|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reference** | **Model** | **Dataset** | **Significant Factor** | **Evaluation Measures** |
| Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015) | Attention-Based Recurrent Sequence Generators (ARSG) | TIMIT corpus | Incorporation of convolutional features for location-awareness. | Phoneme Error Rate (PER):  Baseline model: 18.7% (test set). |
| Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., & An, W. (2018) | Faster R-CNN (Region-based Convolutional Neural Network) | A custom dataset containing over 100,000 image frames from far-field surveillance videos at 25 construction sites collected over one year | High precision and recall rates in detecting non-hardhat use (NHU).  Adaptability to various visual conditions (weather, illumination, occlusions, worker posture). | Precision: 95.7%  Recall: 94.9% |
| Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014) | RNN Encoder-Decoder | English-French translation task from the WMT’14 dataset | Use of adaptive gating mechanisms to improve memory capacity and training efficiency. | BLEU Scores:  Baseline: 33.3  Baseline + RNN Encoder-Decoder: 33.87  Baseline + CSLM + RNN Encoder-Decoder: 34.64 |
| Ben-Alon, L., & Sacks, R. (2017) | Agent-Based Simulation (ABS) | Data collected from field interviews and observations with:  13 superintendents and trade crew leaders across four residential tower projects.  Detailed parameters such as production rates, material supply, labor assignments, and decision-making patterns. | Determines task prioritization based on perceived profitability. Influence of material and design information flows on production control. | project duration (e.g., 156–232 days across scenarios), time distribution for crew activities (e.g., 46% work with design info, 16% waiting, 9% rework in Scenario 9), and rework rates reflecting workflow disruptions |
| Waehrer, G. M., Dong, X. S., Miller, T., Haile, E., & Men, Y. (2007) | Cost Model combining direct, indirect, and quality-of-life costs for occupational injuries. | 2002 data from the Bureau of Labor Statistics (BLS) on occupational injuries and illnesses and Census of Fatal Occupational Injuries (CFOI). | Fatal and nonfatal injuries.  Direct costs (medical expenses), indirect costs (wage and household productivity losses), and quality-of-life costs. | Average Fatal Injury Cost: $4 million.  Average Nonfatal Injury Cost: $42,000 per case.  Construction accounted for 15% of private industry injury costs |
| Liu, Y., Yang, G., Qiao, S., Liu, M., Qu, L., Han, N., Wu, T., Yuan, G., & Peng, Y. (2023) | Transfer Learning Classifier (TLC) | CIFAR10, CIFAR100 (synthetically imbalanced), Caltech101, and HAM10000 datasets. | Dynamically adjusts class distributions using F1-scores.  Enhances generalization and reduces overfitting | Classification Accuracy: TLC+COT achieved 85.96% on CIFAR10 (imbalance ratio 10) and 62.05% on CIFAR100 (imbalance ratio 100). |
| Kong, T., Fang, W., Love, P. E. D., Luo, H., Xu, S., & Li, H. (2021)Click or tap here to enter text. | SiamMask: For object tracking.  Improved Social-LSTM: For trajectory prediction.  PNPoly Algorithm: For predicting unsafe behavior. | The model was tested using video data from the Wuhan Metro project captured through a real-time monitoring system equipped with CCTV cameras. | Prediction involved testing if predicted trajectories entered hazardous areas using the PNPoly algorithm. | mIOU, mAP @ 0.5 IOU, and mAP @ 0.7 IOU were used to evaluate tracking precision |
