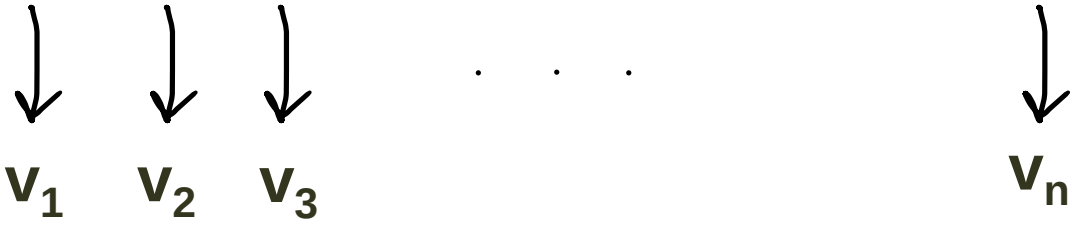
TIMESFORMER

Is Space-Time Attention All You Need for Video Understanding?

Alessia Lombarda – Progetto di Visione Artificiale

IL MECCANISMO DI ATTENZIONE

‘Noah can be annoying but she is a great cat’



$v_1 \cdot v_1 = s_{11}$

$v_1 \cdot v_2 = s_{12}$

$v_1 \cdot v_3 = s_{13}$

…

$v_1 \cdot v_n = s_{1n}$

normalizzazione

