Mutimodal Lecture ASR

Anonymous Authors[1]

Abstract

This document provides a basic paper template and submission guidelines. Abstracts must be a single paragraph, ideally between 4–6 sentences long. Gross violations will trigger corrections at the camera-ready phase.

1. Introduction

With an increasing number of classroom lectures, conference talks and marketing presentations recorded and hosted online, there has been an increasing interest in transcribing such videos. This transcription helps deaf students follow these lectures, enables automatic summarization, plagiarism detection and machine translation of the speech making them more accessible to a wider audience. Furthermore, transcription allows the recordings to be more useful as it can now be searched thoroughly according to the content and not just the keywords and titles.

Unfortunately, automatic speech recognition (ASR) in classrooms and presentations is challenging because the speakers use varied domain specific technical jargon. To make a general ASR system work well for technical speech, it would have to be adapted with in-domain language and knowledgeCite a paper which shows the importance of domain adaptation. Fortunately in such scenarios, the presentation is often accompanied with supporting slides which present content and talking points about the speech. There also exists a strong correlation between the slides and the speech(Martínez-Villaronga et al., 2014). The slides which include text, metadata and figures can thus be used as a context for the speech so that the ASR system can adapt itself to pay more attention to the utterances that are supported by the slide contents. Although a lot of instructors use blackboards to write on in class, modern handwriting recognizers perform poorly for this scenario(Dutta et al., 2018), hence we focus on slides where Optical Character Recognition systems perform well. Slides tend to have disproportionately many words from the subject vocabulary and these are also the most important words to get right. this is very interesting. I would try to stress more this in the paper.Akita et. al. (Akita et al., 2015) showed that on a set of Japanese lectures, 53% out of vocabulary words for an ASR system were present in the slides enabling a substantial improvement in word error rates.

Due to lack of an existing dataset for this task, we compile a new dataset of video lectures that have slides visible in them along with their transcripts. The videos are from different instructors and are from various sources like MIT OCW, NPTEL, Stanford etc and cover a wide variety of topics. In total there are 2532 videos with a total video duration of 1648 hours and more than 11 million manually transcribed words by the original source. In most of the videos, the focus alternates between the instructor and the slide and thus we extract the slide text contents as a part of our pipeline.

Add a example slide image for example this should bleong to the second part: What are your contributions?

Thus in our work, we make the following contributions

- We compile and release a new large dataset for video lectures transcription
- We demonstrate that adaptation of the language model of a general end-to-end speech recognition model with textual information from slides and titles improve its performance.
- (speculative- stretch goal) We demonstrate that embedded images from the slides also improve performance of the speech recognition system

2. Related Work

please AIGAIN try to be consistent with the citation of google NAMEYEARFIRSTWORD This paragram should go to related work... Previous efforts with using slide text as context have focused on modifying the language model of the speech recognizer. Miranda et. al.(Miranda et al., 2013) modified the speech recognition lattice by intersecting it with a lattice created from the slide text. As the lecture presentation is often available in the form of a video stream, an Optical Character Recognition (OCR) system is required

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

to read the words, which is then error corrected and used to create an interpolated language model with a base language model (Martínez-Villaronga et al., 2014) (Akita et al., 2015) (Yamazaki et al., 2007). The topic words extracted from the slides with Probabilistic Latent Semantic Analysis also helps adapt the language model for words that are not in the slides(and and, 2008). One can also fix the transcription in a post-processing step by inferring over an HMM which models the transcript as a sequence of phonemes uttered from slide words interspersed with other words(Swaminathan et al., 2010). RNN Language models show improvements in perplexity when provided topic level information(Chen et al., 2015). Recently, end to end ASR systems with textual contexts (Pundak et al., 2018) show significant improvement over baseline indicating advantage of joint optimization over individually trained components.

It is ok to cite some papers to motivate Accuracy of existing speech recognition system degrades with slides that chiefly has images(Miranda et al., 2013). Even though these slides don't provide textual context but the images also provide context and should improve accuracy. Existing work in fact shows that video provides context to a multimodal speech recognition system and improves its accuracy(Gupta et al., 2017). This intuition was extended to end to end ASR systems where visual adaptive training improved performance over sequence to sequence and CTC models (Palaskar et al., 2018) (Caglayan et al., 2018).

3. Format of the Paper

All submissions must follow the specified format.

3.1. Author Information for Submission

ICML uses double-blind review, so author information must not appear. If you are using LaTeX and the `icml2019.sty` file, use `\icmlauthor{...}` to specify authors and `\icmlaffiliation{...}` to specify affiliations. (Read the TeX code used to produce this document for an example usage.) The author information will not be printed unless `accepted` is passed as an argument to the style file. Submissions that include the author information will not be reviewed.

3.1.1. Self-Citations

If you are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (Langley, 2000), we have shown ...").

Do not anonymize citations in the reference section. The only exception are manuscripts that are not yet published (e.g., under submission). If you choose to refer to such unpublished manuscripts (Author, 2018), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look at the Supplementary Material when writing their review.

3.1.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, in a row, separated by white space, and centered. Author names should not be broken across lines. Unbolded superscripted numbers, starting 1, should be used to refer to affiliations.

Affiliations should be numbered in the order of appearance. A single footnote block of text should be used to list all the affiliations. (Academic affiliations should list Department, University, City, State/Region, Country. Similarly for industrial affiliations.)

