=https://openaccess.thecvf.com/content_cvpr_2014/html/Kuehne_A_Dataset_for_2014_CVPR_paper.html>

* **Key Concepts and Taxonomy:** Breakfast dataset for fine-grained activity recognition.
* **Main Contributions and Findings:** Introduced a benchmark with long videos and labeled sub-actions.
* **Limitations and Future Directions:** Limited domain; low motion variation.
* **Relevance to Primary Paper:** NSVA’s fine-grained labels take cues from the semantic structure here.

**[23] Li and Liang, NeurIPS 2021**

<https://arxiv.org/abs/2101.00190>

* **Key Concepts and Taxonomy:** Prefix tuning for multi-task learning.
* **Main Contributions and Findings:** Lightweight adaptation for transformers across tasks.
* **Limitations and Future Directions:** Needs task-specific prefix design.
* **Relevance to Primary Paper:** Suggested for future NSVA work to enable unified training.

**[24] Li et al., CVPR 2018**

<https://www.google.com/search?q=https://openaccess.thecvf.com/content_cvpr_2018/html/Li_Diving48_A_Fine-Grained_CVPR_2018_paper.html>

* **Key Concepts and Taxonomy:** Diving48 dataset for fine-grained action classification.
* **Main Contributions and Findings:** Dataset with minimal intra-class variance, rich annotation.
* **Limitations and Future Directions:** Limited diversity of motion and camera views.
* **Relevance to Primary Paper:** Provides a blueprint for designing NSVA’s action taxonomy.

**[25] Lin et al., CVPR 2014**

<https://www.google.com/search?q=https://openaccess.thecvf.com/content_cvpr_2014/html/Lin_Temporal_Action_Localization_2014_CVPR_paper.html>

* **Key Concepts and Taxonomy:** Temporal proposal generation for activity detection.
* **Main Contributions and Findings:** Segment proposal generation in untrimmed videos.
* **Limitations and Future Directions:** Does not cover identity or textual grounding.
* **Relevance to Primary Paper:** NSVA action event detection leverages similar segmentation logic.

**[26] Lin et al., ICCV 2019**

<https://arxiv.org/abs/1811.08383>

* **Key Concepts and Taxonomy:** Temporal Shift Module (TSM) for efficient video understanding.
* **Main Contributions and Findings:** Introduced a lightweight temporal module to enable temporal modeling using 2D CNNs.
* **Limitations and Future Directions:** Limited in capturing long-range dependencies.
* **Relevance to Primary Paper:** NSVA’s model considers TSM as an alternative for efficient temporal representation.

**[27] Lin et al., ECCV 2014**

<https://cocodataset.org/>

<https://arxiv.org/abs/1405.0312>

* **Key Concepts and Taxonomy:** MS COCO dataset.
* **Main Contributions and Findings:** Provided a large dataset with object segmentation and caption annotations.
* **Limitations and Future Directions:** Geared toward images, not video.
* **Relevance to Primary Paper:** MS COCO-trained YOLOv5 used for player detection in NSVA.

**[28] Liu et al., ICCV 2021**

<https://arxiv.org/abs/2103.14030>

* **Key Concepts and Taxonomy:** Swin Transformer for hierarchical vision tasks.
* **Main Contributions and Findings:** Developed window-based attention mechanism for efficient computation.
* **Limitations and Future Directions:** Needs exploration in dense temporal tasks.
* **Relevance to Primary Paper:** Proposed for future work to enhance video representation in NSVA.

**[29] Liu et al., CVPR 2022**

<https://arxiv.org/abs/2106.13230>

* **Key Concepts and Taxonomy:** Video Swin Transformer.
* **Main Contributions and Findings:** Extended Swin Transformer to video for action recognition.
* **Limitations and Future Directions:** Performance on sports data yet to be benchmarked.
* **Relevance to Primary Paper:** Suggested as a stronger backbone than TimeSformer.

**[30] Lu et al., TPAMI 2013**

<https://www.google.com/search?q=https://ieeexplore.ieee.org/document/6298925>

* **Key Concepts and Taxonomy:** Player identification in sports videos.
* **Main Contributions and Findings:** Tracked and identified basketball players using CRFs.
* **Limitations and Future Directions:** Tested on controlled games.
* **Relevance to Primary Paper:** Early precursor to identity-aware tracking in NSVA.

