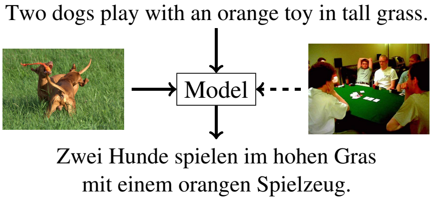
Is Visual Information Needed? (1/2)

- Our text-only English→Hindi was perfect on the “challenge” words.
- Elliott (2018) used MM systems with shuffled, incongruent images.



Two dogs play with an orange toy in tall grass.

Model

Zwei Hunde spielen im hohen Gras mit einem orangen Spielzeug.

  - Only the hierarchical attention was sensitive to images other multi-modal systems performed equally with congruent and incongruent images.

- Caglayan et al. (2019) list other papers where images have not helped much.