w_{11}

w_{12}

w_{13}

…

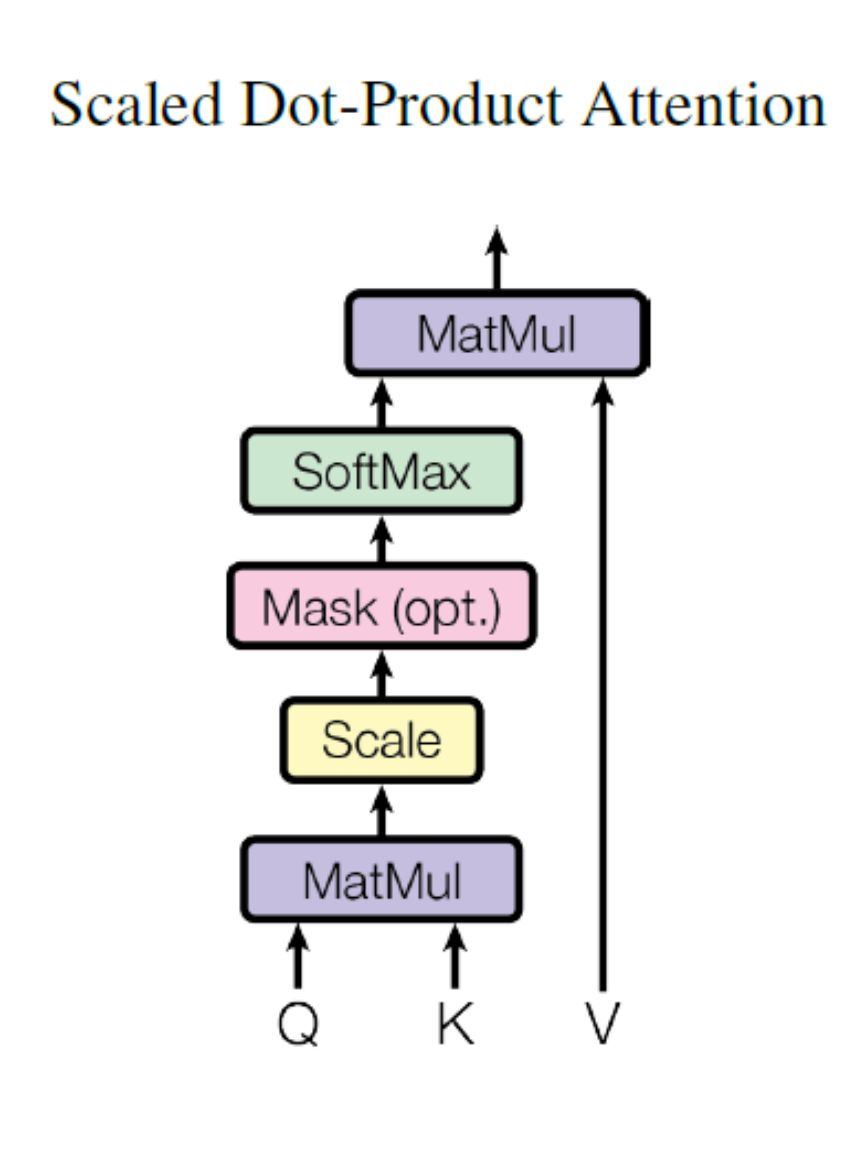
w_{1n}

somma pesata

$$y_1 = \sum_{j=0}^{n} w_{1j} \cdot vj$$

$$y_2 = \sum_{j=0}^{n} w_{2j} \cdot vj$$

$$y_n = \sum_{j=0}^{n} w_{nj} \cdot vj$$

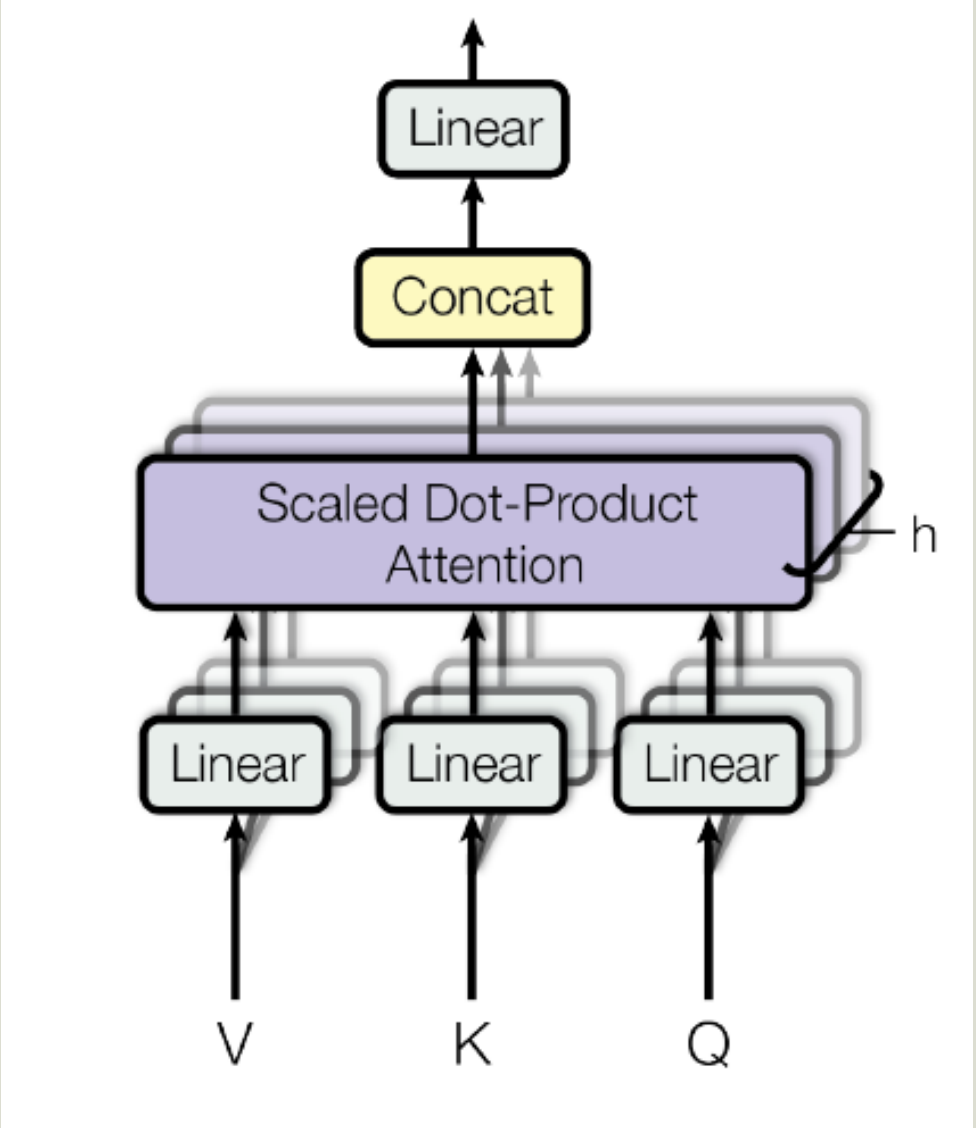


] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Neural Information Processing Systems (NIPS), 2017