| Zhu, F., Shao, L., Xie, J., & Fang, Y. (2016) | Handcrafted Representations:  Histogram of Gradients (HOG), Histogram of Optical Flow (HOF)achieved high accuracy.  Learning-Based Representations:  CNN, 3D CNNs, Two-stream CNNs, and Hybrid models. | UCF-101,  HMDB-51,KTH,  Sports-1M. | Utilize spatial and temporal data for more robust and generalized feature extraction.  Depend on large datasets and computational resources. | IDT + FV: Achieved 85.9% accuracy on UCF-101 and 57.2% on HMDB-51.  IDT + HSV: Improved accuracy to 87.9% on UCF-101 and 61.1% on HMDB-51.  Learning-Based Approaches:  Two-Stream CNN: Scored 88.0% on UCF-101 and 63.2% on HMDB-51.  TDD + FV: Combined with Improved Dense Trajectories, it achieved 91.5% on UCF-101 and 65.9% on HMDB-51. |
| Xu, S., Sun, M., Fang, W., Chen, K., Luo, H., & Zou, P. X. W. (2023) | Bayesian-based Knowledge Tracing (BKT) model | Over 1000 on-site photos and videos.  A question bank containing 139 questions related to 35 knowledge concepts tailored to scaffolders. | Includes demographic information, job trades, learning motivations, and cognitive styles. | Accuracy of 80.9 |
| Shao, B., Hu, Z., Liu, Q., Chen, S., & He, W. (2019) | Frequency analysis, correlation coefficient analysis, and variance analysis to study fatal accident patterns in building construction | 2,348 fatal accidents reported between 2012 and 2016. | Falls account for more than 55% of fatal accidents, followed by "struck by object," "collapse," and "hoisting damage." | Fatal Accident Frequency: Distribution of accidents across time, region, and type.  Mortality Rate per GDP: Indicator to assess economic and safety implications. |
| Li, H., Wu, D., Zhang, W., & Xiao, C. (2024) | YOLO-PL: An improved, lightweight variant of YOLOv4 | SHWD (Safety Helmet Wearing Detection)  SHD,  MHD: Motorcycle helmet detection dataset images. | Optimization for detecting small safety helmets.  Improved robustness against occlusions and environmental noise. | Results on SHWD Dataset:  AP50: 94.23%.  AP75: 59.00%.  Recall: 94.26%. |
| Wei, R., Love, P. E. D., Fang, W., Luo, H., & Xu, S. (2019) | Spatial and Temporal Attention Pooling Network (ASTPN) | A custom-created video dataset from construction sites, featuring 12 pairs of surveillance videos.  Pre-trained on the publicly available iLIDS-VID database containing 600 sequences of 300 individuals. | Removes redundant video data and focuses on relevant features.  Uses Siamese network and Euclidean distance to recognize individuals. | Achieved an average accuracy of 79.2% using k-fold cross-validation.  Fine-tuned model accuracy: 75%, 75%, and 87.5% across different tests.  Non-fine-tuned model accuracy: Averaged 50%. |
| Luo, H., Liu, J., Fang, W., Love, P. E. D., Yu, Q., & Lu, Z. (2020) | YOLOv2 | Over 10,000 labeled image frames containing people and excavators collected from the construction site of the Wuhan Rail Transit System. | Key parameters like the coordinates of bounding boxes and angular changes were utilized.  Continuous video stream processing for detecting hazardous proximities between people and machinery. | People Detection: Precision = 94%, Recall = 86%  Excavator Detection: Precision = 90%, Recall = 89%  Excavator Status Recognition Accuracy: 91% average accuracy in recognizing stationary or moving states. |