Each distinct affiliations should be listed once. If an author has multiple affiliations, multiple superscripts should be placed after the name, separated by thin spaces. If the authors would like to highlight equal contribution by multiple first authors, those authors should have an asterisk placed after their name in superscript, and the term "*Equal contribution" should be placed in the footnote block ahead of the list of affiliations. A list of corresponding authors and their emails (in the format Full Name <email@domain.com>) can follow the list of affiliations. Ideally only one or two names should be listed.

A sample file with author names is included in the ICML2019 style file package. Turn on the `[accepted]` option to the stylefile to see the names rendered. All of the guidelines above are implemented by the LaTeX style file.

3.2. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

3.3. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

3.3.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

3.3.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

3.4. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

[1]Footnotes should be complete sentences.

[2]Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

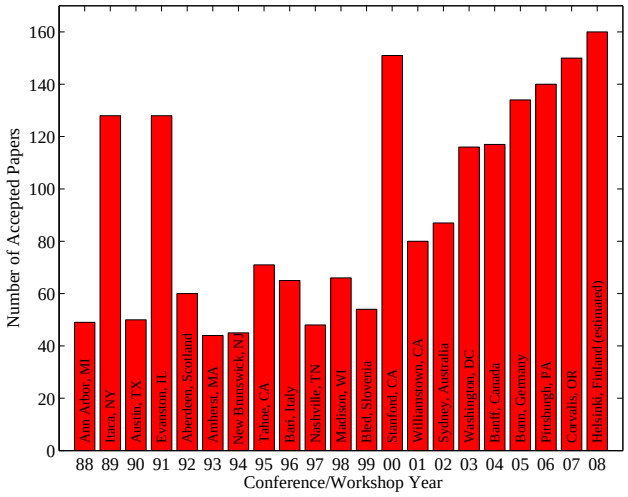

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true.$
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 1. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

3.5. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

3.6. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

3.7. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2019.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (1959). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Samuel, 1959). List multiple references separated by semicolons (Kearns, 1989; Samuel, 1959; Mitchell, 1980). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Michalski et al., 1983).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section 3.1 for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Samuel, 1959), conference publications (Langley, 2000), book chapters (Newell & Rosenbloom, 1981), books (Duda et al., 2000), edited volumes (Michalski et al., 1983), technical reports (Mitchell, 1980), and dissertations (Kearns, 1989).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

3.8. Software and Data

We strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, do not include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look at this material when writing their review.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Akita, Y., Tong, Y., and Kawahara, T. Language model adaptation for academic lectures using character recognition result of presentation slides. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5431–5435, April 2015. doi: 10.1109/ICASSP.2015.7179009.

and and. Automatic lecture transcription by exploiting presentation slide information for language model adaptation.

In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4929–4932, March 2008. doi: 10.1109/ICASSP.2008.4518763.

Author, N. N. Suppressed for anonymity, 2018.

Caglayan, O., Sanabria, R., Palaskar, S., Barrault, L., and Metze, F. Multimodal Grounding for Sequence-to-Sequence Speech Recognition. *arXiv e-prints*, art. arXiv:1811.03865, Nov 2018.

Chen, X., Tan, T., Liu, X., Lanchantin, P., Wan, M., Gales, M. J., and Woodland, P. C. Recurrent neural network language model adaptation for multi-genre broadcast speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Dutta, K., Mathew, M., Krishnan, P., and Jawahar, C. V. Localizing and recognizing text in lecture videos. In *ICFHR*, 2018.

Gupta, A., Miao, Y., Neves, L., and Metze, F. Visual features for context-aware speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5020–5024, March 2017. doi: 10.1109/ICASSP.2017.7953112.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Martínez-Villaronga, A., Agua, M. A., Silvestre-Cerdí, J. A., Andrés-Ferrer, J., and Juan, A. Language model adaptation for lecture transcription by document retrieval. In *Proceedings of the Second International Conference on Advances in Speech and Language Technologies for Iberian Languages - Volume 8854*, IberSPEECH 2014, pp. 129–137, Berlin, Heidelberg, 2014. Springer-Verlag. ISBN 978-3-319-13622-6. doi: 10.1007/978-3-319-13623-3_14. URL https://doi.org/10.1007/978-3-319-13623-3_14.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). *Machine Learning: An Artificial Intelligence Approach, Vol. I*. Tioga, Palo Alto, CA, 1983.

Miranda, J., Neto, J. P., and Black, A. W. Improving asr by integrating lecture audio and slides. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8131–8135, May 2013. doi: 10.1109/ICASSP.2013.6639249.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Palaskar, S., Sanabria, R., and Metze, F. End-to-End Multimodal Speech Recognition. *arXiv e-prints*, art. arXiv:1804.09713, Apr 2018.

Pundak, G., Sainath, T. N., Prabhavalkar, R., Kannan, A., and Zhao, D. Deep context: end-to-end contextual speech recognition. *arXiv e-prints*, art. arXiv:1808.02480, Aug 2018.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.

Swaminathan, R., Thompson, M. E., Fong, S., Efrat, A., Amir, A., and Barnard, K. Improving and aligning speech with presentation slides. In *2010 20th International Conference on Pattern Recognition*, pp. 3280–3283, Aug 2010. doi: 10.1109/ICPR.2010.802.

Yamazaki, H., Iwano, K., Shinoda, K., Furui, S., and Yokota, H. Dynamic language model adaptation using presentation slides for lecture speech recognition. In *INTERSPEECH*, 2007.

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.