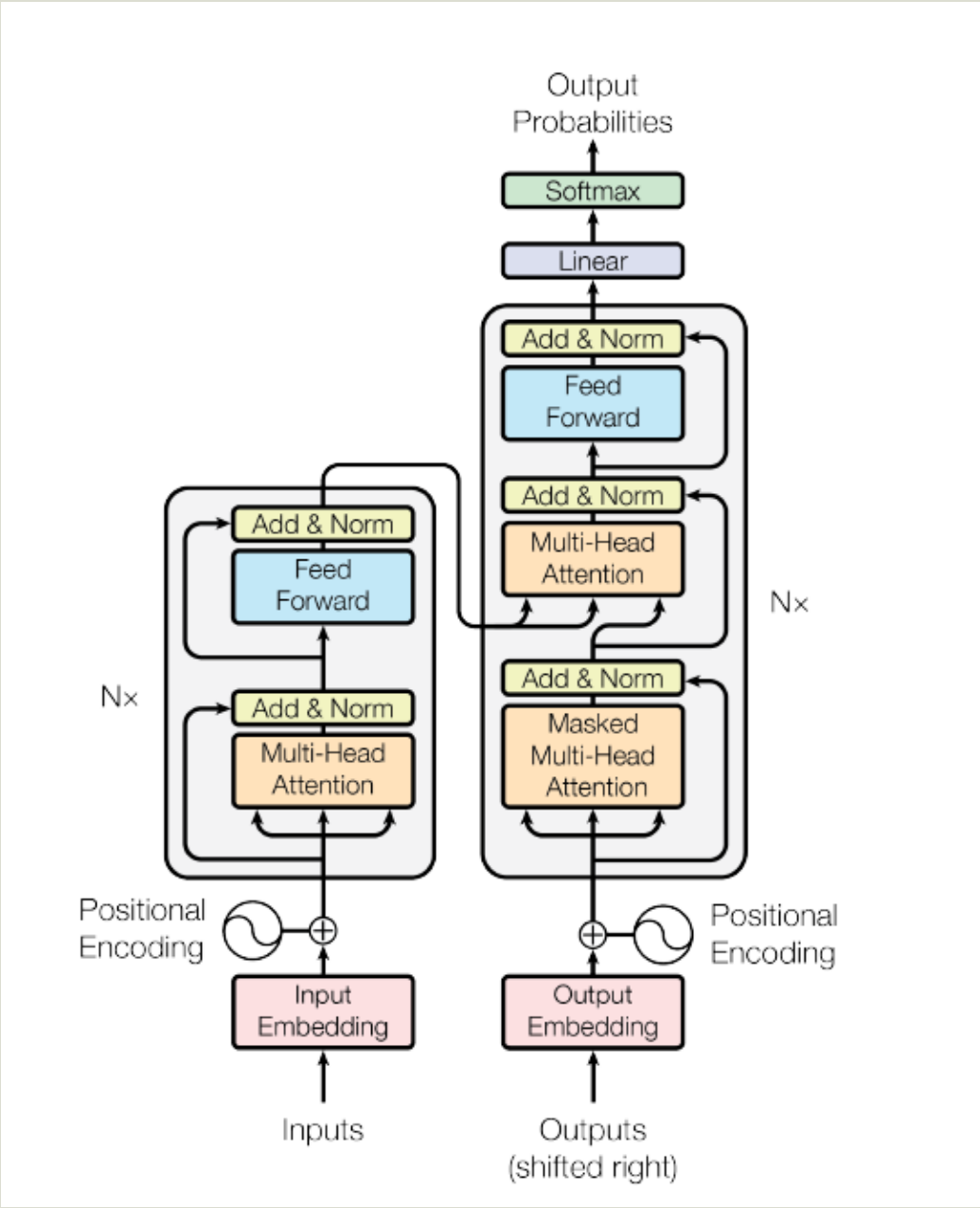
ATTENTION MECHANISMS

- Self-attention schema
- Query, keys, values: parametri
- Apprendimento

MULTI-HEAD ATTENTION



] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Neural Information Processing Systems (NIPS), 2017

