Beyond Text and Back Again

Desmond Elliott
University of Copenhagen
Denmark
de@di.ku.dk

ABSTRACT

A talk with two parts covering three modalities. In the first part, I will talk about NLP Beyond Text, where we integrate visual context into a speech recognition model and find that the recovery of different types of masked speech inputs is improved by fine-grained visual grounding against detected objects [2]. In the second part, I will come Back Again, and talk about the benefits of textual supervision in cross-modal speech–vision retrieval models [1].

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; **Image representations**; **Natural language processing**.

KEYWORDS

neural networks, speech recognition, computer vision, retrieval

ACM Reference Format:
Desmond Elliott. 2021. Beyond Text and Back Again. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3442442.3451896

BIOGRAPHY

Desmond is an Assistant Professor at the University of Copenhagen. He received his PhD from the University of Edinburgh, and was a postdoctoral researcher at CWI Amsterdam, the University of Amsterdam, and the University of Edinburgh, funded by an Alain Bensoussan Career Development Fellowship and an Amazon Research Award. His research interests include multimodal and multlingual machine learning, which has appeared in papers ACL, CoNLL, EMNLP and NAACL. He was involved in the creation the Multi30K and How2 multilingual multimodal datasets and has developed a variety of models that learn from these types of data. He co-organised the How 2 Challenge Workshop at ICML 2019, the Multimodal Machine Translation Shared Task from 2016–2018, and the 2018 Frederick Jelinek Memorial Workshop on Grounded Sequence-to-Sequence Learning.

REFERENCES

[1] Bertrand Higy, Desmond Elliott, and Grzegorz Chrupała. 2020. Textual Supervision for Visually Grounded Spoken Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* Association for Computational Linguistics, Online, 2698–2709.

[2] Tejas Srinivasan, Ramon Sanabria, Florian Metze, and Desmond Elliott. 2020. Fine-Grained Grounding for Multimodal Speech Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* Association for Computational Linguistics, Online, 2667–2677.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3451896