**Motivation**

**Goal:** Represent images and video as emoji.

We transform a multi-modal input into an ideogrammatical representation which can be used for search and summarization.

**Why Emoji?**
- Visual grammar of interaction
- Language independent
- Age accessible
- Widely supported
- Semantically diverse
- Easy form factor for smart phones and watches

**Method**

- **Semantic embedding**, from the text of 100M Flickr photos
- **Multi-modal**, applicable to any embeddable modality
- **Zero-shot**, applicable to any arbitrary set of ideograms

**Visual Modality**

For a target emoji $z$ and input image $x_v$, emoji are scored:

$$S_v(z, x_v) = \sum_{y_v \in \mathcal{Y}_v} \cos(\psi(z), \psi(y_v)) \cdot p(y_v | x_v)$$

where $\psi(\cdot)$ gives a word2vec vector, and $y_v$ is a visual concept.

**Textual Modality**

Given a text input $x_t$, the emoji scores are:

$$S_t(z, x_t) = \max_{y_t \in \mathcal{Y}_t} \cos(\psi(z), \psi(y_t)) \cdot p(y_t | x_t)$$

The final score is a weighted combination across modalities:

$$S(z, x_v, x_t) = \alpha S_v(z, x_v) + (1 - \alpha) S_t(z, x_t)$$

Where $\alpha$ can be manually set or experimentally validated.

**Experimental Results**

**Test Set:** MSCOCO Training (83k images in 80 classes)

**Features:** 15k ImageNet concept probabilities from deep net

Image2Emoji outperforms the supervised visual model.

---

**Search**

Emoji can form semantically dense single ideogram search queries...

...or compose more nuanced queries:

---

**Describe**

Emoji can be used to summarize image collections and video...

---

**Summarize**

Emoji can be used to summarize image collections and video...

...check it out in our demo!