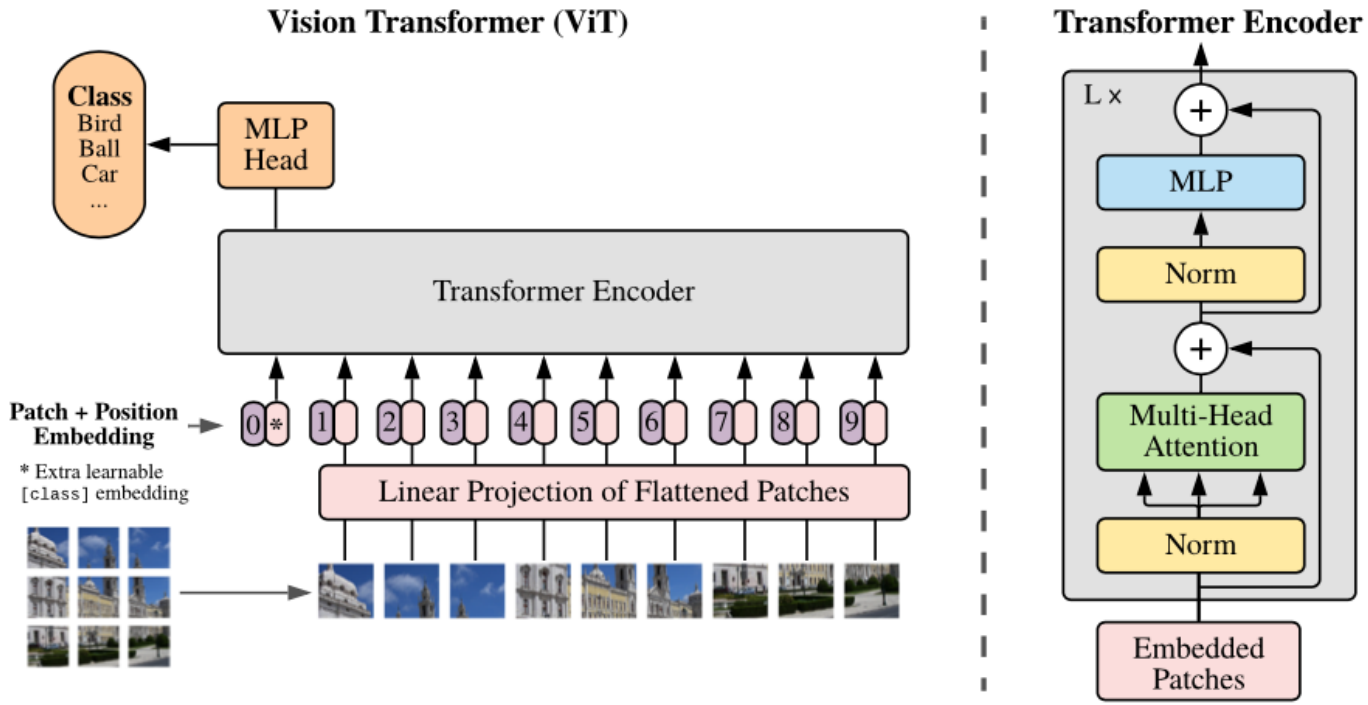


] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Neural Information Processing Systems (NIPS), 2017

STRUTTURA DEL TRANSFORMER

- Encoder e decoder: traduzione di testi
- Positional encoding
- Nx e h: iperparametri