| Cai, J., Zhang, Y., Yang, L., Cai, H., & Li, S. (2020) | Context-Augmented LSTM (Long Short-Term Memory) Model | Collected from three projects, including a publicly available dataset from YouTube and two in-house videotaped building projects.  Dataset included 241 trajectories, augmented to 3,640 tracks for training using sliding windows. | Avoids error accumulation typical in recursive prediction models.  Contextual Information Integration:  Neighboring entity movements.  Group relationships. | Final Displacement Error (FDE):  Recursive Model: 28.32 pixels  Position-based Seq2Seq Model: 9.00 pixels  Context-Augmented Seq2Seq Model: 8.51 pixels  Average Displacement Error (ADE):  Recursive Model: 15.41 pixels  Position-based Seq2Seq Model: 8.95 pixels  Context-Augmented Seq2Seq Model: 9.00 pixels |
| Duan, P., Zhou, J., & Goh, Y. M. (2023) | Complex Network Theory Framework-  Converts construction workers' movement trajectories into spatial–temporal networks to analyze risks and identify patterns in high-risk areas. | Residential Project Construction Site  Data collected using smartphone GPS from 67 workers, with 45 trajectories used for generating a safety risk heatmap and 22 for validation. | Movement between risk levels, measured using network complexity and tightness metrics.  Identified using risk intensity and network measures such as clustering coefficient and modularity. | Complexity Metrics:  Average Strength: Cluster 0 = 154.73, Cluster 1 = 319.37, Cluster 2 = 915.11.  Entropy: Cluster 0 = 0.98, Cluster 1 = 0.91, Cluster 2 = 1.00.  Tightness Metrics:  Modularity: Cluster 0 = 0.76, Cluster 1 = 0.48, Cluster 2 = 0.34.  Transitivity: Cluster 0 = 0.34, Cluster 1 = 0.37, Cluster 2 = 0.29. |
| Xu, M., Di, Y., Ding, H., Zhu, Z., Chen, X., & Yang, H. (2023) | Attentive Graph Neural Process (AGNP) | Real-world traffic speed data from Xuancheng City, Anhui Province, China, covering April 1–27, 2019. The dataset includes 1,196 lanes and 1,563 valid links with traffic speeds aggregated at 5-minute intervals. | Traffic state features such as lane width, free-flow speed, green time ratio, and precipitation.  Graph-structured data with spatial-temporal dependencies. | Mean Absolute Error (MAE):  Missing 10%: 4.013  Missing 40%: 4.132  Missing 70%: 4.236  Root Mean Square Error (RMSE):  Missing 10%: 6.190  Missing 40%: 6.328  Missing 70%: 6.368 |
| Schroff, F., Kalenichenko, D., & Philbin, J. (2015) | FaceNet – A deep convolutional network that learns a compact Euclidean embedding for face recognition, verification, and clustering. | Labeled Faces in the Wild (LFW),  YouTube Faces Database (YTF),  Personal Photos Dataset: Verified clean labels with manually curated data (~12k images).  Large-Scale Training Dataset: Contains 200M face thumbnails from ~8M identities. | Separates embeddings of different identities by a margin while bringing similar embeddings closer. | LFW Dataset Accuracy: 99.63%  YTF Dataset Accuracy: 95.12% |
| Ma, L., Lu, Z., Shang, L., & Li, H. (2015) | Multimodal Convolutional Neural Networks (m-CNN) | Flickr8K: 8,000 images with 5 sentences per image.  Flickr30K: 31,783 images with 5 sentences per image.  Microsoft COCO: Over 110,000 images with 5 sentences per image. | Combines image CNN and matching CNN to jointly learn image and sentence embeddings. | R@K (Recall at K): Measures the fraction of correct results in the top K results.  Med r (Median Rank): Median rank of the first correctly retrieved result.  Results on Microsoft COCO:  Sentence Retrieval:  R@1: 42.8  R@5: 73.1  R@10: 84.1  Med r: 2  Image Retrieval:  R@1: 32.6  R@5: 68.6  R@10: 82.8  Med r: 3 |
