Matching & Ranking and Description Generation Techniques for Video-To-Text

Grethel Coello, Jesús Pérez, Nicolás Bravo and Benjamín Bustos
Millennium Institute Foundational Research on Data
DCC, Universidad de Chile

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

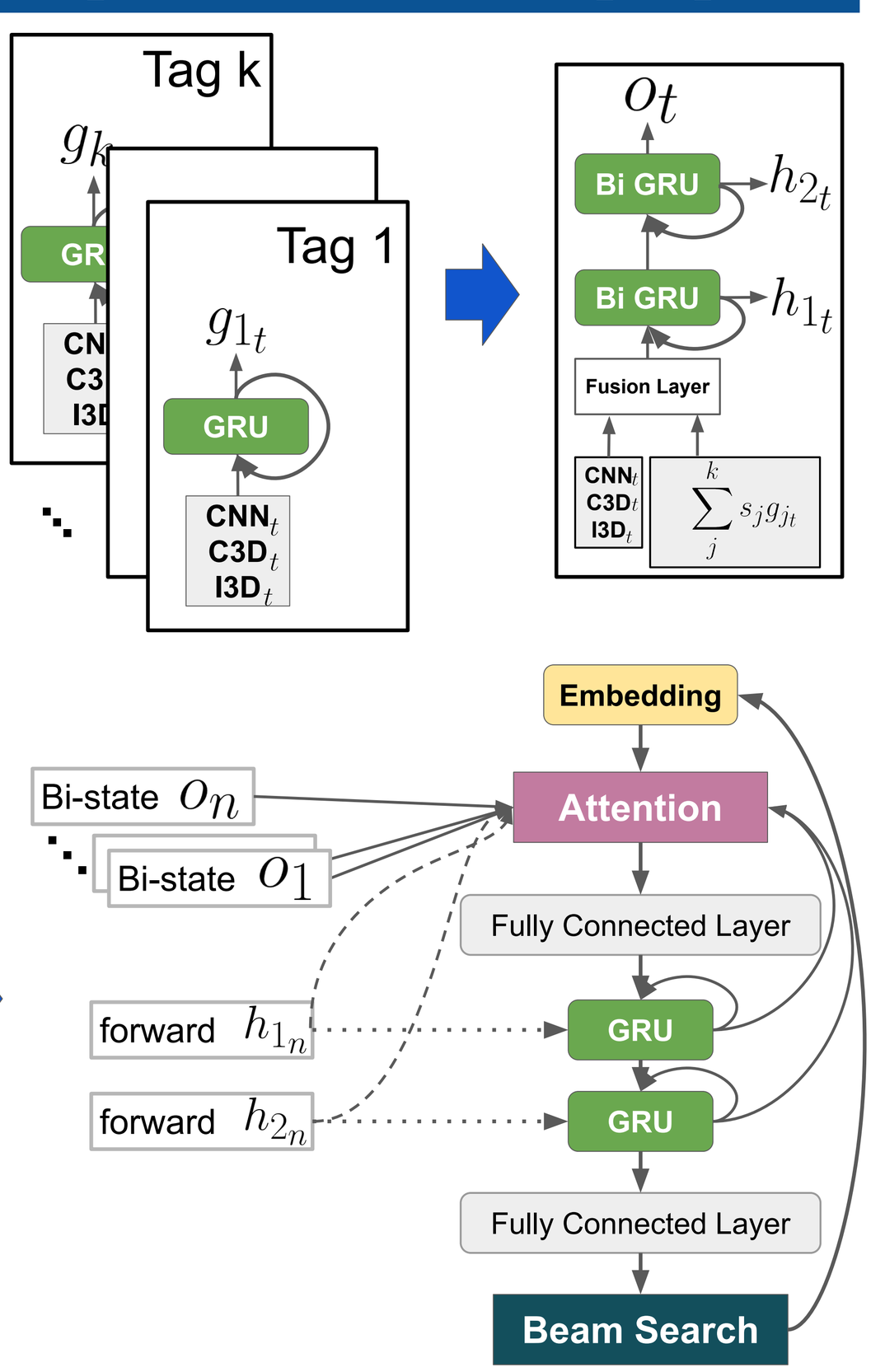
Millennium Institute Foundational Research on data