**[31] Lu et al., CVPR 2011**

<https://www.google.com/search?q=https://ieeexplore.ieee.org/document/5995400>

* **Key Concepts and Taxonomy:** CRF-based player identification.
* **Main Contributions and Findings:** Combined player detection and identity inference.
* **Limitations and Future Directions:** Scalability to real games not evaluated.
* **Relevance to Primary Paper:** Influences NSVA’s goal of automatic player labeling.

**[32] Luo et al., arXiv 2020**

<https://arxiv.org/abs/2002.06353>

* **Key Concepts and Taxonomy:** UniVL—Unified Video and Language pretraining.
* **Main Contributions and Findings:** Unified transformer model for video-text generation.
* **Limitations and Future Directions:** Pretraining bottlenecks on specialized domains.
* **Relevance to Primary Paper:** Serves as NSVA’s architectural backbone.

**[33] Medress et al., AI 1977**

<https://www.google.com/search?q=https://www.sciencedirect.com/science/article/pii/0004370277900499>

* **Key Concepts and Taxonomy:** Beam search algorithm.
* **Main Contributions and Findings:** Described early approaches to constrained search in speech.
* **Limitations and Future Directions:** Requires heuristic tuning.
* **Relevance to Primary Paper:** Used for caption generation inference in NSVA.

**[34] Miech et al., CVPR 2019**

<https://www.di.ens.fr/willow/research/howto100m/>

<https://arxiv.org/abs/1906.03327>

* **Key Concepts and Taxonomy:** HowTo100M dataset.
* **Main Contributions and Findings:** Constructed large-scale instructional video dataset for video-text learning.
* **Limitations and Future Directions:** Domain limited to tutorials.
* **Relevance to Primary Paper:** Used for pretraining NSVA’s encoder-decoder models.

**[35] NBA.com**

<https://www.nba.com/stats>

* **Key Concepts and Taxonomy:** Official source of professional NBA video and statistics.
* **Main Contributions and Findings:** Provided metadata, play-by-play logs, and high-res clips.
* **Limitations and Future Directions:** Not curated for machine learning tasks.
* **Relevance to Primary Paper:** Primary source for building the NSVA dataset.

**[36] Pan et al., CVPR 2020**

<https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPR_2020/html/Pan_Spatiotemporal_Graph_for_Video_Captioning_with_Knowledge_Distillation_CVPR_2020_paper.html>

* **Key Concepts and Taxonomy:** Spatiotemporal graph modeling with knowledge distillation.
* **Main Contributions and Findings:** Encoded visual content with scene graphs for captioning.
* **Limitations and Future Directions:** High complexity for dense sports scenes.
* **Relevance to Primary Paper:** Influences NSVA’s interest in multi-object features.

**[37] Papineni et al., ACL 2002**

<https://aclanthology.org/P02-1040.pdf>

* **Key Concepts and Taxonomy:** BLEU metric for evaluating machine translation.
* **Main Contributions and Findings:** Proposed n-gram-based precision scoring.
* **Limitations and Future Directions:** Favors shorter outputs.
* **Relevance to Primary Paper:** Used as a standard metric to evaluate generated captions.

**[38] Park et al., ECCV 2020**

<https://www.google.com/search?q=https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1735_ECCV_2020_paper.php>

* **Key Concepts and Taxonomy:** Multi-sentence identity-aware video description.
* **Main Contributions and Findings:** Used person identification cues in multi-sentence video narration.
* **Limitations and Future Directions:** Not applied to professional sports.
* **Relevance to Primary Paper:** Supports the idea of identity-focused captioning in NSVA.

**[39] Pei et al., CVPR 2019**

<https://openaccess.thecvf.com/content_CVPR_2019/html/Pei_Memory-Attended_Recurrent_Network_for_Video_Captioning_CVPR_2019_paper.html>

* **Key Concepts and Taxonomy:** Memory-attended recurrent model for video captioning.
* **Main Contributions and Findings:** Used external memory module to enhance visual-textual alignment.
* **Limitations and Future Directions:** LSTM-based; limited long-term dependencies.
* **Relevance to Primary Paper:** Forms a baseline in NSVA evaluations.

**[40] Qi et al., TCSVT 2019**

<https://www.google.com/search?q=https://ieeexplore.ieee.org/document/8693892>

* **Key Concepts and Taxonomy:** Sports video captioning with motion and group modeling.
* **Main Contributions and Findings:** Modeled volleyball team interactions.
* **Limitations and Future Directions:** Dataset not public.
* **Relevance to Primary Paper:** Compared directly with NSVA’s broader domain and accessibility.