| Ma, Z., Lu, Y., & Foster, D. (2015) | Augmented Approximate Gradient (AppGrad) | Mediamill: Annotated video dataset with 30,000 samples,  MNIST Handwritten text dataset,  Penn Tree Bank: Text dataset for word co-occurrence with 500,000 samples,  URL Reputation Dataset: Contains 2 million samples and is highly dimensional (100,000+ features). | The algorithm avoids the costly whitening step and leverages a memory-efficient design with an optimal storage complexity. | Proportion of Correlations Captured (PCC)  Results:  Mediamill: PCC exceeded 95% for both AppGrad and its stochastic variant.  MNIST: Similar performance above 90%.  Penn Tree Bank: Achieved near-complete PCC (close to 100%) . |
| Wang, L., Li, Y., & Lazebnik, S. (2015) | A two-branch deep neural network with multiple layers and nonlinearities | Flickr30K: Contains 31,783 images and 5 corresponding descriptive sentences each,  MSCOCO Dataset. | The proposed method combines bi-directional ranking loss (to align images with text) with structure-preserving constraints (to cluster semantically similar samples within each modality), improving embedding quality and retrieval performance. | Results on Flickr30K:  R@1: 40.3 (image-to-sentence) and 29.7 (sentence-to-image).  R@10: 79.9 (image-to-sentence) and 72.1 (sentence-to-image).  Results on MSCOCO (1000 test images):  R@1: 50.1 (image-to-sentence) and 39.6 (sentence-to-image).  R@10: 89.2 (image-to-sentence) and 86.9 (sentence-to-image). |
| Luong, M.-T., Pham, H., & Manning, C. D. (2015) | Attention-based Neural Machine Translation (NMT) | WMT’14 and WMT’15 English-German translation datasets | The integration of attention mechanisms (global and local) significantly improves alignment and translation quality. | WMT’14 :  Baseline: 14.0 BLEU.  Global Attention: 18.1 BLEU (+4.1).  Local Predictive Attention: 19.0 BLEU (+0.9).  Ensemble Model: 23.0 BLEU (state-of-the-art).  WMT’15 :  Ensemble Model: 25.9 BLEU |
| Vendrov, I., Kiros, R., Fidler, S., & Urtasun, R. (2015) | Normal caption-image retrieval | Hypernym Prediction: WordNet hierarchy (82192 concepts, 838073 edges).  Caption-Image Retrieval: Microsoft COCO (113,287 training images, 5000 validation, 5000 test).  Textual Entailment: SNLI | Use of partial order structure in embedding spaces. | Hypernym Prediction:  Accuracy: 90.6%  Caption-Image Retrieval:  Recall@1 = 46.7% and Med r = 2  Textual Entailment:  Accuracy: 88.6% |
| Reed, S., Akata, Z., Schiele, B., & Lee, H. (2016) | The study uses Deep Symmetric Structured Joint Embedding (DS-SJE) and Deep Asymmetric Structured Joint Embedding (DA-SJE) | Caltech-UCSD Birds 200-2011 (CUB),Flowers Dataset. | Introduction of a deep structured joint embedding framework for jointly embedding images and text. | CUB Dataset (Zero-Shot Classification Accuracy):  DS-SJE: Word-CNN-RNN achieved 56.8%  DA-SJE: Word-CNN-RNN achieved 54.3%.  DS-SJE: Word-CNN-RNN achieved 48.7%.  Oxford Flowers Dataset:  Zero-Shot Classification: Word-CNN-RNN achieved 65.6%. |
| Nam, H., Ha, J.-W., & Kim, J. (2016) | Dual Attention Networks (DANs)  Two Variants:  Reasoning-DAN (r-DAN) for multimodal reasoning,  Matching-DAN (m-DAN) for multimodal matching | Visual Question Answering (VQA) Dataset,Flickr30K Dataset. | Simultaneously focus on specific image regions and words in text. | For Visual Question Answering (VQA):  Accuracy-64.3% (open-ended) and 69.1% (multiple-choice)  For Image-Text Matching (Flickr30K):  Image-to-Text: R@1 = 55.0%, R@5 = 81.8%, R@10 = 89.0%, MR = 1.  Text-to-Image: R@1 = 39.4%, R@5 = 69.2%, R@10 = 79.1%, MR = 2. |