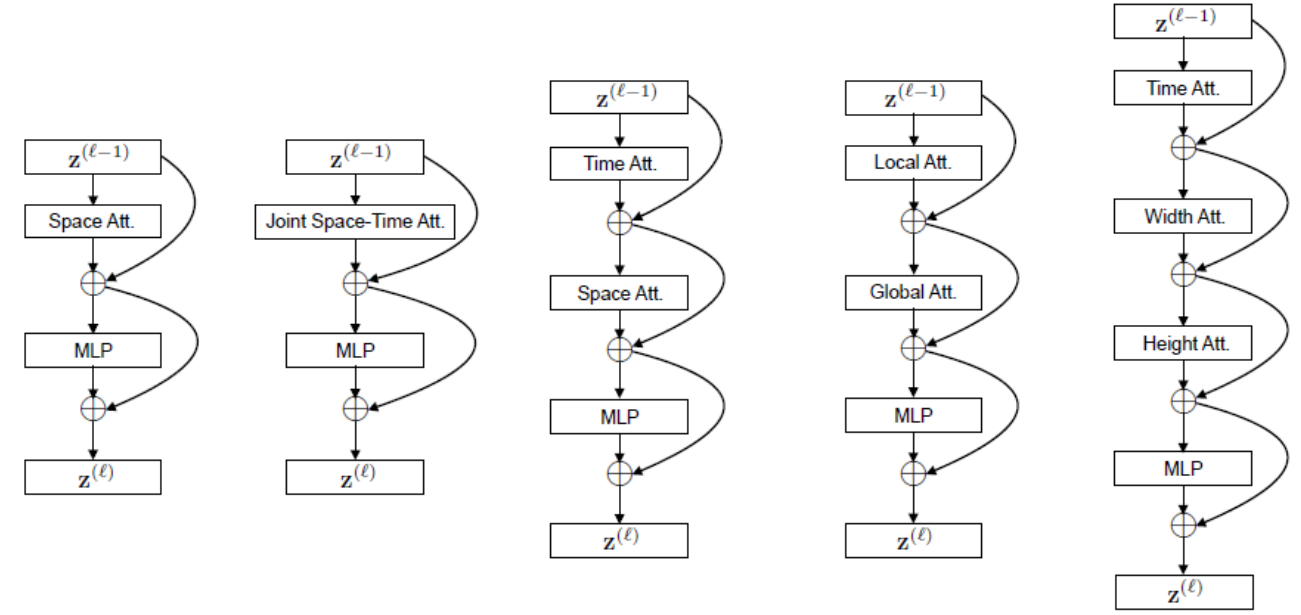
VISION TRANSFORMER



- ViT utilizzato per classificare immagini

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations. Jun Feng, Minlie Huang, Yang Yang, and Xiaoyan Zhu. 2016.

TIMESFORMER: ATTENTION MECHANISMS



- Pipeline
- Uso di diversi schemi di self-attention

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095, 2021

TIMESFORMER VS 3D-CNN

- Minori bias iniziali quindi maggior numero di famiglie di funzioni rappresentabili, maggiore capacità di apprendimento
- Catturano dipendenze spaziotemporali globali e locali
- Modelli più interpretabili
- Più veloci nel training e nel test
- Applicabili a video più lunghi

ESPERIMENTI EFFETTUATI - DATASET

UCF11 Youtube Action Data Set[1]	11 categorie di azioni, 1160 video
	Per ogni categoria i video son raggruppati in 25 gruppi con più di 4 clip per gruppo. I video nello stesso gruppo possono condividere alcune caratteristiche comuni come stesso attore, sfondo simile, simile punto di osservazione
	Grandi variazioni nel movimento della camera, posa e scala degli oggetti, punto di osservazione, sfondo, condizioni di illuminazione (così come in UCF101)
UCF101[2]	101 categorie di azioni, 13320 video
	Per ogni categoria i video son raggruppati in 25 gruppi con 4-7 clip per gruppo.I video nello stesso gruppo possono condividere alcune caratteristiche comuni come stesso attore, sfondo simile, simile punto di osservazione
	Le categorie di azioni possono essere suddivise in 5 gruppi: interazione uomo-oggetto, movimento del corpo, interazione uomo-uomo, suonare uno strumento musicale, sport

[1] https://www.crcv.ucf.edu/data/UCF_YouTube_Action.php, [2] https://www.crcv.ucf.edu/data/UCF101.php

ESPERIMENTI EFFETTUATI – PREPROCESSING E TRAINING DEL MODELLO

- Utilizzo di 1 GPU (8 nel paper) e Batch size ridotta a 4 per la JointST attention
- Preprocessing
- Training:
  - 15% test set, 85% train set
  - K-Fold CrossValidation (k=5) per Train e Validation set
  - 7 epoche massime

RISULTATI E CONSIDERAZIONI

Dataset	DivST		SpaceOnly		JointST	
	Top1	Top5	Top1	Top5	Top1	Top5
UCF11	90.61	91.08	90.33	91.08	89.48	91.08
UCF101	84.08	84.69	83.58	84.63	83.58	84.64

Dataset	DivST	
	Top1	Top5
K400	77.9	99.32
K600	79.1	94.4
SSv2	59.1	85.6
HowTo100M	56.8	-

- DivST ottiene le migliori performance
- JointST richiede un tempo di esecuzione maggiore
- La performance peggiora all'aumentare della dimensione del dataset
- In termini di accuratezza non ci sono grandi differenze tra i meccanismi di attenzione considerati

https://github.com/facebookresearch/TimeSformer

CONCLUSIONE E POSSIBILI SVILUPPI

- Sviluppi:
  - Sperimentare Transfer Learning a partire dai modelli pretrainati
  - Sperimentare su video più complessi (più lunghi o con maggiore risoluzione) e dataset più grandi, maggior numero di epoche
  - Ipotizzare possibili variazioni ai meccanismi di attenzione proposti