**[41] Radford et al., ICML 2021**

<https://arxiv.org/abs/2103.00020>

* **Key Concepts and Taxonomy:** CLIP—Contrastive pretraining for vision and language.
* **Main Contributions and Findings:** Unified vision-language model trained with natural supervision.
* **Limitations and Future Directions:** Weak performance on fine-grained events.
* **Relevance to Primary Paper:** Reinforces the benefit of multimodal pretraining for NSVA.

**[42] Shao et al., CVPR 2020**

<https://sdolivia.github.io/FineGym/>

<https://arxiv.org/abs/2004.06704>

* **Key Concepts and Taxonomy:** FineGym dataset.
* **Main Contributions and Findings:** Provided hierarchical labels for gymnastics actions.
* **Limitations and Future Directions:** Limited to a niche sport.
* **Relevance to Primary Paper:** Inspiration for NSVA’s multi-level action taxonomy.

**[43] Sharma et al., ACL 2018**

<https://aclanthology.org/P18-1238.pdf>

* **Key Concepts and Taxonomy:** Conceptual Captions dataset.
* **Main Contributions and Findings:** Curated image captions from alt-text with hypernymed concepts.
* **Limitations and Future Directions:** Lack of verbs and events.
* **Relevance to Primary Paper:** Captions used for baseline evaluations in NSVA.

**[44] Shi et al., MM 2020**

<https://dl.acm.org/doi/10.1145/3394171.3413809>

* **Key Concepts and Taxonomy:** Semantic alignment for narrated video.
* **Main Contributions and Findings:** Modeled temporal and semantic alignment using transformer encoders.
* **Limitations and Future Directions:** Not tailored to multi-agent sports.
* **Relevance to Primary Paper:** Offers guidance for NSVA’s decoder structure.

**[45] Soomro et al., arXiv 2012**

<https://www.crcv.ucf.edu/data/UCF101.php>

<https://arxiv.org/abs/1212.0402>

* **Key Concepts and Taxonomy:** UCF101 action recognition dataset.
* **Main Contributions and Findings:** Established a standard benchmark for evaluating video classification.
* **Limitations and Future Directions:** Amateur sports, limited annotation.
* **Relevance to Primary Paper:** Benchmarks like UCF101 highlight the professional depth of NSVA.

**[46] Strudel et al., ICCV 2021**

<https://arxiv.org/abs/2105.05633>

* **Key Concepts and Taxonomy:** Segmenter—a transformer for semantic segmentation.
* **Main Contributions and Findings:** Applied sequence-to-sequence modeling to pixel segmentation.
* **Limitations and Future Directions:** High compute cost.
* **Relevance to Primary Paper:** Aids in NSVA’s use of segmentation in player positioning.

**[47] Sun et al., ICCV 2019**

<https://arxiv.org/abs/1904.01766>

* **Key Concepts and Taxonomy:** VideoBERT model.
* **Main Contributions and Findings:** Joint video-language model using masked token prediction.
* **Limitations and Future Directions:** Works better on procedural than dynamic content.
* **Relevance to Primary Paper:** Inspired transformer fusion for captioning in NSVA.

**[48] Vaswani et al., NeurIPS 2017**

<https://arxiv.org/abs/1706.03762>

* **Key Concepts and Taxonomy:** Attention is all you need.
* **Main Contributions and Findings:** Introduced the transformer architecture.
* **Limitations and Future Directions:** Initially for NLP.
* **Relevance to Primary Paper:** Core architecture for NSVA's modeling.

**[49] Vedantam et al., CVPR 2015**

<https://arxiv.org/abs/1411.5726>

* **Key Concepts and Taxonomy:** CIDEr evaluation metric.
* **Main Contributions and Findings:** Proposed consensus-based caption evaluation.
* **Limitations and Future Directions:** Sensitive to lexical diversity.
* **Relevance to Primary Paper:** Used to evaluate NSVA caption generation.

**[50] Venugopalan et al., NAACL 2015**

<https://arxiv.org/abs/1411.4555>

* **Key Concepts and Taxonomy:** Video captioning using RNNs.
* **Main Contributions and Findings:** Early work combining CNN features with LSTM-based decoder.
* **Limitations and Future Directions:** Weak long-term modeling.
* **Relevance to Primary Paper:** Serves as historical baseline in NSVA comparisons.