| Huang, Y., Wang, W., & Wang, L. (2016) | Selective Multimodal Long Short-Term Memory network (sm-LSTM) | Flickr30k Dataset,Microsoft COCO Dataset. | Focuses on pairwise instance-aware saliency maps to predict which image regions and sentence phrases to attend to. | Flickr30k:  Image Annotation:  R@1 = 42.4, R@10 = 79.9.  Image Retrieval:  R@1 = 28.2, R@10 = 68.4.  Microsoft COCO:  Image Annotation:  R@1 = 52.4, R@10 = 90.8.  Image Retrieval:  R@1 = 38.6, R@10 = 84.6. |
| Wang, L., Li, Y., Huang, J., & Lazebnik, S. (2017). | Two-Branch Neural Networks:  Embedding Network: Maps images and text into a joint latent space  Similarity Network: Aggregates features using an element-wise product | Flickr30K,  Microsoft COCO (MSCOCO). | Embedding Network:  Uses triplet sampling and neighborhood constraints to learn semantic similarity.  Similarity Network:  Simplifies computation by directly predicting similarity scores for image-text pairs. | Flickr30K:  Image-to-Sentence :  Embedding Network: R@1 = 43.2, R@10 = 79.8.  Sentence-to-Image:  Embedding Network: R@1 = 31.7, R@10 = 72.4.  MSCOCO:  Image-to-Sentence :  Embedding Network: R@1 = 54.9, R@10 = 92.2.  Sentence-to-Image :  Embedding : R@1 = 43.3, R@10 = 87.5. |
| Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017) | VSE++ (Visual-Semantic Embeddings with hard negatives) | MS-COCO and Flickr30K datasets | Incorporating hard negatives in the ranking loss to enhance retrieval performance. | Results on MS-COCO (1K test images):  Caption Retrieval: R@1 = 64.6%, R@5 = 90.0%, R@10 = 95.7%.  Image Retrieval: R@1 = 52.0%, R@5 = 84.3%, R@10 = 92.0%.  Results on Flickr30K:  Caption Retrieval: R@1 = 52.9%, R@5 = 80.5%, R@10 = 87.2%.  Image Retrieval: R@1 = 39.6%, R@5 = 70.1%, R@10 = 79.5%. |
| Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2017) | Bottom-Up and Top-Down Attention Model | MS-COCO 2014 captions dataset,  Visual Question Answering (VQA)- VQA v2.0 dataset | Integration of bottom-up attention (via Faster R-CNN) with top-down attention to focus on salient image regions, improving interpretability and performance across tasks. | Image Captioning:  MS-COCO Karpathy split:  BLEU-4 = 36.3, METEOR = 27.7, CIDEr = 120.1, SPICE = 21.4.  MS-COCO test server:  BLEU-4 = 36.9, METEOR = 27.6, CIDEr = 117.9, SPICE = 21.5.  Visual Question Answering (VQA):  VQA v2.0 test-standard:  Overall Accuracy = 70.34%, Yes/No = 86.6%, Number = 48.64%, Other = 61.15%. |
| Li, S., Xiao, T., Li, H., Yang, W., & Wang, X. (2017) | Identity-Aware Textual-Visual Matching with Latent Co-Attention (Two-Stage CNN-LSTM Framework). | CUHK-PEDES,  Caltech-UCSD Birds (CUB),  Oxford-102 Flowers | CNN-LSTM network utilizing a Cross-Modal Cross-Entropy (CMCE) loss to minimize intra-identity discrepancies and maximize inter-identity differences. | CUHK-PEDES:  Top-1 Accuracy = 25.94%  Top-10 Accuracy = 60.48%  CUB Dataset:  Image-to-Text (Top-1 Accuracy) = 61.5%, Text-to-Image (AP@50) = 57.6%  Oxford-102 Flowers:  Image-to-Text (Top-1 Accuracy) = 68.4%, Text-to-Image (AP@50) = 70.1% |
| Huang, Y., Wu, Q., & Wang, L. (2017) | Semantic-Enhanced Image and Sentence Matching Model | Flickr30k,  MSCOCO | Semantic concepts extraction through multi-regional multi-label CNN.  Semantic order learning with a context-gated sentence generation scheme. | Accuracy:  Flickr30k: 62.3  MSCOCO: 81.8 |
| Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2018) | RU-Net (Recurrent U-Net),  R2U-Net (Recurrent Residual U-Net) | DRIVE (Retina Blood Vessel Segmentation)  STARE (Retina Blood Vessel Segmentation)  CHASE\_DB1 (Retina Blood Vessel Segmentation)  Kaggle Skin Cancer Segmentation Dataset  LUNA (Lung Segmentation Dataset) | Incorporation of recurrent convolutional layers for feature accumulation.  Residual units to address vanishing gradient issues in deep networks.  Efficient segmentation with minimal network parameters. | Dice Coefficient (DC)  Area Under the Curve (AUC)  DRIVE Dataset: DC: 0.9784, AUC: 0.9784  STARE Dataset: DC: 0.9914, AUC: 0.9914  CHASE\_DB1 Dataset: DC: 0.9815, AUC: 0.9815  Skin Cancer: DC: 0.8616, AUC: 0.9419  Lung Segmentation: DC: 0.9918, AUC: 0.9889 |
| Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. ( 2018) | Stacked Cross Attention Network (SCAN) | MS-COCO,  Flickr30K | Novel Stacked Cross Attention Mechanism to discover comprehensive visual-semantic alignments.  Incorporates Bottom-Up Attention using Faster R-CNN for detecting salient image regions. | Flickr30K Dataset:  Sentence Retrieval (R@1): 67.9  Image Retrieval (R@1): 45.8  MS-COCO Dataset :  Sentence Retrieval (R@1): 72.7  Image Retrieval (R@1): 58.8 |
| Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and  Yun Fu(2017) | Visual Semantic Reasoning Network (VSRN),  Graph Convolutional Networks (GCN) | MS-COCO,  Flickr30K | Combines matching and sentence generation for learning image-text alignments.  Uses GCN to build semantic relationships between image regions. | MS-COCO:  Caption Retrieval: R@1 = 53.0, R@5 = 81.1, R@10 = 89.4.  Image Retrieval: R@1 = 40.5, R@5 = 70.6, R@10 = 81.1.  Flickr30K Test Set:  Caption Retrieval: R@1 = 71.3, R@5 = 90.6, R@10 = 96.0.  Image Retrieval: R@1 = 54.7, R@5 = 81.8, R@10 = 88.2. |
| Weili Fang, Lieyun Ding, Hanbin Luo, and Peter E.D. Love(2020) | Faster R-CNN,region proposal network (RPN) and classification layers for object detection. | Custom Dataset:  Created from images and videos of construction sites in Wuhan, China. | Faster R-CNN extracts features from worker images to identify their presence.  The CNN performs specific safety harness detection by focusing on cropped regions. | Worker Detection (Faster R-CNN):  Precision: 99%  Recall: 95%  Harness Detection (CNN):  Precision: 80%  Recall: 98% |
| Lieyun Ding, Weili Fang, Hanbin Luo, Peter E.D. Love, Botao Zhong, and Xi Ouyang (2019) | Hybrid Deep Learning Model (CNN + LSTM) | Custom Dataset:  Video recordings of construction workers performing actions related to ladder climbing. | Combines spatial feature extraction (via CNN) with temporal sequence analysis (via LSTM) for robust action detection. | CNN + LSTM accuracy: 97%. |
| Fang, W., Zhong, B., Zhao, N., Love, P. E. D., Luo, H., Xue, J., & Xu, S. (2019). | Mask R-CNN | Custom Dataset: 2018 images of construction sites in Wuhan, China, capturing individuals walking on and around structural supports.  Pre-trained on Microsoft COCO dataset for feature extraction. | Mask R-CNN achieved high precision (75%) and recall (90%) for detecting unsafe behaviors, specifically people traversing structural supports. | Precision: 75% for unsafe behavior recognition  Recall: 90% for unsafe behavior recognition |
