Motivation

• Concept-based approach
  1. Numerous concepts.
  2. Exponentially increasing number of concept relations
     → Impossible to prepare models that detect all the concepts and relations.

• Embedding approach
  Map visual features of a shot and textual features of a topic into a common space.
  → Their similarity can be directly computed.

Overview

Coarse-grained embedding (VSE++)
• Map frames in a shot and the text description of a topic into a common space.
• It is useful to evaluate the overall relevance of a shot to the topic.

Results

i. C_D_kindai_kobe.19_1 is an ensemble of VSE++M, VSE++F and SCAN
   VSE++M trained on MS-COCO, VSE++F trained on Flickr 30k, SCAN trained on MS-COCO
ii. C_D_kindai_kobe.19_2 is an ensemble of VSE++M, VSE++F and SCAN where the
    feature of each region is L2-normalised
iii. C_D_kindai_kobe.19_3 is comprised only of SCAN
iv. C_D_kindai_kobe.19_4 is an ensemble of VSE++M and VSE++F
v. N_D_kindai_kobe.19_5 is comprised only of SCAN with L2-normalization

Analysis on failure cases

• For topics involving phrases
  This caption should be split into "black man" and "singing", but it is split into "black", "man" and "singing".
  → Many retrieved shots show black background or clothes and not black man who singing

• For topics involving words which are not included in dataset’s vocabulary
  "Drone" exists in neither MS-COCO’s nor Flickr30k’s vocabulary.
  → Many retrieved shots show a flying bird and a person who is parachuting or hang-gliding

Future work

• Develop a fast matching method that efficiently filters out many irrelevant shots by considering the structure of a topic’s text description and the relation among regions.
• Exploit Web images annotated with words which are not included in captions of dataset’s vocabulary.