Motivation

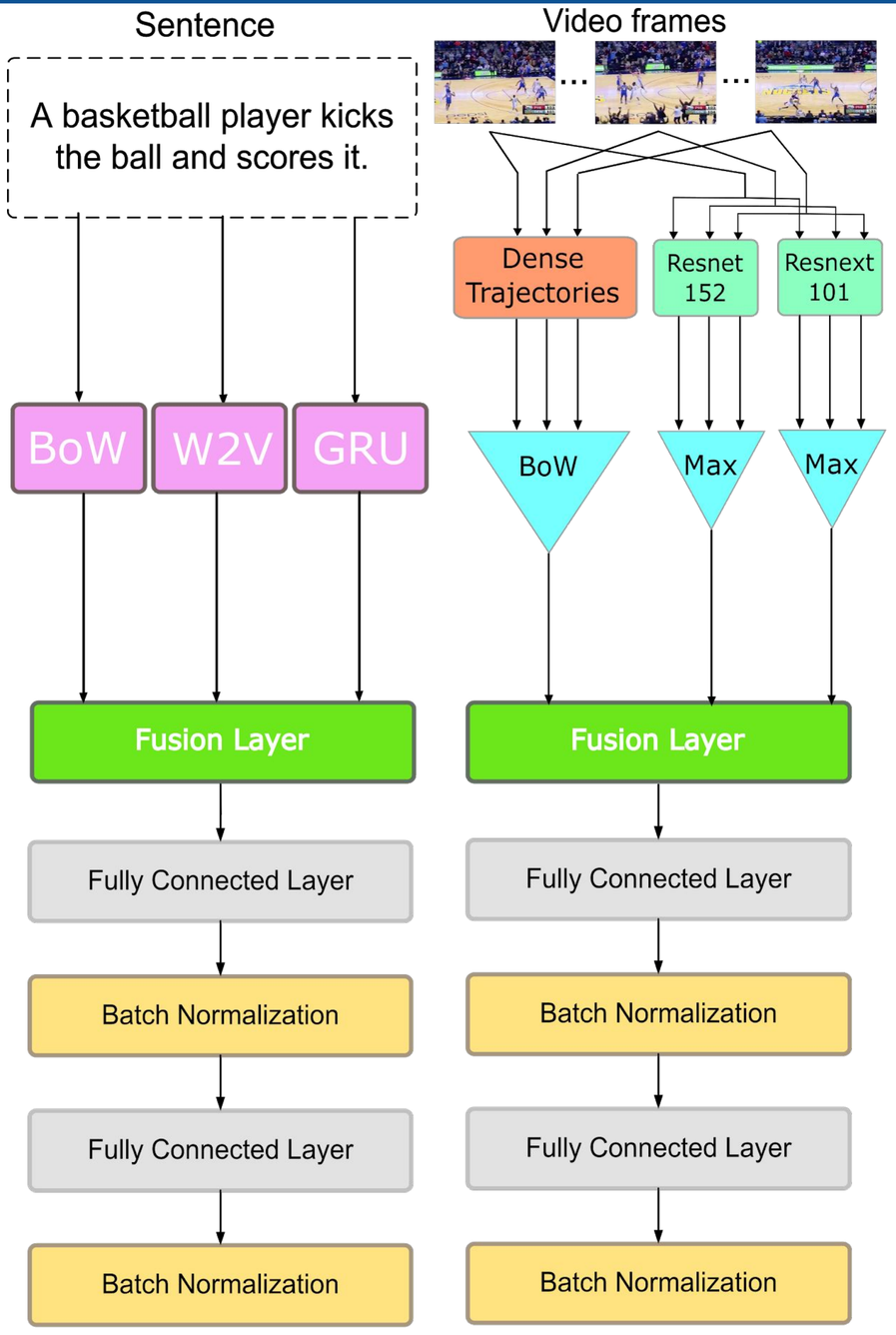
Automatic annotation of videos using natural language text descriptions has been a long-standing application of computer vision and natural language processing.
It could be useful for **video summarization** in the form of natural language, facilitating the **search and browsing** of video archives using such descriptions, **describing videos to the visually impaired** and **Human-Robot interactions**

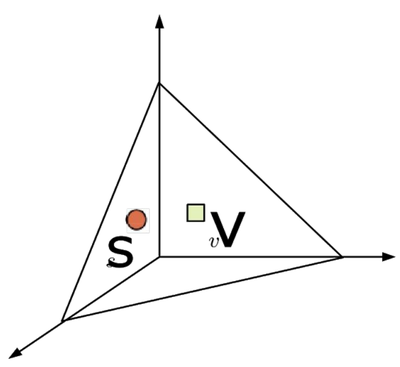
Tasks

Description Generation proposal



Matching & Ranking proposal

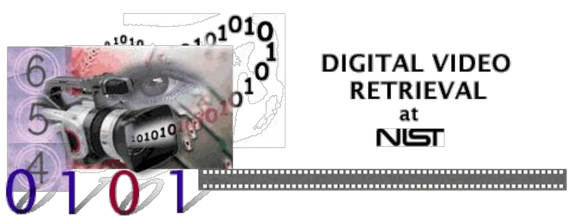




Training with
Triplet Ranking Loss
$\max_{s'}[\alpha + d(v,s) - d(v,s')]$
using Cosine Similarity
as distance function d

Datasets

MSVD: 1970 Youtube videos, +70k video-sentence pairs.
MSR-VTT: 10k Youtube videos, +150k video-sentence pairs.
TGIF: 100k Tumbrl GIFs with its corresponding sentence.



Participating in NIST TRECVID 2019, in conjunction with ORAND S.A.