| Choudhry, R. M. (2014) | Behavior-Based Safety (BBS) Management System | Safety performance data collected over nine weeks from three construction projects in Hong Kong. | Personal protective equipment (PPE), housekeeping, access to heights, plant and equipment, and scaffolding. | Housekeeping: From 83.7% to 92.9%  Scaffolding: From 83% to 93.3% |
| Huang, F., Zhang, X., Zhao, Z., & Li, Z. (2019) | Bi-Directional Spatial-Semantic Attention Network (BSSAN)  Combines Word-to-Regions (W2R) and Object-to-Words (O2W) attention mechanisms. | Flickr30K,  MSCOCO | Exploits fine-grained, bi-directional correlations between image regions and text semantics for improved matching. | Flickr30K Results:  Image-to-Text Retrieval: R@1 = 48.3%, R@5 = 77.0%, MedR = 2  Text-to-Image Retrieval: R@1 = 36.2%, R@5 = 65.0%, MedR = 3  MSCOCO Results:  Image-to-Text Retrieval: R@1 = 64.2%, R@5 = 88.4%, MedR = 2  Text-to-Image Retrieval: R@1 = 50.8%, R@5 = 79.2%, MedR = 3 |
| Fang, W., Love, P. E. D., Luo, H., & Ding, L. (2020) | Convolutional Neural Networks (CNNs) and Deep Learning Integration for Behavior-Based Safety (BBS) Programs | MSCOCO | Incorporation of computer vision and deep learning to improve BBS in construction.  Focus on safety performance through observation, feedback, and prediction of unsafe behaviors. | Accuracy :  88.9% |
| Xu, X., Wang, T., Yang, Y., Zuo, L., Shen, F., & Shen, H. T. (2020) | Cross-Modal Attention with Semantic Consistency (CASC) | Flickr30k,  MSCOCO | CASC combines local alignment (cross-modal attention) and global semantic consistency (multilabel prediction). | MSCOCO :  Image Retrieval: R@1: 68.5%, R@5: 90.3%, R@10: 95.4%.  Sentence Retrieval: R@1: 52.7%, R@5: 86.2%, R@10: 94.8%.  Flickr30k:  Image Retrieval: R@1: 56.6%, R@5: 85.8%, R@10: 93.8%.  Sentence Retrieval: R@1: 43.2%, R@5: 73.9%, R@10: 85.9%. |
| Yan, F., & Mikolajczyk, K. (2015) | Deep Canonical Correlation Analysis (DCCA) | Flickr8K,  Flickr30K,  IAPR TC-12 | The implementation of DCCA in a GPU environment, addressing computational challenges like overfitting and high-dimensional feature representation, leading to state-of-the-art performance in image-caption matching tasks. | Flickr8K:  Image Annotation R@1 = 17.9, R@5 = 40.3, R@10 = 51.9,  MR = 9.  Image Retrieval :  R@1 = 12.7, R@5 = 31.2, R@10 = 44.1, MR = 13.  Flickr30K:  Image Annotation : R@1 = 16.7, R@5 = 39.3, R@10 = 52.9, MR = 8.  Image Retrieval : R@1 = 12.6, R@5 = 31.0, R@10 = 43.0, MR = 15.  IAPR TC-12:  Image Annotation Precision: P@1 = 0.302, P@5 = 0.114, MAP = 0.426.  Image Retrieval Precision: P@1 = 0.295, P@5 = 0.120, MAP = 0.415. |
| Karpathy, A., & Fei-Fei, L. (2015) | Convolutional Neural Networks (CNNs) for image regions and Bidirectional Recurrent Neural Networks (BRNNs) for sentence segments | Flickr8K  Flickr30K  MSCOCO | The introduction of a novel alignment approach that connects image regions to sentence fragments, enabling the generation of image descriptions at both full-image and region-specific levels. The model integrates structured objectives and multimodal embeddings to improve inter-modal correspondences. | Image-Sentence Alignment (Flickr30K):  Image Annotation (R@1): 22.2 (BRNN model)  Image Retrieval (R@1): 15.2 (BRNN model)  Median Rank: 4.8 (BRNN model)  Generated Descriptions (MSCOCO):  BLEU Scores: B-1 = 62.5, B-2 = 45.0, B-3 = 32.1, B-4 = 23.0  METEOR: 19.5  CIDEr: 66.0  Region-Level Descriptions (New Dataset):  BLEU-4 = 14.8 (Region-level RNN) |