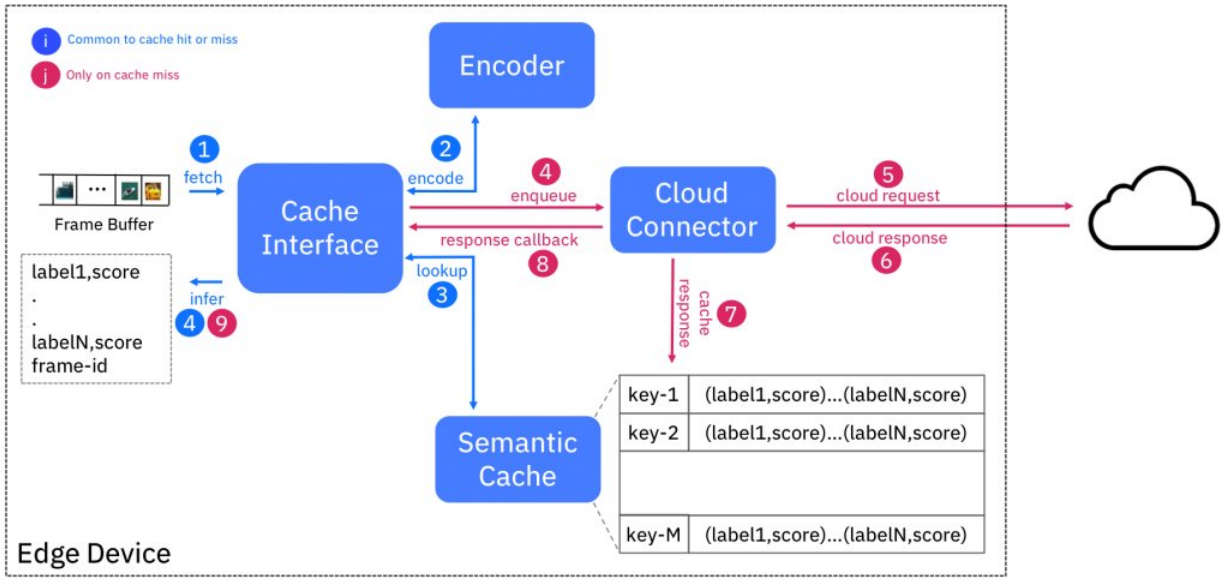
Semantic cache for AI-enabled image analysis

August 28 2018, by Srikumar Venugopal



Block diagram of semantic cache service. Credit: IBM

The availability of high-resolution, inexpensive sensors has exponentially increased the amount of data being produced, which could overwhelm the existing Internet. This has led to the need for computing capacity to process the data close to where it is generated, at the edges of the network, in lieu of sending it to cloud datacenters. Edge computing, as this is known, not only reduces the strain on bandwidth but also reduces latency of obtaining intelligence from raw data. However, availability of resources at the edge is limited due to the lack of economies of scale that

make cloud infrastructure cost-effective to manage and offer.

The potential of edge computing is nowhere more obvious than with video analytics. High-definition (1080p) video cameras are becoming commonplace in domains such as surveillance and, depending on the frame rate and data compression, can produce 4-12 megabits of data per second. Newer 4K resolution cameras produce raw data on the order of gigabits per second. The requirement for real-time insights into such video streams is driving the use of AI techniques such as deep neural networks for tasks including classification, object detection and extraction, and anomaly detection.

In our Hot Edge 2018 Conference Paper "Shadow Puppets: Cloud-level Accurate AI Inference at the Speed and Economy of Edge," our team at IBM Research – Ireland experimentally evaluated the performance of one such AI workload, object classification, using commercially available cloud-hosted services. The best result we could secure was a classification output of 2 frames per second which is far below the standard video production rate of 24 frames per second. Executing a similar experiment on a representative edge device (NVIDIA Jetson TK1) achieved the latency requirements but used up most of the resources available on the device in this process.

We break this duality by proposing the Semantic Cache, an approach that combines the low latency of edge deployments with the near-infinite resources available in the cloud. We use the well-known technique of caching to mask latency by executing AI inference for a particular input (e.g. video frame) in the cloud and storing the results on the edge against a "fingerprint", or a hash code, based on features extracted from the input.

This scheme is designed such that inputs being semantically similar (e.g. belonging to the same class) will have fingerprints that are "near" to each

other, according to some distance measure. Figure 1 shows the design of the cache. The encoder creates the fingerprint of an input video frame and searches the cache for fingerprints within a specific distance. If there is a match, then the inference results are provided from the cache, thus avoiding the need to query the AI service running in the cloud.

We find the fingerprints analogous to shadow puppets, two dimensional projections of figures on a screen created by a light in the background. Anyone who has used his/her fingers to create shadow puppets will attest that the absence of detail in these figures does not restrict their ability to be the foundation for good storytelling. The fingerprints are projections of the actual input that can be used for rich AI applications even in the absence of original detail.

We have developed a complete proof of concept implementation of the semantic cache, following an "as a service" design approach, and exposing the service to edge device/gateway users via a REST interface. Our evaluations on a range of diverse edge devices (Raspberry Pi 3/ NVIDIA Jetson TK1/TX1/TX2) have demonstrated that the latency of inference has been reduced by 3 times and the bandwidth usage by at least 50 percent when compared to a cloud-only solution.

Early evaluation of a first prototype implementation of our approach showcases its potential. We are continuing to mature the initial approach, prioritizing on experimenting with alternative encoding techniques for improved precision, while also extending the evaluation to further datasets and AI tasks.

We envisage this technology to have applications in retail, predictive maintenance for industrial facilities, and video surveillance, among others. For example, the semantic cache could be used to store fingerprints of product images at checkouts. This can be used to prevent store losses due to theft or mis-scanning. Our approach serves as an

example of seamlessly switching between cloud and edge services to deliver best-of-breed AI solutions on the edge.

More information: Shadow Puppets: Cloud-level Accurate AI Inference at the Speed and Economy of Edge. www.usenix.org/conference/hote … esentation/venugopal

This story is republished courtesy of IBM Research. Read the original story here.

Provided by IBM